

基于WordBERT和BiLSTM的政策工具自动分类方法研究

Automatic Classification Method of Policy Tools Based on WordBERT and BiLSTM

霍朝光^{1,2} 霍帆帆¹ 王婉如³ 余芊蓉¹ 杨冠灿¹
HUO Chaoguang HUO Fanfan WANG Wanru YU Qianrong YANG Guancan

(1.中国人民大学信息资源管理学院,北京,100872; 2.中国人民大学数字人文研究院,北京,100872; 3.首都经济贸易大学财政税务学院,北京,100070 / 1.School of Information Resources Management, Renmin University of China, Beijing, 100872; 2.Institute of Digital Humanities, Renmin University of China, Beijing, 100872; 3.School of Public Finance and Taxation, Capital University of Economics and Business, Beijing, 100070)

摘要:【目的/意义】政策工具是政府为实现政策目标,将其行政理念转为切实行动的手段和方法,是政策分析的重要维度。鉴于当前政策工具分析仍停留在手工分类阶段,存在标准不统一、难以复现、规模小、成本高等一系列问题,提出构建政策工具自动分类模型。【研究设计/方法】系统梳理了现有的政策工具分类框架,在 Rothwell 和 Zegveld 政策工具分类体系基础上,提出基于 WordBERT 和 BiLSTM 构建政策工具自动分类模型,以数据治理和数字经济政策数据集为例,自主构建数据集,进行三组实验,验证模型优劣。【结论/发现】研究结果表明提出的政策工具自动分类模型效果最好,准确率达到73.91%,为繁琐的政策工具分类提供了一种相对高效的自动分类方法。【创新/价值】针对政策工具自动分类这一学科难题,提出利用无监督表示学习和有监督机器学习等算法,进行政策工具自动分类,以锻造政策工具自动量化分析新模块。

关键词: 政策工具; 自动分类; WordBERT; BiLSTM; 政策计量

中图分类号: G202 **DOI:** 10.13366/j.dik.2023.03.129

引用本文: 霍朝光, 霍帆帆, 王婉如, 等. 基于 WordBERT 和 BiLSTM 的政策工具自动分类方法研究 [J]. 图书情报知识, 2023, 40 (3): 129-138. (Huo Chaoguang, Huo Fanfan, Wang Wanru, et al. Automatic Classification Method of Policy Tools Based on WordBERT and BiLSTM [J]. Documentation, Information & Knowledge, 2023, 40 (3): 129-138.)

Abstract: [Purpose/Significance] Policy tools are the means and tools for government to implement govern targets and visions, which are important research domain for policy analysis. Considering that the current analysis of policy tools is still in the stage of manual classification, and there are a series of problems such as inconsistent standard of coding check, hard to reproduce, small scale, high cost and so on, this paper proposes to build an automatic classification model of policy tools. [Design/Methodology] This paper systematically sorted out the existing policy tool classification framework, and on the basis of Rothwell and Zegveld policy tool classification system, proposed an automatic policy tool classification model based on WordBERT and BiLSTM. Taking the the datasets of data governance and digital economy policy as an example, we independently constructed the data set and carried out three sets of experiments to verify the advantages and disadvantages of the model. [Findings/Conclusion] We find the automatic classification model of policy tools proposed in this paper work best, with a precision of 73.91%, which provide a highly efficient automatic classification method for the tedious work of policy tools classification. [Originality/Value] Aiming at the difficult problem of automatic classification of policy tools, this paper proposes to utilize unsupervised representation learning and supervised machine learning algorithms for automatic classification, so as to provide a strong tool for policiometrics analysis.

Keywords: Policy tools; Automatic classification; WordBERT; BiLSTM; Policiometrics

1 引言

政策工具 (Policy Tools/ Policy Instruments, 又称政府工具、治理工具) 是指在国家治理过程中, 政府为实现政策目标, 将其行政理念转为切实行动的手段和方法, 是政府发挥引导、规范、提倡、支持、约束、防止、惩罚等作用的具体措施^[1]。政策工具集中体现在国

家发布的政策文件中, 是政策文本分析和研究的重要维度。例如以“政策工具”为主题词, 在北京大学核心期刊目录和南京大学核心期刊目录CNKI数据库中检索, 可以发现自2011年以来共计发文3,842篇, 2020年、2021年发文均在500篇以上。关于政策工具的研究呈逐年上升的趋势, 其研究体量和增长趋势如图1所示。

【基金项目】本文系中国人民大学科学研究基金项目 (中央高校基本科研业务费专项资金资助) “基于跨语言知识图谱的政策对冲分析” (22XNF053) 的研究成果之一。(This is an outcome of the Scientific Research Foundation Project of Renmin University of China “Policy Hedging Analysis based on Cross-language Knowledge Graph” (22XNF053) supported by Basic Scientific Research Expenses Fund of Central Universities.)

【通讯作者】王婉如 (ORCID: 0009-0001-1879-2013), 博士, 讲师, 研究方向: 数字经济与财税政策, Email: accwang@163.com. (Correspondence should be addressed to WANG Wanru, Email: accwang@163.com, ORCID: 0009-0001-1879-2013)

【作者简介】霍朝光 (ORCID: 0000-0002-5063-0938), 博士, 副教授, 研究方向: 政策信息学与文本挖掘, Email: huochaoguang@126.com; 霍帆帆 (ORCID: 0000-0001-8392-8995), 博士研究生, 研究方向: 政策信息分析与数据治理, Email: huo_ff@163.com; 余芊蓉 (ORCID: 0000-0002-0963-5059), 硕士研究生, 研究方向: 机器学习, Email: yuqianrong77@ruc.edu.cn; 杨冠灿 (ORCID: 0000-0003-3322-6092), 博士, 副教授, 研究方向: 专利分析与机器学习, Email: yanggc@ruc.edu.cn.

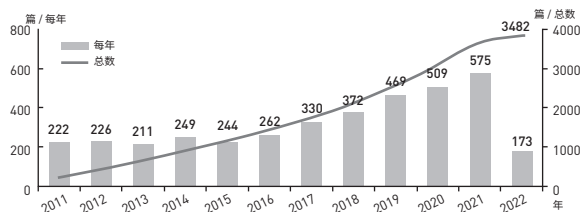


图1 政策工具相关研究统计

Fig.1 The Statistics of Research Related to Policy Tools

任何领域或行业政策都需要分析其使用的政策工具。关于政策工具分析的需求，几乎覆盖科技、金融^[2]、环境^[3]、气候^[4]、卫生^[5]、医药、生物、教育^[6]、养老等所有行业和领域。例如黄萃等对我国69份少数民族双语教育政策进行了手工分类，分析了能力建设工具、象征和劝诫工具、权威工具、激励工具和系统变革等政策工具的使用情况^[7]。孙建军等对我国131份智慧城市政策中的政策工具进行了手工分类，基于R-Z框架对其中的供给型、环境型、需求型政策工具进行了分析，基于Doelen框架对其中的管制规制类、激励诱因类、信息交互类等政策工具进行了分析^[8]。卢小宾等对我国278份信息公开核心政策进行政策工具分类，对供给型、环境型、需求型等下属的二级政策工具的使用情况进行了分类和分析^[9]。张蓓等对我国30份应急管理政策进行政策工具分类和分析^[10]。林德明等通过专家咨询等手工分类的方式，从三份知识产权战略纲领性文件和九份国家知识产权战略的年度推进计划中析出政策工具，并将政策工具和战略目标文本分别进行向量表示，尝试通过语义相似度计算来实现政策工具与战略目标的自动匹配^[11]。

政策工具是政策量化分析的重要版块，然而现有的政策工具分类（如图2所示）仍存在诸多不足。其一，政策工具分类工作仍停留在手工处理阶段，以我国核心期刊近5年数据为例，在662篇政策工具分类研究中，政策工具的分类工作几乎全部都是由手工分类完成（统计时间2022年4月22日）。其二，虽然多人协同手工分类工作通常以Cohen's kappa系数^[12]、Holsti一致性百分比^[13]等为依据，判断不同分类者对同一段文本独立分类的一致性，但是仍然可以发现不同分类工作信度参差不齐，信度为82%、88%、92.7%、95.63%等各种情况均有出现，以80%的信度阈值判断分类工作的合理性仍存在一定争议。其三，手工分类无法避免认知偏差，即使具有较高专业素养的人员依然会存在认知偏差，不同人员所分类的结果定有所不同，即使邀请同一专家对数据多次重复分类，其结果也会存在一定的偏差。质性分析的固有局限，严重影响到分类结果的复现率，甚至导致很多分析结果出现偏差、难以复现。其四，手工分类工作需要耗费大量时间和精力，不仅要求数据分类人员对所有政策工具以及子工具有深刻认识，将十几种政策工具熟记于心，还需要其对相关行业和领域的政策有全面的了解，对数据分类人员要求较高。尤其针对大量的政策数据，工作相当枯燥繁琐，数据分类人员很难完成大规模的政策工具分类，这限制了政策工具的大规模解析。这也正是以往研究中分类样本往往非常有限的主要原因。

由此可见，政策工具分类迫切需要提升分类速度和效率，需要利用计算机技术实现政策工具的自动分类。本文提出构建政策工具自动分类模型，利用文本

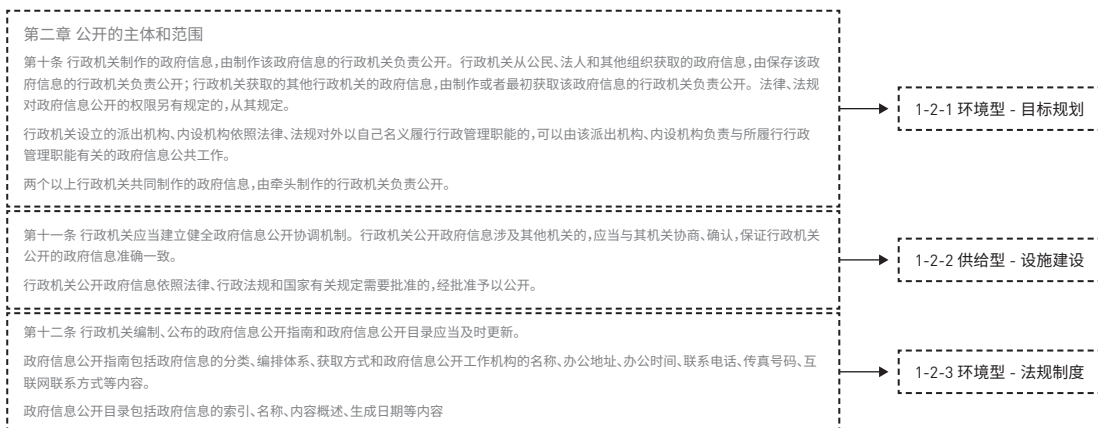
图2 政策工具手工分类示例^[9]

Fig.2 Examples of Manual Classification of Policy Tools

(technical support)、公共服务(public enterprise)、设施建设(infrastructure development)五个子类。人才培养/教育是指政府通过开展教育和培训,培养相关专业人才,为相关工作开展提供人才保障;资金支持是指政府通过拨付专项资金,对相关工作开展提供资金支持;科技支持是指政府通过提供技术支持推动相关工作的开展;公共服务是指政府通过提供相应的基本配套服务,来保障工作顺利开展;设施建设是指政府通过建设和完善基础设施,为相关工作开展提供基本物质保障。

环境型(environmental oriented)政策工具主要包括目标规划(target planning)、法规管制(legal regulatory)、金融支持(financial support/taxation)、政策支持(political support)、产权保护(property rights protection)五个子类。其中,法规管制是指政府通过制定法规、制度等强制性措施,加强对相关工作的规范和监督;目标规划是指政府为实现目标通过行政手段或活动对相关工作开展作出总体规划或描述、制定具体实

施计划;金融支持是指政府通过提供必要金融支持为相关工作营造良好的政策环境;税收优惠/激励是指政府通过制定减税、补贴、延期付款等税收方面优惠,激活市场活力;政策支持是指政府通过提供必要的政策支持来为相关工作营造良好的政策环境。

需求型(demand oriented)政策工具主要包括政府采购(government procurement)、公私合作/服务外包(service outsource)、示范工程(demonstration project)、鼓励引导(guidance and encouragement)四个子类。其中,政府采购是指中央或地方政府通过财政资金向第三方企业或营利组织购买物资或服务;公私合作/服务外包是指政府与社会资本合作拉动相关工作的开展,或者直接将某项工作外包给相关企业和公司;示范工程是指政府通过建设示范项目等带动相关工作的实施和开展;鼓励引导是指政府采取措施鼓励、引导(奖励、表彰)公众相关行为的形式来带动相关工作的开展。本研究主要针对这14个政策工具子类进行自动分类,各政策工具类型含义如表1所示。

表1 政策工具类型划分标准
Table 1 The Classification Standards of Policy Tool Types

工具类型	子工具	子工具含义
供给型	人才培养	政府通过开展教育和培训,培养相关专业人才,为相关工作开展提供人才保障。
	资金支持	政府通过拨付专项资金,对相关工作开展提供资金支持
	科技支持	政府通过提供技术支持推动相关工作的开展
	公共服务	政府通过提供相应的基本配套服务,来保障工作顺利开展
	设施建设	政府通过建设和完善基础设施,为相关工作开展提供基本物质保障
环境型	法规管制	政府通过制定法规、制度等强制性措施,加强对相关工作的规范和监督
	目标规划	政府为实现目标通过行政手段或活动对相关工作开展作出总体规划或描述、制定具体实施计划
	金融支持	政府通过提供必要金融支持为相关工作营造良好的政策环境
	税收优惠 / 激励	政府通过制定减税、补贴、延期付款等税收方面优惠,激活市场活力
	政策支持	政府通过提供必要的政策支持来为相关工作营造良好的政策环境
需求型	政府采购	中央或地方政府通过财政资金向第三方企业或营利组织购买物资或服务
	公私合作 / 服务外包	政府与社会资本合作拉动相关工作的开展
	示范工程	政府通过建设示范项目等带动相关工作的实施和开展
	鼓励引导	采取措施鼓励、引导(奖励、表彰)公众相关行为以带动相关工作开展

3 研究设计

本文研究的政策工具自动分类属于典型的多类别 (multiclass) 文本分类。多类别分类强调超过两个类别的分类任务, 例如根据目前的政策工具类型框架, 则有供给型、环境型、需求型三大类, 有人才培养、资金投入、科技投入等 14 个子类, 其假设每一个政策单元有且只有一个类型标签^[9]。相反, 多标签 (multilabel) 文本分类则强调给每个样本分配一个或多个标签, 例如一条新闻可以既属于体育类也属于文娱类, 一份政策文件可能包含多个政策工具。但是经过切割后的政策单元主要围绕一个政策工具展开, 因此对于每一个政策单元, 假设其只包含一个标签, 进行多类别文本分类, 以实现政策工具自动分类, 其研究流程如图 4 所示。

3.1 数据收集和标注

本文以国内目前相对较为权威和全面的法律法规政策库“北大法宝”为依据, 分别构建、标注数据治理和数字经济两个数据集。其一, 检索筛选出数据治理相关政策 1,097 份 (截至 2021 年 12 月 15 日), 根据其效力级别进一步选取法律、行政法规、部门规章、国务院规范性文件、部门规范性文件等核心政策 440 份, 进行政策工具编码。其二, 检索筛选出各地方数字经济政策 290 份 (截至 2021 年 12 月 15 日), 此部分对全部

政策进行编码。在政策单元切分方面, 主要依据段落、小标题以及长段落中的分号来进行, 因为如果仅仅以段落来切分政策单元, 则一个段落中极可能出现多个政策工具, 即变为多标签分类问题导致偏差。本研究中的政策工具编码工作分别来源于其他两项研究, 由主导人分别带领小组进行编码, 各编码工作平行进行, 最后由各自领域的两位专家进行校验和筛选, 以保证标注数据的质量, 经统计, 编码一致性高达 98%。最终, 分别获得包含有 2433 条、6477 条政策工具的两个较高质量的数据集^①。

3.2 模型构建

关于政策工具自动分类模型, 本文强调通过无监督文本表示学习和有监督深度学习分类的思路构建模型, 从而达到完全自动分类的效果。在无监督文本表示学习方面, 考虑到中文的特殊性, 如果使用预训练模型, 则只有基于中文预训练的才可掌握中文表达的规律, 因此本文提出基于在中文语料上训练出的 WordBERT-ZH 预训练模型。在深度学习分类算法方面, 本文提出基于 BiLSTM, 在 WordBERT-ZH 训练出的句子向量基础上, 接入全连接层进行分类, 具体政策工具自动分类模型架构如图 5 所示。

WordBERT-ZH 是 Bert 在中文语料下的一个变体。Bert (Bidirectional Encoder Representation from Transformers) 是谷歌 AI 提出的一种基于 Transformer

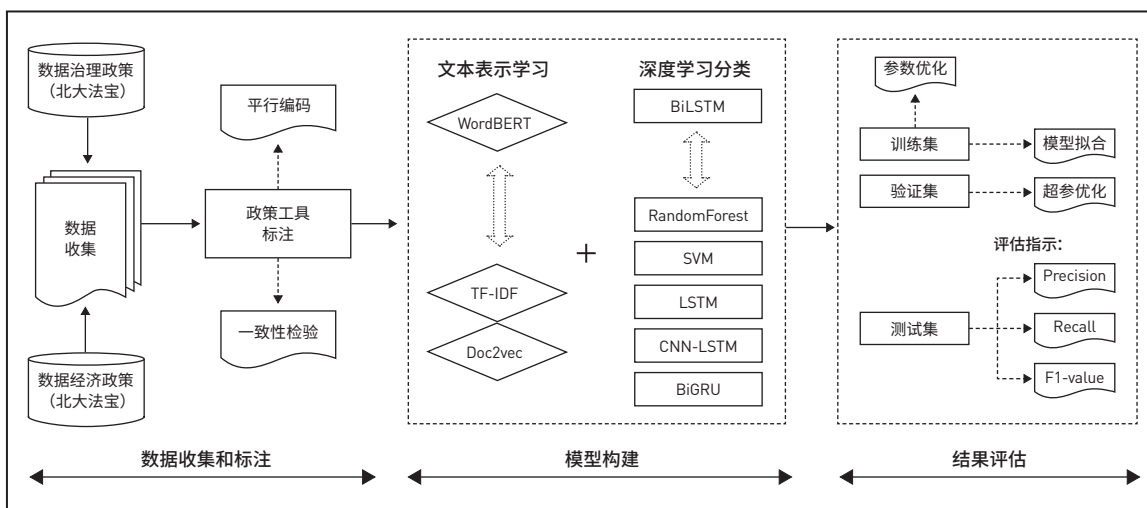


图4 基于 WordBERT 和 BiLSTM 的政策工具自动分类方法研究流程

Fig.4 The Research Process of Automatic Classification Method of Policy Tools Based on WordBERT and BiLSTM

① https://github.com/ChaoguangHuo/policy_tools_classification

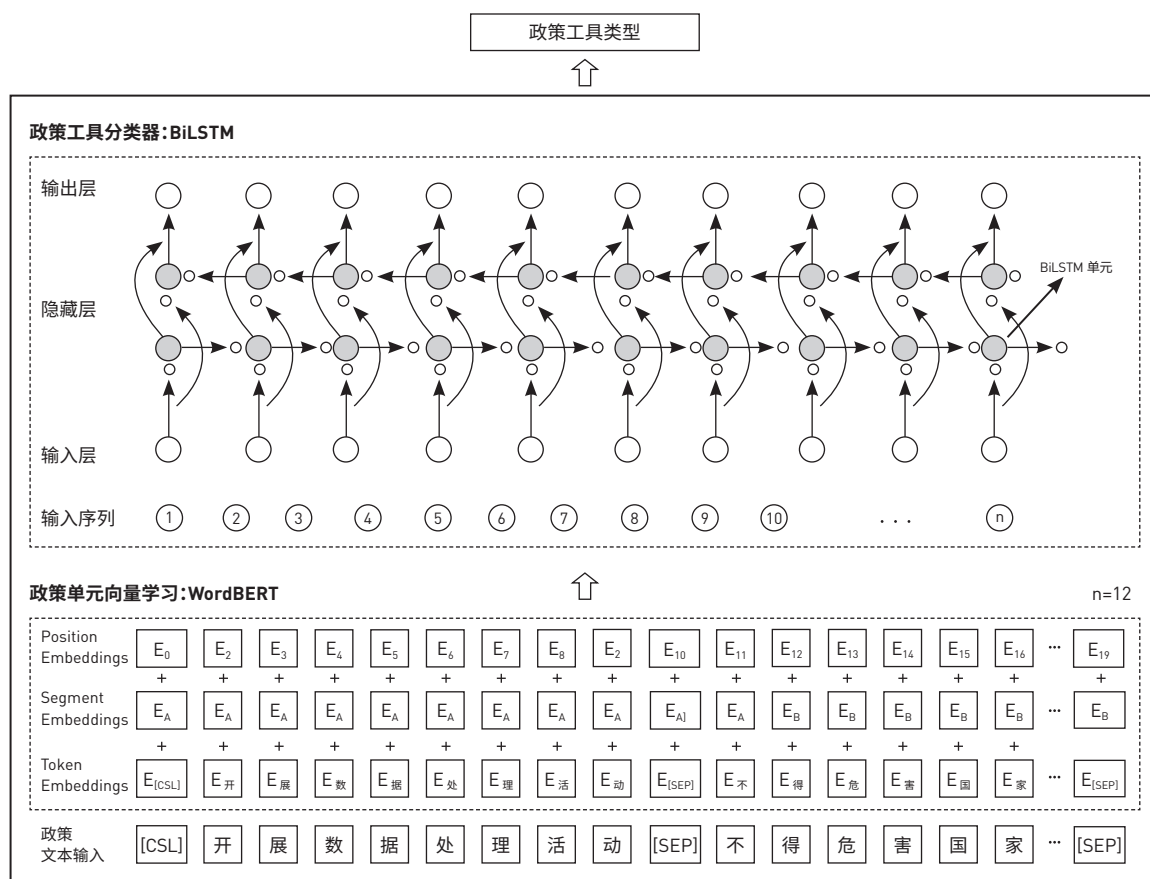


图5 基于 WordBERT 和 BiLSTM 的政策工具自动分类模型

Fig.5 The Automatic Classification Model of Policy Tools Based on WordBERT and BiLSTM

编码器的语言预训练向量表示模型,发布之初在11种自然语言处理测试任务中均取得最佳的效果,主要有BERT-Base和BERT-Large两个代表性模型^[21]。WordBERT-ZH是由Feng等人于2022年提出的一系列无需拆分单词的Bert改进版,尤其针对中文语料训练出了WordBERT-ZH预训练模型,在分类、推理以及关键字识别等任务方面均超出以往BERT-Base、RoBERTa-base、WoBERT、MarkBERT等预训练模型^[22]。基于WordBERT的政策工具分类需要以下两步,第一步,利用大规模未标注的中文政策文本语料进行自监督训练,综合词语在各种情境中的表达,充分学习政策文本的语言特征,得到文本的深层次向量表示,进而得到预训练模型,不过鉴于普通算力无法训练出此大规模的预训练模型,而学术与业界也尚无基于政策文本的预训练模型,因此本文选择基于普通文本的预训练模型WordBERT-ZH;第二步,结合具体任务进行微调,即将预训练得到的网络参数作为初始模型,输

入分类任务所标注的数据集,进一步调整部分参数,使其在所分类任务中达到较优的结果^[23]。本研究在WordBERT-ZH模型(模型参数高达3.26亿)基础上,输入政策工具标注数据集进行微调,设置样本输入最大长度500,以充分学习一个政策单元的所有语义关系信息。

BiLSTM由正向和反向LSTM组成。LSTM是深度学习递归神经网络(Recurrent Neural Network,RNN)的变体,强调将时序概念引入到神经网络中,以增强神经网络对长期时序数据的记忆能力,通过输入门、输出门、遗忘门三个门控制机制,处理长距离依赖的时间序列数据^[24]。LSTM在学习时间序列数据方面,表现较好,因此从理论上对于处理具有较长语义关系的文本应该也有不错的结果^[25]。LSTM虽然解决了长期时序数据的记忆方面的问题,但是其在处理时序数据时是单向的,对于需要双向考虑的政策单元文本序列数据可能会存在一定的不足。因此本文采纳BiLSTM算法^[26],强

调从两个方向对政策单元序列进行训练,不仅考虑到词汇之前的语义信息,还考虑到词汇之后的语义信息,即充分考虑词汇所在的上下文信息,实现时序数据的双向记忆。

针对政策工具分类研究任务,在处理政策文本时,如果依据常规停用词表直接去掉停用词^[27],势必会影响政策语义表达的完整度,例如如果直接去掉“为了”停用词,那么原本是目标规划一类的政策工具,可能就会被误识别为政策支持,甚至被识别为法规管制。“根据”“为了”“朝着”“依照”“遵照”“鉴于”“即令”“对于”“按照”等常规文本处理时的停用词,对于政策文本来说是理解政策内容的重要提示词,对于政策表达具有重要意义^[28]。由此可见常规文本处理时所用的停用词表,对于政策文本具有较大弊端,尤其是对于TF-IDF这种完全基于词频而缺乏语义的文本向量学习方法^[29],因此本文在对本文进行向量化表示时,不作停用词处理,以此保证政策工具表达的完整性,这也是本文提出基于WordBERT构建政策工具自动分类模型的主要原因。

3.3 实验设置

数据治理政策和数字经济政策是两个具有不同特征的政策工具标注数据集,其中数据治理政策工具标注数据集全部是中央相关部门出台的政策,由法律、行政法规、部门规章、国务院规范性文件、部门规范性文件等组成,其文体段落分明、内容简洁,总条目相对较少,本文共计标注有2,433个政策工具样本;数字经济政策标注数据集则全部是地方政府相关部门出台的政策,由地方规范性文件、地方工作文件、地方性法规等组成,其文本目录较多,并且层层嵌套,内容更加具体,总条目相对较多,共计有6,477个政策工具样本。因此,本文对两个数据集分别进行实验,以检验不同表示学习和机器学习分类算法对不同层面政策文本的学习和识别能力,最后再对总数据集进行实验。对于每一个数据集,本文均按照6:2:2的比例划分训练集、验证集、测试集,依此分别进行三组实验。

鉴于在政策工具分类方法方面,尚无成熟基线方法,因此本文选择将构建的政策工具自动分类模型,同其他文本表示学习和机器学习分类算法等组合形成的模型进行对比,例如TF-IDF、Doc2vec等文本表示学习算法^[30],以及RandomForest、SVM、LSTM、

CNN-LSTM、BiGRU等分类算法进行对比分析。本研究实验设备为集成服务器,最大运行内存为94G,存储硬盘8T,包含5块Nvidia Titan RTX显卡,单块GPU内存24G。其中TF-IDF平均用时较短,在几分钟内即可训练完毕,Doc2vec在几十分钟内即可训练完毕,WordBERT-ZH每次训练均在24小时以上,甚至更长时间。经过近半个月的模型调参、模型训练,最终确定较优参数设置,完成三组对比实验,相关数据以及详细代码请见GitHub。

4 结果评估

4.1 评估指标

政策工具的自动分类属于典型的多类别文本分类问题,因此以精确率(Precision)、召回率(Recall)、F1(F1-score)三个指标综合评估分类模型优劣最为直观,也最具说服力。其中准确率是指分类器对整个样本的判断能力,即正确的分类样本数与总样本数的比例;召回率是指分类器判定为正且判断正确的样本数与真实正样本数的比例;F1值是指综合精确率Precision和召回率Recall的调和平均数,各评估指标计算方法如下:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

4.2 模型评估

本文对数据治理政策数据集、数字经济政策数据集以及两个数据的合集,分别进行三组实验。在每组实验中,分别检验TF-IDF、Doc2vec、WordBERT-ZH等不同表示学习算法对不同层面和不同规模政策文本的特征学习能力,分别检验Random Forest、SVM等传统机器学习分类算法以及LSTM、BiLSTM、CNN-LSTM、BiGRU等深度学习分类算法对政策工具类型特征的拟合能力。

第一组实验,如表2所示,在数据治理政策数据集上,TF-IDF+RandomForest、TF-IDF+BiLSTM组合模型效果相对较好,其准确率分别达到64.72%和64.77%,

表2 模型分类结果评估 (数据治理政策数据集)
Table 2 The Evaluation Result of Model Classification (Data Governance Policy Dataset)

方法组合	Precision	Recall	F1-value
TF-IDF+ RandomForest	64.72%	66.39%	64.19%
TF-IDF + SVM	61.95%	65.15%	61.85%
TF-IDF + LSTM	62.91%	67.22%	64.69%
TF-IDF+ BiLSTM	64.77%	68.87%	66.25%
Doc2vec + RandomForest	41.30%	51.13%	45.28%
Doc2vec + SVM	49.44%	51.96%	49.78%
Doc2vec + LSTM	50.28%	50.93%	49.36%
Doc2vec + BiLSTM	52.45%	52.16%	50.24%
WordBERT-ZH + CNN-LSTM	58.67%	58.51%	55.71%
WordBERT-ZH + BiGRU	66.48%	66.48%	64.90%
WordBERT-ZH + BiLSTM	65.84%	65.84%	64.58%

表3 模型分类结果评估 (数字经济政策数据集)
Table.3 The Evaluation Result of Model Classification (Digital Economy Policy Dataset)

方法组合	Precision	Recall	F1-value
TF-IDF + RandomForest	60.98%	60.49%	57.97%
TF-IDF + SVM	60.88%	57.33%	54.95%
TF-IDF + LSTM	63.24%	63.43%	62.91%
TF-IDF + BiLSTM	63.59%	63.50%	62.96%
Doc2vec + RandomForest	50.64%	43.29%	39.24%
Doc2vec + SVM	50.08%	50.07%	48.75%
Doc2vec + LSTM	44.59%	46.06%	42.39%
Doc2vec + BiLSTM	45.87%	47.84%	44.96%
WordBERT-ZH + CNN-LSTM	65.63%	65.12%	63.60%
WordBERT-ZH + BiGRU	69.29%	66.32%	65.58%
WordBERT-ZH + BiLSTM	73.48%	70.55%	69.86%

召回率分别为66.39%和68.87%，当限定TF-IDF表示学习算法时，无论是传统机器学习分类算法，还是深度学习算法，其差异相对较小。本文提出的WordBERT-ZH+BiLSTM政策工具自动分类模型，虽然准确率不是最高的，但综合来看效果也相当不错。

第二组实验,如表3所示,在数字经济政策数据集上,WordBERT-ZH+BiGRU、WordBERT-ZH+BiLSTM模型明显优于其他模型,准确率分别为69.29%和73.48%,召回率分别为66.32%和70.55%。当限定BiLSTM机器学习算法时,TF-IDF和WordBERT-ZH两种表示学习算法的最终准确率相差高达10个百分点,召回率也相差7个百分点。对于此包含6477个样本的

数据集, WordBERT-ZH表示学习算法是明显优于TF-IDF、Doc2vec等算法的, 深度学习算法也愈加优于传统机器学习分类算法。由此可见只有在一定规模的数据集上, 大型预训练模型才更能发挥其对语义理解和学习的优势, 例如本文所用的WordBERT-ZH总参数为3.26亿个, 微调参数118.40万个, 只有输入足够多的数据使其充分进行参数微调, Bert系列模型才能发挥其优势。

第三组实验,如表4所示,在数据治理和数字经济政策两个数据合集上,基于WordBERT-ZH的组合效果最好,TF-IDF次之,Doc2vec较差。由此可见,TF-IDF自身虽有不足,但在政策工具自动分类方面,其模型简单,训练速度快,仍具有一定应用空间;本研究对Doc2vec向量维度训练以及各种参数调整方面,在目前参数设置的基础上,其结果相对较差,可见对政策文本进行简单向量表示的处理方式,在政策工具分类方面具有较大局限性;WordBERT-ZH与深度学习算法的组合,明显优于其他方法组合,当数据样本达到一定规模时,Bert的优势就会彰显,虽然其训练速度较慢、训练成本较高,但是其在准确率和召回率等方面的优势是显著可见的。并且随着数据样本的增多,深度学习算法较传统机器学习分类算法,也具有明显优势。综合所有实验结果可以发现,本文提出的基于WordBERT-ZH和BiLSTM的政策工具自动分类模型,明显优于其他模型组合,准确率达到73.91%,召回率达到71.29%,F1值达到71.46%,相对于其他方法,是目前最为有效的一种政策工具自动分类方法。

5 结语

政策工具是政府为实现政策目标, 将其行政理念转为切实行动的手段和方法, 是政策量化分析的重要版块。鉴于当前政策工具分析仍停留在手工分类阶段, 存在标准不统一、难以复现、规模小、成本高等一系列问题, 本文提出利用表示学习、机器学习分类等算法, 对政策文件中的政策工具进行自动分类。本文系统梳理了现有的政策工具分类框架, 在Rothwell和Zegveld政策工具分类体系基础上, 提出构建基于WordBERT-ZH和BiLSTM的政策工具自动分类模型, 并对模型进行检验。

以数据治理和数字经济政策数据集为例，经过三

表4 模型分类结果评估(数据治理和数字经济两个数据集)

Table 4 The Evaluation Result of Model Classification (Data Governance and Digital Economy Policy Dataset)

方法组合	Precision	Recall	F1-value
TF-IDF + RandomForest	64.45%	63.70%	62.27%
TF-IDF + SVM	63.96%	63.25%	62.97%
TF-IDF + LSTM	65.79%	66.27%	65.87%
TF-IDF + BiLSTM	65.82%	66.27%	65.90%
Doc2vec + RandomForest	48.22%	42.45%	40.27%
Doc2vec + SVM	48.49%	47.92%	48.10%
Doc2vec + LSTM	43.74%	52.99%	47.75%
Doc2vec + BiLSTM	48.24%	53.81%	50.03%
WordBERT-ZH + CNN-LSTM	63.74%	64.14%	63.60%
WordBERT-ZH + BiGRU	69.20%	66.69%	67.53%
WordBERT-ZH + BiLSTM	73.91%	71.29%	71.46%

组实验,本文发现对于数据治理小数据集,无论是小型TF-IDF还是大型预训练WordBERT-ZH模型,其分类结果差异较小。但是随着数据样本的增加,在数字经济政策数据集和两类政策合集时,WordBERT-ZH表示学习算法明显优于TF-IDF、Doc2vec等算法,并且

深度学习算法也愈加优于传统机器学习分类算法。由此可见,在数据达到一定规模,大型预训练模型才能更加充分进行参数微调,以更好发挥其在语义理解和学习方面的优势。在三组实验中,相较于其他组合模型,本文提出的基于WordBERT-ZH和BiLSTM的政策工具自动分类模型效果最好,准确率达到73.91%,召回率达到71.29%,F1值达到71.46%,是目前政策工具自动识别领域较为有效的一种方法。

虽然本文使用了两个数据集,但是样本量仍然是比较有限的,限制了准确率的进一步提升。目前政策工具标注数据开放程度较低,虽几经联系以往作者所标注的政策工具,但仍然很难获取到相关数据,政策工具自动分类领域研究仍需构建并开放大规模高质量的标注数据,以此才有望实现更高精度的政策工具分类,促成对海量政策文件的全局分析。此外限于目前所用预训练模型WordBERT-ZH并不是专门针对政策文本进行训练的,未来仍需要专门围绕政策文本进行预训练,只有基于更有针对性的模型和更多的标注数据,政策工具识别准确率才有望进一步得到提升。

作者贡献说明

霍朝光:研究设计,模型构建,进行实验,论文撰写;

霍帆帆:数据标注,论文修改;

王婉如:数据标注;

余芊蓉:实验支持;

杨冠灿:研究设计。

支撑数据

支撑数据可开放获取,获取地址为:https://github.com/ChaoguangHuo/policy_tools_classification。

参考文献

- [1] Jordan A,Wurzel R K W,Zito A. The Rise of 'New' Policy Instruments in Comparative Perspective:Has Governance Eclipsed Government?[J]. Political Studies,2005,53(3):477-496.
- [2] 黄新平,黄萃,苏竣.基于政策工具的我国科技金融发展政策文本量化研究[J].情报杂志,2020,39(1):130-137.(Huang Xinping,Huang Cui,Su Jun. Textual and Quantitative Research on Chinese Science and Technology Finance Development Policies Based on Policy Tools[J]. Journal of Intelligence,2020,39(1):130-137.)
- [3] Chai S,Zhang Z C,Ge J P. Evolution of Environmental Policy for China's Rare Earths:Comparing Central and Local Government Policies[J]. Resources Policy,2020,68:101786.
- [4] Ulibarri N,Ajibade I,Galappaththi E K,et al. A Global Assessment of Policy Tools to Support Climate Adaptation[J]. Climate Policy,2022,22(1):77-96.
- [5] 霍朝光,钱毅,祁天娇.基于开放公文的新冠肺炎政策知识图谱构建与分析[J].档案学通讯,2021(2):53-62.(Huo Chaoguang,Qian Yi,Qi Tianjiao. The Construction and Analysis of Epidemic Prevention Policy Knowledge Graph Based on Open Administrative Documents[J]. Archives Science Bulletin,2021(2):53-62.)
- [6] 周付军,胡春艳.政策工具视角下“双一流”政策工具选择研究——基于政策工具和建设要素双维度的分析[J].教育学报,2019,15(3):84-93.(Zhou Fujun,Hu Chunyan. Research on the Choice of Double First-Class Initiative Policy Tools from the Perspective of Policy Tools:Based on

