基于BERT模型的科技政策文本分类研究*

沈自强[1,2] 李晔[1,2] 丁青艳[3] 王金颖[1,2] 白全民[1,2]
(1. 齐鲁工业大学（山东省科学院）经济与管理学部，济南 250014；2. 山东省科技发展战略研究所，济南 250014；
3. 山东省计算中心（国家超级计算济南中心），济南 250014)

摘要：在智慧政务的应用背景下，利用深度学习的方法对海量的科技政策文本数据进行自动分类，可以降低人工处理的成本，提高政策匹配的效率。利用BERT深度学习模型对科技政策进行自动分类实验，通过TextRank算法和TF-IDF算法提取政策文本关键词，将关键词与政策标题融合后输入BERT模型中以优化实验，并对比不同深度学习模型的分类效果来验证该方法的有效性。结果表明，通过BERT模型，融合标题和TF-IDF政策关键词的分类效果最佳，其准确率可达94.41%，证明利用BERT模型在标题的基础上加入政策关键词能够提高政策文本自动分类的准确率，实现对科技政策文本的有效分类。

关键词：科技政策；文本分类；BERT模型；关键词提取
中图分类号：TP311　　DOI：10.3772/j.issn.1673-2286.2022.01.002
引文格式：沈自强，李晔，丁青艳，等. 基于BERT模型的科技政策文本分类研究[J]. 数字图书馆论坛，2022(1)：10-16.

十九届五中全会通过的《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》中强调“完善国家创新体系，加快建设科技强国”[1]。科技政策是建设科技强国的重要保障，科技政策文本是体现政策工具的主要载体。党的十八大提出创新驱动发展战略后，我国日益重视自主创新工作，各部门围绕科技创新工作密集出台了大量科技政策文件。然而，频繁的政策发布，使得政策文本的数据量越来越大，对政策文本进行筛选、分析的人力成本激增[2]。通过分析发现，这些科技政策由不同政府部门在不同的时间发出，分布在不同的政务网站上，但由于缺少高效的信息处理手段，高校院所、企业、科技人才等创新主体在配对合适自己的科技政策时存在成本高、耗时久，乃至找不到的困难，影响了科技政策的实施效果。因此，有必要预先为科技政策文本合理地设定分类标签，从而提高检索和配对的效率，帮助各类创新主体找到与自身需求相关的科技政策信息。

在政府政务信息化、智慧化转型趋势下，大数据赋能的政策信息的智能查询、匹配、推送等个性化服务需求日益凸显。自然语言处理、深度学习等计算机信息技术的快速发展为满足这一需求提供了条件。但是在政策文本挖掘领域，许多新技术的应用仍处于探索阶段，并且由于政策文本的长短不一、信息密度大、分类体系不统一等特点，研究人员在借助信息技术手段对政策文本进行自动分类时会遇到困难，尚未形成得到广泛认可的技术方案。针对上述情况，本文以科技政策为出发点，采用谷歌公司AI团队在2018年发布的预训练语言表示模型——BERT（Bidirectional Encoder Representations from Transformers）模型[3]，结合关键词提取技术，融合政策标题与关键词信息作为模型训练语料，对科技政策文本进行自动分类研究，从而实现更加准确快速的科技政策文本归类，降低人力成本，

* 本研究得到山东省高等学校青创科技支持计划“智能时代的产业变革：技术、制度与创业导向”（编号：2020RWG009）资助。

助力政务大数据发展。

1 相关研究

科技政策是由国家或地方政府机构为促进经济社会发展，基于社会需求在不同阶段制定颁布的，一系列用于规制和激励全社会从事知识发现、积累，以及应用于技术创新行为的政策集合，包括规划计划、法规条例、决定、办法、措施，以及相应的实施细则、意见建议等[4]。近年来，中国进入了科技政策颁布的密集期，出台了一系列促进科技发展的配套政策。对于科技政策的分类可以从多个角度进行划分。从政策的发文机构来看，涉及国务院、科技部、工信部、财政部、商务部、国税局等，以及各地方政府机构[4]；从政策针对的创新主体来看，涉及企业、高校院所、科技人才、科技中介、创新平台、产学研联合体等[5]；从政策涉及的领域来看，涉及第一产业、第二产业和第三产业[6]。目前对科技政策的研究通常采用政策工具视角下的分类方式，将科技政策划分为供给型政策（人才支持、资金支持、技术支持、基础设施建设等）、环境型政策（法规管制、目标规划、金融支持、税费减免、知识产权等）和需求型政策（政府采购、外包、贸易管制、海外机构管理等）三大类[7-8]。

文本分类是指从原始文本数据中提取特征，并基于这些特征预测文本数据的类别，作为有效的信息检索和挖掘技术，其在管理文本数据中起着至关重要的作用[9]。传统机器学习方法如朴素贝叶斯、支持向量机等技术表现出的分类效果相对较差[10]。随着深度学习的发展，文本的表征方式从空间向量模型发展到word2vec词向量模型，基于FastText、CNN、RNN、LSTM等神经网络语言模型的文本分类技术得到广泛应用，并涌现出各种变体[11]，随后ELMo、BERT等通用预训练语言模型的出现有效提高了文本分类等自然语言处理任务的实验效果。目前，针对中文文本分类任务的研究主要包括社交文本的情感分析[12-13]、新闻文本的分类任务[14-15]和专利的自动分类[16]等。在政策文本分类领域，杨锐等[17]将能源政策划分为投资开发与建设类、科技与产业装备类、安全生产管理类和市场调节与监管类，通过Doc2Vec提取主题信息并将其融入卷积神经网络的方法有效提升了自动文本分类的效果。胡吉明等[18]从政策涉及的产业领域角度进行分类，利用LDA模型和改进的TextRank模型增强政策文本的表示效果，采用CNN-BiLSTM-Attention的集成模型来提升政策文本分类的效果和准确度。张雨等[19]在科技政策知识图谱研究中根据政策内容训练Bi-LSTM模型对科技政策文本进行情感分类，将政策文本按照句子级别划分扶持型、禁止型、普通型三类。虽然上述研究人员开始尝试将深度学习的技术应用在政策文本领域以达到批量自动分类的目的，但由于政策文本具有结构复杂、信息密度大且内涵分布不均衡等特点，因此，这样的研究仍然较少。此外，已有研究在政策分类的标签划分上较为简单，在文本的特征提取上也没有应用BERT、XLNet等新兴起的预训练语言模型，对文本语义的理解还有待提升。

通过对各类科技政策文本进行深入解读后发现，政策文本的语义特征是其重要特征之一[20]，如“组织开展离岸创新人才认定和引才用才补贴申请工作的通知”，这个标题中涉及了人才和资金支持两个方面，从政策语义来说，其表达的含义是人才认定和人才激励，因此属于人才支持类的政策。BERT模型具有较强的文本语义理解能力，在训练过程中可以更好地获得了一个句子的语义表达[21]。段瑞雪等[22]将BERT模型应用于文本分类、机器阅读理解和文本摘要3个下游任务中，并通过对比实验展示了BERT模型的优越性。因此，本研究从科技政策文本出发，在梳理过往分类标准和分析科技政策文本特点的基础上，结合BERT深度学习模型，对8 761条科技政策文本进行分类研究，提取出9个目标类别进行实验，融合科技政策标题和关键词作为训练语料训练模型，以提升分类实验的准确度，实现科技政策文本的自动分类。

2 研究思路

2.1 研究框架

基于BERT深度学习模型的科技政策文本分类方法的研究框架如图1所示，主要包括科技政策文本数据采集与预处理、科技政策文本关键词提取、科技政策文本分类训练三个环节。首先采集科技政策文本数据，确定分类维度并进行人工标注，再对科技政策文本数据进行清洗，得到样本数据；接着对科技政策的正文文本进行关键词提取，将关键词拼接在政策标题后面作为训练文本，形成数据集；最后对3个实验数据集划分训练集和测试集，构建BERT模型，分别进行实验，并对实验结果进行对比分析。

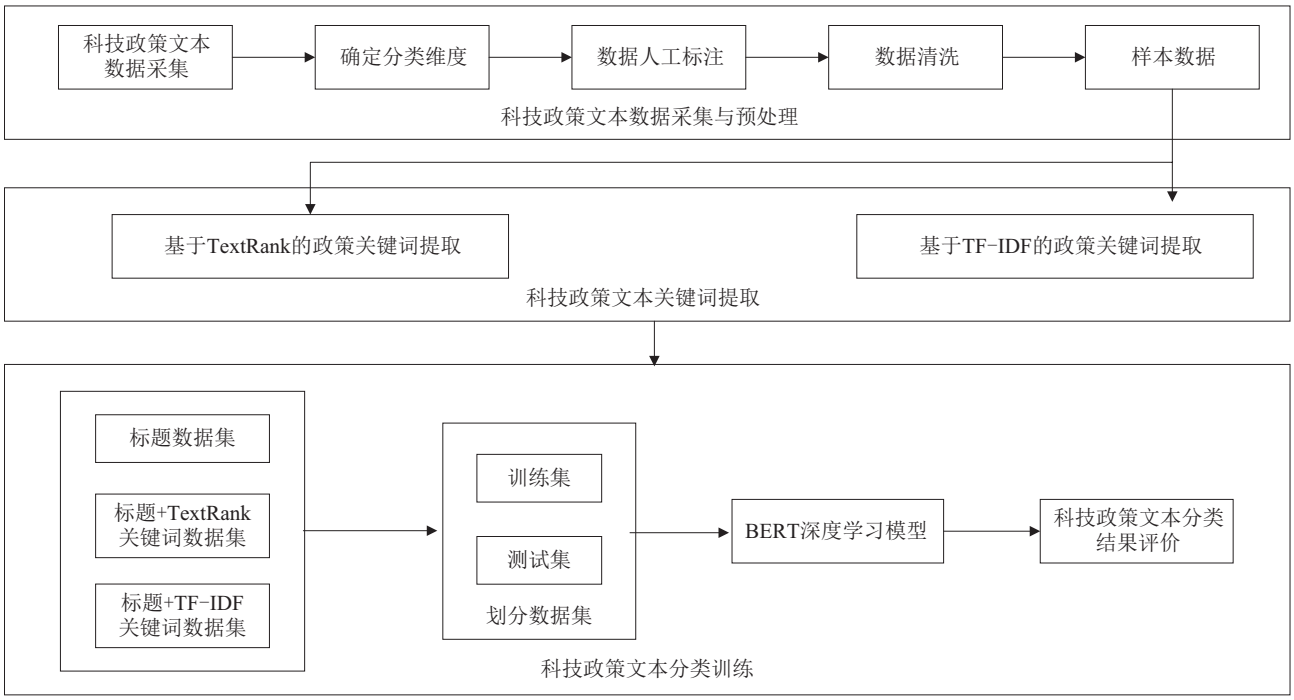

图1 科技政策文本分类框架图

2.2 理论模型

本文采用BERT模型对科技政策文本进行特征提取并自动分类。BERT模型的结构由多层Transformer模型组成，用于文本的特征提取。Transformer的核心是注意力机制，其核心思想是计算一个句子中的每个词与这句子中所有词的相互关系，并认为词与词之间的相互关系在一定程度上反应了这句话中不同词之间的关联性和重要程度，再利用这些相互关系来调整每个词的重要性（权重）就可以获得每个词的新表征。这个新表征既包含了该词本身，还包含了其他词与这个词的关系，因此模型能够获得词语的上下文语义信息。

BERT模型由两个阶段组成，第一阶段是模型在大量通用语料上的预训练过程，可以学习到通用的语义信息；第二阶段是模型在专业语料库上的微调过程，能够得到专业领域内的语义新特征，从而在具体任务中有良好的表现。

2.3 算法流程

基于BERT模型的科技政策文本分类算法运行流程总体上分为数据读入、特征转化、模型构建和模型运行4个步骤。①读入准备好的科技政策文本实验数据集，每一条数据按照文本、tab分隔符、标签的形式存储在文档中；②将读入的数据转化为特征向量的形式，并记录到TFRecord格式文件中，特征向量包含字向量、分段向量和位置向量3个部分；③创建和配置BERT模型，读取TFRecord格式文件并将特征向量转化成标准的模型输入，这一步主要通过BERT的Transformer层来实现；④根据设置好的参数运行构建好的模型，用训练集的标准输入进行训练，用测试集的标准输入进行评估，输出分类结果。

3 实验过程

3.1 数据采集与预处理

本文实验的科技政策数据来源于国家中小企业政策信息发布平台[23]和其他相关机构网站，采用网络爬虫技术和八爪鱼软件进行科技政策采集，共采集科技政策10 388条。对采集的科技政策数据进行预处理，清除空值、去重、去除掉各种与科技政策内容无关的信息。通过专家讨论并借鉴过往文献[6]的分类标准，在需求型、供给型和环境型下确定科技政策分类类别。由于需求型政策数量较少，而对类别不均衡数据分类不是本文研究的重点，因此从供给型和环境型两个方面

提取出9个目标类别对科技政策完成了人工标注，共计8 761条，如表1所示。

表1 9类政策文本数量分布

政策文本类别	数量/条	备 注
人才支持	1 009	人才引进、人才培训、人才激励等
资金支持	1 040	财政投入、资金投入等
技术支持	839	技术引进、技术交易、技术奖励、技术创新等
法规管制	1 270	监督管理、法律规章、管理办法等
目标规划	1 025	发展规划、行动计划等
知识产权	893	知识产权、专利、品牌保护等
创新平台建设	916	开发区、园区、基地、平台等
税费减免	906	财税管理、税收优惠
金融支持	863	投融资、贷款 、保险、质押担保等

3.2 关键词提取

由于科技政策文本篇幅较长，不利于分类模型的训练[18]，往往需要提取出用于分类的关键信息，而关键词是对政策信息最大限度的浓缩，因此本文通过TextRank算法[24-25]以及TF-IDF算法[26]提取科技政策文本的关键词，将两种方法提取的关键词分别拼接在政策标题后面作为实验数据集。在采用两种算法进行科技政策文本关键词提取时，所提取的关键词数量值均设置为20。

3.3 分类实验设置

本文通过Tensorflow深度学习框架实现BERT科技政策文本多分类任务，按照4:1的比例将科技政策文本数据集划分为训练集和测试集。模型运行时加载的预训练语言模型是BERT官方的BERT-Base-Chinese模型，该模型的Transformer共有12层，词向量维度为768维，多头注意力机制参数是12，其他实验参数设置如表2所示。在BERT模型上进行3次实验，实验1对科技政策标题数据集进行实验；实验2将科技政策标题与TextRank算法提取出的关键词拼接后输入BERT模型进行实验；实验3将科技政策标题与TF-IDF算法提取出的关键词拼接后输入BERT模型进行实验。

此外，为了验证BERT模型的有效性并比较不同深度学习模型在科技政策文本分类上的效果，选取FastText、LSTM-Attention、TextCNN 3个模型进行对比试验。将所采集到的全部科技政策文本作为语料库，利用word2vec训练文本词向量，词向量维度为300维，对训练文本加载训练好的词向量模型进行科技政策文本的向量表示，分别输入上述3个模型中进行分类实验，每个批次的训练样本数及最大句子长度与BERT模型做相同设置，模型训练次数epochs均设置为20，dropout率均为0.5，损失函数采用交叉熵损失函数，激活函数采用softmax，最后对不同模型的分类结果进行对比分析。

表2 实验参数设置

参 数	参数解释	参数值
num_train_epochs	训练数据被训练的次数	3
batch_size	训练时每个batch的样本数量	32
learning_rate	学习率	5×10^{-5}
max_seq_length_1	标题最大句子长度	60
max_seq_length_2	标题拼接关键词最大句子长度	100
num_keyword	关键词提取的词语数量	20

4 实验结果与对比分析

对于科技政策文本分类的效果采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1值4个指标进行评价[27]，全部实验结果如表3所示。

对比4种深度学习模型的实验结果可以看出，基于BERT模型的科技政策文本分类效果最好，各个指标结果均优于其他模型，除BERT模型之外，TextCNN模型在科技政策文本分类效果上也有较好的表现，各评价指标也超过了90%，相比TextCNN模型，LSTM-Attention模型的分类结果则较差，而FastText模型的表现是最差的，由此得出模型分类结果为“BERR>TextCNN>LSTM-Attention>FastText”，可以验证BERT模型在科技政策文本分类上的优越性。此外，在每个模型上，将两种算法提取出的政策关键词与政策标题拼接后作为训练文本进行训练时，分类效果在总体上均有所提升，这种提升在FastText模型上最为明显。将TF-IDF算法提取出的政策关键词与标题进行融合后的分类实验效果均为最佳，在BERT模型上融合标题和TF-IDF关键词进行训练时的准确率和F1值能够达到94.41%和94.59%，是最佳实验，比仅使用标题进行训练时准确率和F1值分别提升了1.21个百分点和1.38个百分点。由此得出训练文本的分类结果为“Title+TF-

表3 实验结果

%

实 验	训练文本	准确率	精准率	召回率	F1
FastText	Title	74.36	75.33	73.78	74.44
	Title +TextRank	79.63	80.22	79.67	80.11
	Title +TF-IDF	82.19	82.56	82.33	82.33
LSTM-Attention	Title	88.99	88.78	89.44	88.89
	Title +TextRank	89.22	89.56	89.33	89.44
	Title +TF-IDF	91.04	91.00	91.11	91.22
TextCNN	Title	90.93	91.67	90.56	90.89
	Title +TextRank	91.16	91.44	91.33	91.22
	Title +TF-IDF	91.79	92.00	91.67	91.78
BERT	Title	93.20	93.09	93.19	93.21
	Title +TextRank	93.95	94.20	94.05	94.12
	Title +TF-IDF	94.41	94.54	94.65	94.59

IDF>Title+TextRank>Title”，可以验证科技政策文本关键词能够作为重要信息指导模型进行分类，TF-IDF关键词比TextRank关键词更具有指导意义。

在BERT模型中，采用标题融合TF-IDF关键词进行训练时的分类效果是最佳的，每个政策类别的分类结果如图2所示。可以看出，“人才支持”“税费减免”和“目标规划”3个类别的分类效果最好，其F1值均超过了95%；“创新平台建设”“知识产权”和“技术支持”3个类别次之；“法规管制”“金融支持”“资金支持”3个类别的分类效果较差，但各评价指标在90%左右，其中“法规管制”分类效果较差，因为其作为常用的政策工具，经常用于规范科技创新的各个领域，容易与其他政策类别相重合，出现分类错误的可能性较大。总的来看，在BERT模型上，融合标题与TF-IDF关键词的方法能够

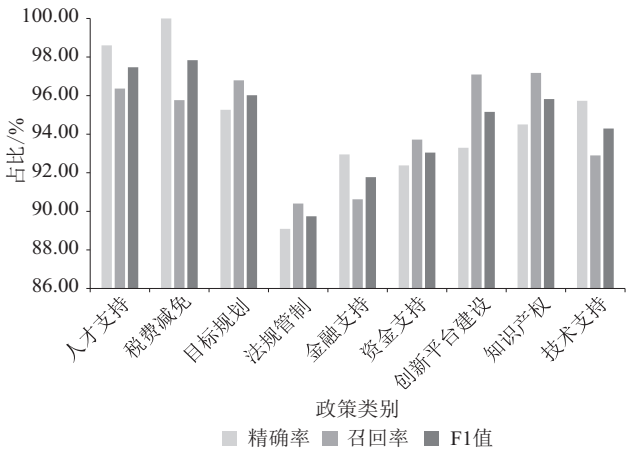

图2 各政策类别的分类结果

较为准确地实现对科技政策文本的自动分类。

5 结论与展望

随着信息技术的发展，政务大数据研究成为热点方向。通过大数据技术赋能政策文本的查询、匹配、推送等智能化服务的基础步骤之一就是实现政策文本的自动分类。本文为政策领域的文本分类提供了参考实例，在政策工具视角下提出一种基于BERT模型和关键词提取技术相结合的科技政策文本分类方法，致力于大量科技政策文本的自动分类，得出如下结论：首先，在FastText、LSTM-Attention、TextCNN、BERT 4个模型的对比实验中，验证了BERT模型在科技政策文本分类领域上的优越性；然后，通过TF-IDF算法和TextRank算法对政策正文进行关键词提取，将关键词与政策标题进行拼接融合后进行分类训练有效提升了分类效果，证实了政策文本关键词对科技政策文本分类具有指导意义，且TF-IDF关键词的指导意义更大；最后，该方法实现了对科技政策文本较为准确的自动分类，最佳实验的准确率和F1值能够达到94.41%和94.59%，在“人才支持”“税费减免”和“目标规划”3个政策类别上的识别效果最好。然而本文依然存在一些不足之处，缺少对政策类别分布不均衡这一问题的考虑，以及如何较为完整地提取政策文本的关键内容还有待深入研究。

政策文本自动分类具有广阔的发展空间和应用前景，未来可以往两个方面进行拓展。一是制定政策分类标准体系，构建专业的政策分类语料库。目前分类标准

不清晰和语料库建设成本的局限，使政策文本自动分类的维度是片面的、低实用性的，难以满足政策分析者的需要，因此需要加强跨学科合作，注重分类标准与应用背景的结合，构建适用于政策文本分类的专用语料库，为技术发展提供基础条件。二是将深度学习与自然语言处理技术在政策文本分类领域做进一步拓展。根据政策的文本特点，引入和开发新兴技术，需要探究多任务和多标签的文本分类技术、长文本分类技术等在政策分类上的应用，搭建智能政策服务平台，以满足政府、企业、高校院所和科技人才等创新主体对前沿政策信息的获取与捕捉。

参考文献

[1] 中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议 [EB/OL]. [2021-07-12]. http://www.gov.cn/zhengce/2020-11/03/content_5556991.htm.

[2] 郑新曼，董瑜. 政策文本量化研究的综述与展望 [J]. 现代情报，2021，41（2）：168-177.

[3] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding [J/OL]. arXiv [2021-11-20]. https://arxiv.53yu.com/abs/1810.04805v2.

[4] 章刚勇. 基于大数据的中国科技政策体系研究：理论与实践 [J]. 中国软科学，2018（6）：172-180.

[5] 宋娇娇，孟溦. 上海科技创新政策演变与启示——基于1978—2018年779份政策文本的分析 [J]. 中国科技论坛，2020（7）：14-23.

[6] 赵筱媛，苏竣. 基于政策工具的公共科技政策分析框架研究 [J]. 科学学研究，2007（1）：52-56.

[7] ROTHWELL R，ZEGVELD W. Reindustrialization and technology [M]. London：Logman Group Limited，1985.

[8] 徐硼，罗帆. 政策工具视角下的中国科技创新政策 [J]. 科学学研究，2020，38（5）：826-833.

[9] LI Q，PENG H，LI J，et al. A Survey on text classification：from shallow to deep learning [J/OL]. arXiv [2021-11-20]. https://doi.org/10.48550/arXiv.2008.00364.

[10] 万家山，吴云志. 基于深度学习的文本分类方法研究综述 [J]. 天津理工大学学报，2021，37（2）：41-47.

[11] 贾澎涛，孙炜. 基于深度学习的文本分类综述 [J]. 计算机与现代化，2021（7）：29-37.

[12] 蔡庆平，马海群. 基于Word2Vec和CNN的产品评论细粒度情感分析模型 [J]. 图书情报工作，2020，64（6）：49-58.

[13] 王婷，杨文忠. 文本情感分析方法研究综述 [J]. 计算机工程与应用，2021，57（12）：11-24.

[14] 蓝雯飞，徐蔚，汪敦志，等. 基于LSTM-Attention的中文新闻文本分类 [J]. 中南民族大学学报（自然科学版），2018，37（3）：129-133.

[15] 胡玉兰，赵青杉，陈莉，等. 面向中文新闻文本分类的融合网络模型 [J]. 中文信息学报，2021，35（3）：107-114.

[16] 杨辰，王楚涵，陶琬莹，等. 基于专利的技术机会识别：深度学习领域的案例分析 [J]. 科技管理研究，2021，41（12）：172-176.

[17] 杨锐，陈伟，何涛，等. 融合主题信息的卷积神经网络文本分类方法研究 [J]. 现代情报， 2020，40（4）：42-49.

[18] 胡吉明，付文麟，钱玮，等. 融合主题模型和注意力机制的政策文本分类模型 [J]. 情报理论与实践，2021，44（7）：159-165.

[19] 张雨，吴俊. 科技政策知识图谱构建研究 [J]. 数字图书馆论坛，2021（8）：31-38.

[20] 郑新曼，董瑜. 基于科技政策文本的程度词典构建研究 [J]. 数据分析与知识发现，2021，5（10）：81-93.

[21] 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.

[22] 段瑞雪，巢文宇，张仰森. 预训练语言模型BERT在下游任务中的应用 [J]. 北京信息科技大学学报（自然科学版），2020，35（6）：77-83.

[23] 国家中小企业政策信息互联网发布平台. 政策文件库 [EB/OL]. [2021-07-12]. https://www.sme-service.cn/#/initialArticle.

[24] 陈芬. 基于Word2Vec与TextRank的关键词抽取研究 [D]. 武汉：华中师范大学，2020.

[25] MIHALCEA R，TARAU P. TextRank：Bringing Order into Texts [EB/OL]. [2021-07-12]. https://www.researchgate.net/publication/200042361_TextRank_Bringing_Order_into_Text.

[26] GERARD S，CHRISTOPHER B. Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management，1988，24（5）：513-523.

[27] 王艳，王胡燕，余本功. 基于多特征融合的中文文本分类研究 [J]. 数据分析与知识发现，2021，5（10）：1-14.

作者简介

沈自强，男，1997年生，硕士研究生，研究方向：科技政策、自然语言处理。
李晔，男，1981年生，博士，研究员，通信作者，研究方向：深度学习、信息技术，E-mail：liye@sdas.org。
丁青艳，女，1981年生，博士，研究员，研究方向：管理系统工程、复杂系统优化。
王金颖，女，1986年生，博士，助理研究员，研究方向：科技政策、创新管理。
白全民，男，1983年生，博士，副研究员，研究方向：科技政策、创业与战略管理。

Research on Science and Technology Policy Text Classification Based on BERT Model

SHEN ZiQiang[1,2] LI Ye[1,2] DING QingYan[3] WANG JinYing[1,2] BAI QuanMin[1,2]
(1. School of Economics and Management, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, P. R. China;
2. Institute of Science and Technology for Development of Shandong, Jinan, 250014, P. R. China;
3. Shandong Computer Science Center (National Super Computer Center in Jinan), Jinan 250014, P. R. China)

Abstract: In the context of the application of smart government, this article uses deep learning methods to automatically classify massive amounts of scientific and technological policy text data in order to reduce the cost of manual processing and improve the efficiency of policy matching. This paper used the BERT deep learning model to automatically classify science and technology policies. It extracted the keywords of the policy text through the TextRank algorithm and the TF-IDF algorithm, then integrated the policy titles and policy keywords into the BERT model, so as to optimize the experiment and improve the effect and accuracy of policy text classification. It also made a comprehensive comparative analysis of the classification effect on different deep learning models to show the superiority of this method. The results show that the classification effect of combining the title and TF-IDF policy keywords is the best through the BERT model, and the accuracy rate can reach 94.41%, which proves that adding policy keywords on the basis of the title can improve the accuracy of automatic classification of policy texts on BERT model. Our research achieves an efficient classification of science and technology policy texts.

Keywords: Science and Technology Policy; Text Classification; BERT Model; Keyword Extraction

（收稿日期：2021-12-08）

书 讯

《汉语主题词表》

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有作用，曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要，由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，自2009年开始进行重新编制工作，拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表（工程技术卷）》共收录优选词19.6万条，非优选词16.4万条，等同率0.84，在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表（自然科学卷）》共收录专业术语12.4万条，包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域，收词系统、完整，语义关系丰富、严谨，每条词汇都有相应的学科分类号表现其专业属性，并与同义英文术语对应。同时，建立《汉语主题词表》网络服务系统，提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘，是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，分为13个分册，总定价3 880元。《汉语主题词表（自然科学卷）》已于2018年5月由科学技术文献出版社出版，分为5个分册，总定价1 247元。两卷均可分册购买。