高技术通讯 2023 年 第33 卷 第7 期:692-703

doi:10.3772/j.issn.1002-0470.2023.07.003

基于深度学习的自然资源政策文本分类研究①

胡容波②****　郭　诚****　王锦浩****　方金云③*

(*中国科学院计算技术研究所　北京 100190)

(**自然资源部信息中心　北京 100036)

(***中国科学院大学　北京 100190)

摘　要　政策文本分类是一项涉及自然语言处理(NLP)、机器学习、政策解析等多领域的综合性技术,在政策管理、研究以及信息服务等方面有重要应用。首先,针对目前政策文本领域公共资源较少的问题,提出结合领域知识和 NLP 构建政策文本分类数据集的半自动化方法,构建了句子级自然资源政策文本分类数据集;其次,挖掘政策文本自身特点,提出基于深度学习的标题信息自适应增强政策文本分类方法,并在现有主流深度学习模型上进行扩展应用;最后,在自然资源政策文本分类数据集上的实验表明,应用该方法后,5 个常用深度学习分类模型的准确率获得了 3% 以上提升,宏平均 F_1 值获得了 5% 以上提升。

关键词　政策文本; 文本分类; 深度学习; 自然资源; 延迟决策; 数据集构建

0　引　言

近年来,各领域的政策法规都在不断增长和完善,在现代化治理体系中发挥着越来越重要的作用。政策法规大多是用自然语言表示的文本文件,本文将其简称为政策文本。对政策文本进行人工处理需要丰富的专业知识,时间成本、人力成本高昂并且容易出错[1]。为了实现对自然资源政策文本的高效管理与应用,本文提出句子级自然资源政策文本自动分类方法。

政策文本分类是一个新兴的自然语言处理(natural language processing,NLP)任务,具有重要应用价值。比如,对政策文本句子中的相关措施进行分类,可以评估、监测和改善政策[2];对政策文本句子的阅读难度进行分类,可以改进立法[3];对政策文本的业务领域进行分类,可以实现更加智能的法规语义检索和推荐[4];对政策文本中的义务性、禁止性、许可性等条款进行分类,可以辅助合规性审查[5];对政策文本包含的政策元素进行分类,可用于法规知识建模和信息系统需求工程[6]。

然而,由于自然语言具有抽象性、组合性、歧义性、进化性等特点[7],而政策法规中又有复杂的概念、规则、原则等要素,对政策文本进行自动处理仍是一件具有挑战性的工作。多年来,研究者们已经开发出了基于规则[8]、基于传统机器学习[3-4]以及基于深度学习[2,5]的各种政策文本分类方法。其中,基于深度学习的方法具有端到端学习、分类精度高等优点,目前已成为主流方法。然而,深度学习要取得较好效果离不开大规模有标签数据集的支撑[9]。目前在政策文本分类领域,数据集等公共资源有限。为此,本文采用半自动化方法,结合 NLP 和领域知识构建了句子级自然资源政策文本分类数据集。

已有的政策文本分类方法主要是将通用文本分类方法迁移应用到政策文本分类任务上,忽略了对政策文本自身特点的挖掘和利用。以自然资源政策

① 北京科技攻关(A201908230146)和河北省重点研发计划(20310106D)资助项目。

② 男,1979 年生,博士生;研究方向:智能搜索技术;E-mail:hurongbo@sina.com。

③ 通信作者,E-mail:fangjy@ict.ac.cn。

(收稿日期:2022-02-16)

法规为例,政策文本具有以下特点。(1)大部分政策文本都具有非常明确的业务特征。比如“土地开发、保护、建设活动应当坚持规划先行”中,“土地开发”提供了比较明确的业务特征信息。(2)部分政策文本并不包含具有明确业务指向的信息,包括没有业务特征信息或特征信息可指向多个业务类别。(3)随着管理职能的整合以及综合施策逐渐成为常态,在同一份政策文件中,有时会包含多个业务类别的文本。此外,在政策法规篇章级别都有文件标题,文件标题大致规定了政策法规在篇章级别的主题。

对于政策文本的特点(1),采用深度学习模型就可以取得较好的分类效果;对于特点(2),可以考虑引入文件标题信息进行辅助分类;对于特点(3),引入标题信息有利有弊,如果全部增加标题信息,当政策文本业务类别与标题业务类别不一致时反而会引入噪声。因此,为了提高模型的整体分类性能,需要设计灵活的算法以合理利用标题信息。

受文献[10]启发,本文提出基于深度学习的标题信息自适应增强(title adaptive enhancement, TAE)政策文本分类方法。TAE 以常见的深度学习网络为基石,构建孪生网络结构,在推理阶段以自适应方式引入标题信息以增强政策文本表示,进而提高分类精度。在自然资源政策文本分类数据集上的实验结果表明,增加 TAE 方法后,5 个常用深度学习分类模型的准确率和宏平均 F_1 值分别获得了 3% 和 5% 以上的提升。

本文的主要贡献总结为以下 3 点。

(1)提出结合 NLP 和领域知识的政策文本分类数据集半自动化构建方法,并构建了句子级自然资源政策文本分类数据集。

(2)提出基于深度学习的 TAE 政策文本分类方法,并构建了基于该方法的自然资源政策文本分类模型。

(3)在自然资源政策文本分类数据集上进行了广泛实验,各基线模型在增加 TAE 方法后,分类结果指标均获得明显提升。

1 相关工作

本节详细阐述与本文工作相关的历史工作,包括政策文本数据集构建、政策文本分类方法以及三向决策分类方法。

1.1 政策文本分类数据集

目前政策文本分类公开数据集较少,描述相关数据集构建过程的文献也不多。

文献[2]从气候观测组织获取了 165 份 html 格式的世界各国自主贡献英文文档,构建了各国气候政策文本数据集(英文)。该文采用半自动化方法标注数据,先由领域专家根据文档内容设定 11 个主题,再利用文档中的嵌套标题、子标题和表结构为句子生成弱标签,最后根据专家定义的业务主题进行标签映射。本文也采用半自动化方法构建数据集,但本文所获取的文档中并无可利用的句子级标签结构。

文献[11]构建了法律数据集(希腊语)。该文从希腊内政部管理的法律数据库与管理服务门户获得数据,包括 47 卷、389 章、2285 专题,共 47 563 篇文档。数据集由文档内容及其主题信息、发布年份、文档类型构成,均直接从原始文档提取,数据标注相对容易。

在中文领域,文献[12]从中国政府网的政策文件库获得数据,选取文本数量较多的前 6 个类别的 5292 条政策文本进行实验,但政策文本为篇章级。本文构建的是句子级政策文本数据集,难以直接从政策文件库中提取类别标签。

1.2 政策文本分类方法

政策文本分类技术可分为基于规则的方法、基于传统机器学习的方法和基于深度学习的方法。

文献[8]采用模式匹配(基于规则)的方法对荷兰法律法规进行分类,共建立了 88 个模式,对 592 个荷兰法律句子进行分类。基于规则的分类方法需要人工建立匹配模式,模式不足或模式过宽都容易导致分类出错,且模型的泛化能力有限。

传统机器学习方法是一种浅层学习方法,在准确性和稳定性方面比基于规则的方法具有明显优势[1]。一些研究者提出将朴素贝叶斯(naive Bayes, NB)、支持向量机(support vector machine, SVM)、逻辑回归(logistic regression, LR)等传统机器学习算法应用于金融法规分类[4]、博彩业法规分类[13]、农业

法规分类[14]等。传统机器学习方法需要进行繁琐的特征工程,且有效性受到特征提取的限制。

与传统机器学习方法相比,基于卷积神经网络(convolutional neural networks, CNN)、循环神经网络(recurrent neural network, RNN)等结构的深度学习模型可以自动进行特征提取,且文本分类性能较高,被应用于政策文本篇章[12]、合同条款句子[15]、建筑法规句子[5]等分类中。近年来,预训练语言模型(pre-trained language model,PLM)在 NLP 上的应用取得突破性进展,基于转换器的双向编码表征(bidirectional encoder representations from transformers, BERT)[16]微调已成为常见的政策文本分类应用范式[2,11]。

然而,这些方法只是将通用文本分类方法迁移应用到政策文本分类领域,缺乏对政策文本自身特征的利用。文献[17]提出将政策文件的标题和内容按权重 0.7 和 0.3 合并后进行分类,未考虑不同情况下引入标题信息的适应性。

1.3 三向决策分类方法

传统文本分类方法通常只判断待分类文本是否属于某一类别,非黑即白,对区分度小(不确定性高)的样本容易产生误判。三向决策(three-way decisions)[18]将决策区域划分为 3 个不相交的区域,包括接受决策区域、延迟决策区域和拒绝决策区域。如果有足够的信息,可直接决策,即接受或拒绝。否则,可以选择延迟决策,等待更多信息来执行二次分类。

文献[19]提出了一种三向增强卷积神经网络模型 3W-CNN,利用 NB-SVM 作为增强模型对置信度较弱的预测进行延迟决策,提高了情绪分类的准确率。文献[20]将该方法应用于中小企业管理政策文本分类中。该方法分为 2 个阶段。第 1 阶段采用 CNN 模型对政策文本进行分类,对于容易区分的样本直接输出分类结果。第 2 阶段采用传统机器学习方法对难以区分的样本进行二次分类。由于第 2 阶段的传统机器学习方法承担了增加信息并延迟决策的任务,因此其性能提升受限于特征工程的有效性。

2 自然资源政策文本分类数据集构建

本节详细阐述了句子级自然资源政策文本分类数据集的构建方法,包括数据来源、数据基本处理以及数据标注等。

2.1 数据来源

数据来源为自然资源部门户网站的政策法规库专栏[21]。该栏目包括与自然资源管理相关的法律、司法解释、行政法规、部门规章、部门规范性文件及部门其他文件等。栏目提供基本的篇章级业务分类,包括综合管理、土地管理、自然资源确权登记等 8 个业务类别。

2.2 数据获取和基本处理

(1)从政策法规库专栏获取自然资源政策法规文件,共 1722 份,大部分为 html 格式。应用 lxml 库的 etree 模块解析 html 文件,获取文件内容、标题以及文件篇章级业务类别信息。根据 html 标签将文件内容自动分段。1 份政策法规为 1 个 json 文件,数据结构如下所示。

```
{
"title":"文件标题",
"label":"文件业务类别",
"content":[
    {"paragraph":"段落 1"},
    {"paragraph":"段落 2"},
    …,
    {"paragraph":"段落 n"}
  ]
}
```

(2)删除重复文件。

(3)对篇章级误分类文件进行人工调整。

(4)对段落进行分句,删除小于 10 个字的句子,删除文件抬头、文件落款等,按业务类别将单个文件合并形成 8 个 json 文件,数据结构为

{"sentence":"句子","title":"句子所属文件标题","label":"类别标签"}

其中,类别标签为缺省的篇章级业务类别,后续将根据实际内容进行调整。8 个 json 文件共有 63 358 个政

策文本句子,字数少于 128 的句子有 59 819 个,占 94.41%。句子长度分布如图 1 所示。

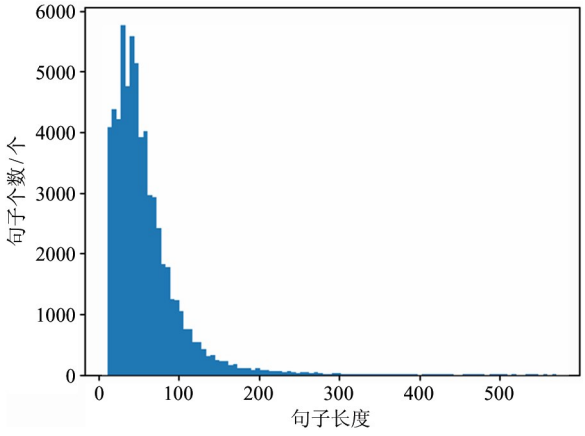

图 1　自然资源政策文本句子长度统计

2.3　数据标注

本文采用半自动化方法进行句子级政策文本数据标注,主要流程如图 2 所示。

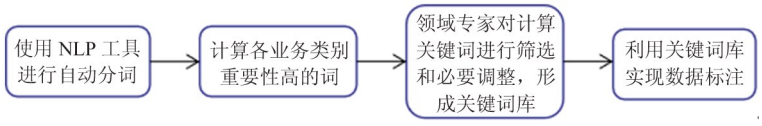

图 2　自然资源政策文本数据标注流程

(1)对 7 个业务类别(不含综合管理)的政策文本句子按业务类别合并,分别作为 7 个业务类别的语料,使用 jieba 库进行分词,分别计算 7 个业务类别去掉停用词后的词频(term frequency,TF)和逆文档频率(inverse document frequency,IDF)。计算公式为

$$tf_i = N(t_i, d_j) \tag{1}$$

式中,t_i 表示词 i,d_j 表示业务类别 j 的语料,tf_i 表示 t_i 在 d_j 中出现的次数。tf_i 越高代表 t_i 对该业务类别的重要性越大。

$$idf_i = \log \frac{|D|}{df_i} \tag{2}$$

式中,df_i 表示 7 个业务类别的语料中包含 t_i 的语料个数,最高为 7,df_i 越高,其包含的分类有效信息越低。$|D|$ 为 7,表示共有 7 个业务类别语料。idf_i 越高代表 t_i 对业务类别的区分度越大。

(2)根据计算结果,选择 TF 和 IDF 都高的词,根据领域知识进行筛选和必要调整,构建 7 个业务类别的关键词库,如表 1 所示(限于篇幅,未全部列出)。

表 1　自然资源政策文本标注关键词库

业务类别	关键词
土地管理	土地利用,土地使用,土地转用,土地储备,土地整治,土地复垦,土地市场,土地开发,土地出让,土地转让……
自然资源确权登记	登记簿,不动产,确权,权籍,使用权登记,所有权登记,抵押权登记,自然资源登记,土地权属,土地权利证书……
地质	地质调查,地质勘查,地质勘探,地勘,地质找矿,地质资料,地质钻孔,地质图,地质科技,地质服务……
地质环境管理	地质环境,地质灾害,古生物化石,地质遗迹,地质公园,矿山公园,矿山生态修复,地面沉降,滑坡,地面塌陷……
矿产资源管理	矿产资源,勘查开采,勘查开发,勘探开发,矿产勘查,矿产开采,矿业,矿权,采矿,探矿,划定矿区,储量……
海洋管理	海洋,海岛,填海,用海,用岛,深海,海标,海域使用,海域评估,海底区域,海底地形,海底电缆,领海基点……
测绘地理信息管理	测绘,地理信息,地图,国家版图,地理国情,定线测量,大地测量,测量标志,航空摄影,基准站,审图……

(3)利用关键词库对句子级文本进行重新标注。标注规则为:如果仅匹配到一个类别,直接标注为该类别;匹配到多个类别的,标注为综合管理;未匹配到的,按缺省业务类别(即篇章级业务类别)标注。

对标注结果重新按业务类别合并,形成 8 个业务类别的标注数据。统计信息如图 3 所示。

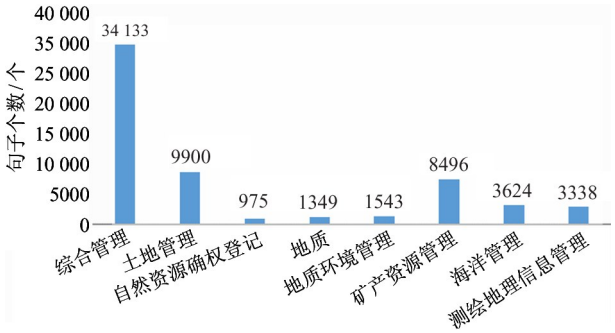

图 3　自然资源政策文本数据集分类统计

3　模型与方法

TAE 方法主要应用于模型推理阶段，通过深度学习网络获得政策文本表示后，根据分类概率的不确定性以自适应的方式选择是否引入标题信息以增强文本表示，进而提升最终分类精度。

3.1　任务定义

本研究任务形式化定义为：对于输入的政策文本句子 $x = (x_1, x_2, \cdots, x_L)$，预测其业务类别 $y \in Y$。为避免相同标题同时参与训练、验证和测试，仅在模型推理阶段可以使用政策文本所属标题信息 $t = (t_1, t_2, \cdots, t_M)$。其中，$L$ 为政策文本句子长度，M 为标题文本长度，Y 为类别标签集合。

3.2　模型总体架构

图 4 为 TAE 方法的整体框架。该方法以深度学习网络（如 CNNs、RNNs、Transformers 等）为模型基石，使用深度学习网络作为编码器来获得政策文本表示（representation）以及标题文本表示，使用 Softmax 分类器进行分类。

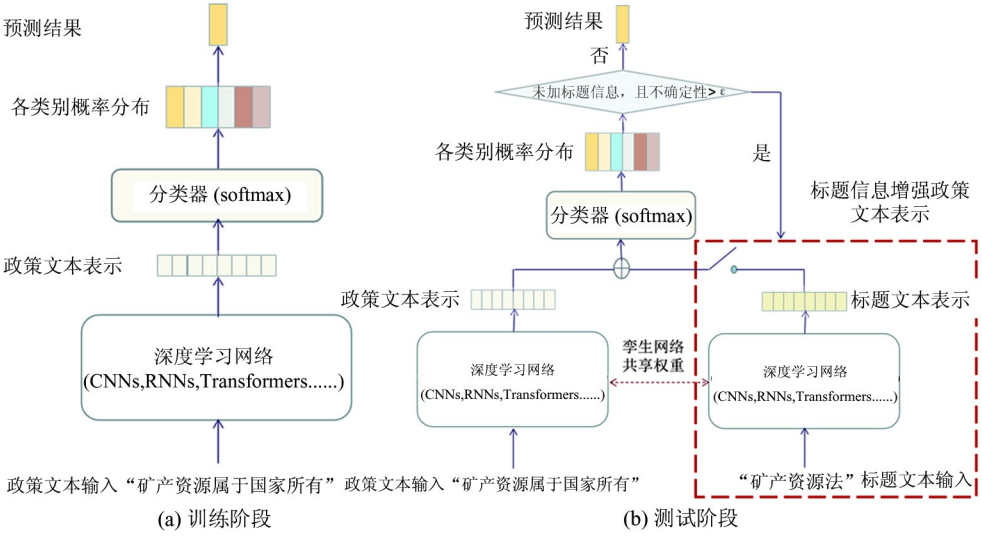

图 4　TAE 政策文本分类方法整体框架

训练阶段采用深度学习网络 + 分类器模型进行训练。推理阶段采用孪生网络结构，按照三向决策方法对分类不确定性超过阈值的政策文本进行延迟决策，借助标题信息增强政策文本表示后进行二次分类。

3.3　模型训练

给定政策文本 x，经过深度学习网络编码后，映射为政策文本表示向量 $\boldsymbol{h}_x \in \Re^d$，其中 d 为深度学习网络输出的隐向量维度。

$$\boldsymbol{h}_x = f(x) \tag{3}$$

将 $\boldsymbol{h}_x$ 送入分类器分类。分类器由一个全连接层构成，用于将 d 维向量映射到 N 维，N 是业务类别个数。对映射结果再进行 Softmax 计算得到预测概率：

$$\boldsymbol{p} = Softmax(\boldsymbol{W}\boldsymbol{h}_x + \boldsymbol{b}) \tag{4}$$

其中，$\boldsymbol{p}$ 是一个概率向量，表示模型对政策文本 x 在各个类别上的预测概率。$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别为全连接层的权重矩阵和偏置项。

以交叉熵损失作为模型优化的目标函数：

$$CE = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{N} y_j^{(i)} \log p_j^{(i)} \tag{5}$$

式中，m 为样本个数，N 为类别个数；$y_j^{(i)}$ 表示第 i

个样本在第 j 类上的真实结果,属于该类为 1,否则为 0;$p_j^{(i)}$ 表示模型对第 i 个样本属于第 j 类的预测概率。

3.4 模型测试

对于文本分类来说,分类器输出的概率分布在一定程度上也显示模型对该样本分类预测的确定性。比如概率分布[0.7,0.1,0.1,0.1]显然比[0.4,0.3,0.2,0.1]的不确定性低。文献[10]指出预测概率的不确定性越低,预测结果的准确性越高。模型测试时,通过式(4)计算出预测概率后并不立即输出预测类别,而是先对该预测概率的不确定性进行计算。基于三向决策分类方法,对于不确定性低的直接输出预测类别;对于不确定性超过阈值的进行延迟决策,即借助标题信息增强政策文本特征表示后再次分类。

文献[10]针对 BERT 多层 Transformers 编码器计算成本高的问题,提出了一种自适应调节算法,通过计算样本在当前层预测的不确定性来决定是否将其送入下一层编码器继续处理。该算法基于熵(entropy)计算样本预测的不确定性,熵越大,随机变量的不确定性就越大。为了规范化处理,该方法将样本预测不确定性定义为预测概率的熵与均匀分布熵的比值。本文采用与文献[10]相同的计算方法量度样本预测概率的不确定性:

$$Uncertainty = \frac{\sum_{i=1}^{N} p(i)\ \log p(i)}{\log \frac{1}{N}} \tag{6}$$

式中,$p(i)$ 为模型预测政策文本属于第 i 个类别的概率,N 是类别个数。

对不确定性超过阈值的样本引入标题信息进行后处理。通过孪生网络获得政策文本所属文件的标题表示向量 $\boldsymbol{h}_t \in \Re^d$。$\boldsymbol{h}_t$ 和 $\boldsymbol{h}_x$ 维度相同,均为 d 维向量。

$$\boldsymbol{h}_t = f'(x) \tag{7}$$

将 $\boldsymbol{h}_x$ 和 $\boldsymbol{h}_t$ 进行相加融合,得到标题信息增强后的政策文本表示 $\boldsymbol{h}_f$:

$$\boldsymbol{h}_f = \boldsymbol{h}_x + \boldsymbol{h}_t \tag{8}$$

将 $\boldsymbol{h}_f$ 送入分类器进行二次分类,获得新的分类概率:

$$\boldsymbol{p}' = \mathrm{Softmax}(\boldsymbol{W}\boldsymbol{h}_f + \boldsymbol{b}) \tag{9}$$

取概率最大值所对应的业务类别作为新的预测类别:

$$\tilde{y}_i = \arg\max_{y}(\{p'_i,\ \forall y \in Y\}) \tag{10}$$

式中,$\tilde{y}_i$ 为模型对第 i 个样本的最终预测类别,p'_i 为第 i 个样本的预测概率分布。

算法 1 给出了政策文本分类 TAE 算法伪代码。

算法 1 政策文本分类 TAE 算法

输入:政策文本句子 x,标题文本 t,不确定性阈值 ε
输出:政策文本句子所属业务类别 $\tilde{y}$
(1) 通过深度学习网络计算政策文本句子的向量表示 $\boldsymbol{h}_x$,式(3)
(2) 计算政策文本句子在各类别上的预测概率 $\boldsymbol{p}$,式(4)
(3) 计算政策文本句子预测结果的不确定性 *Uncertainty*,式(6)
(4) if *Uncertainty* > ε then
(5) 通过深度学习网络计算标题文本的向量表示 $\boldsymbol{h}_t$,式(7)
(6) 计算标题信息增强的政策文本表示 $\boldsymbol{h}_f$,式(8)
(7) 根据新的政策文本表示计算预测概率 $\boldsymbol{p}'$,式(9)
(8) else
(9) $\boldsymbol{p}' = \boldsymbol{p}$
(10) end if
(11) 计算预测类别 $\tilde{y}$,式(10)

4 实验与结果分析

本节详细介绍对 TAE 方法的评估实验,并给出相关分析。

4.1 实验数据集

数据集的不同划分直接影响模型的最终性能[22]。本文对第 2 节所形成数据集中的每一业务类别样本按 8:1:1 的比例进行划分,组合成训练集、验证集和测试集。随机划分 10 次,形成 10 组数据集。对所有模型,在这 10 组数据集上进行 10 次评估。

4.2 对比模型

本文选择在文本分类任务中广泛应用的深度学习模型作为基线模型,并在其基础上增加 TAE 方法进行对比分析。

(1) TextCNN[23] 模型基于 CNN 及 word2vec 对

句子级文本进行分类,擅长提取句子中的 n 元语法作为关键信息,在短文本领域应用广泛,但长距离建模能力有限,且对语序不敏感。

(2)TextRNN[24] 模使用 RNN 对电影评论文本进行分类。RNN 及其变体擅长捕获文本序列信息,其递归结构非常适合处理变长文本,是 NLP 任务中最常用的结构之一。

(3)TextRCNN[25] 模型使用循环卷积神经网络(recurrent convolutional neural networks,RCNN)对句子级及文档级文本进行分类,可以有效捕捉上下文信息。

(4)DPCNN[26] 模型使用深度金字塔卷积神经网络(deep pyramid convolutional neural networks,DPCNN)进行文本分类,通过不断加深网络,可以抽取长距离的文本依赖关系。

(5)BERT[16] 模型是谷歌公司提出的 PLM,在大规模语料上采用掩码语言模型(masked language model,MLM)、下一句预测(next sentence prediction,NSP)对双向多层 Transformer 进行预训练,能够生成深度双向语言表征。预训练后,只需要添加一个额外的输出层进行微调,就可以在包括文本分类在内的各种下游任务中取得优异性能。

4.3 评价指标

对于单个类别的分类性能,采用召回率(recall)、精确率(precision)和 F_1 值作为评价指标。

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{11}$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{12}$$

$$F_{1i} = \frac{2P_i \times R_i}{P_i + R_i} \tag{13}$$

式中,R_i、P_i 和 F_{1i} 分别表示第 i 类的召回率、精确率和 F_1 值,TP_i、FP_i 和 FN_i 分别表示模型预测的第 i 类真正例、假正例和假负例个数。

对于模型整体性能,采用准确率(accuracy)、宏平均 F_1 值和加权平均 F_1 值进行评价。

$$Accuracy = \frac{\sum_{i=1}^{N}(TP_i + TN_i)}{\sum_{i=1}^{C}(TP_i + FP_i + TN_i + FN_i)} \tag{14}$$

$$Macro\_F_1 = \frac{2P_{\text{marco}} \times R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}} \tag{15}$$

式中,$Accuracy$、$Macro\_F_1$ 分别表示模型的准确率和宏平均 F_1 值。$P_{\text{marco}} = \frac{\sum_{i=1}^{N} P_i}{N}$,$R_{\text{marco}} = \frac{\sum_{i=1}^{N} R_i}{N}$,$TN_i$ 表示模型预测的第 i 类真负例个数,N 为类别个数。

$$Weighted\_F_1 = \frac{2P_{\text{weighted}} \times R_{\text{weighted}}}{P_{\text{weighted}} + R_{\text{weighted}}} \tag{16}$$

式中,$Weighted\_F_1$ 表示模型的加权平均 F_1 值。

$P_{\text{weighted}} = \frac{\sum_{i=1}^{N}(P_i \times w_i)}{N}$,$R_{\text{weighted}} = \frac{\sum_{i=1}^{N}(R_i \times w_i)}{N}$,$w_i = \frac{Number_i}{Number}$,$Number_i$ 为第 i 类样本个数,$Number$ 为样本总数,N 为类别个数。

4.4 实验设置

对于 TextCNN、TextRNN、TextRCNN、DPCNN 均使用文献[27]开源的中文词向量(人民日报 Word + Character + Ngram 300d)进行初始化。TextCNN 的卷积核大小设置为 2、3、4,每个尺寸的卷积核数量为 256;DPCNN 的卷积核数量为 256;TextRNN 的 LSTM 隐藏层大小为 128,LSTM 层数为 2;TextRCNN 的 LSTM 隐藏层大小为 256,LSTM 层数为 1。以上模型均选择 Adam 作为优化器,learning_rate 为 0.001,pad_size 为 128,batch_size 为 128,epoch 为 20。对于 BERT,使用 BERT-Base-Chinese 预训练模型,隐藏层大小为 768,dropout 为 0.1,batch-size 大小为 32,pad_size 为 128,选择 AdamW 作为优化器,learning_rate 为 0.000 05,epoch 为 3。主实验不确定性阈值 ε 取 0.2。

实验环境:操作系统为 Linux,CPU 为 12 核 Intel(R) Xeon(R) Gold 5320 CPU@2.20 GHz,内存为 32 GB,GPU 为 1 块 RTX A4000,显存为 16 GB。

4.5 实验结果及与基线模型对比

本文报告了 TAE 和其他基线方法在 10 组随机划分的自然资源政策文本分类数据集上的详细测试性能以及 TAE 方法相对基线模型的性能提升(见表 2)。表中各模型的准确率、宏平均 F_1 值、加权平均 F_1 值为各模型 10 次评估的平均值 ± 标准差,粗体字表示每组内的较好结果。可以得出如下结论。

表 2　TAE 方法与基线模型的对比实验结果

方法	准确率	宏平均 F_1 值	加权平均 F_1 值
TextCNN	0.9196 ±0.0052	0.8893 ±0.0069	0.9188 ±0.0052
TextCNN + TAE	**0.9584 ±0.0024**	**0.9448 ±0.0063**	**0.9583 ±0.0025**
提升	3.88%	5.55%	3.95%
TextRNN	0.9039 ±0.0095	0.8656 ±0.0165	0.9031 ±0.0094
TextRNN + TAE	**0.9458 ±0.0090**	**0.9241 ±0.0142**	**0.9457 ±0.0090**
提升	4.19%	5.85%	4.26%
TextRCNN	0.9172 ±0.0044	0.8848 ±0.0092	0.9167 ±0.0044
TextRCNN + TAE	**0.9562 ±0.0025**	**0.9395 ±0.0052**	**0.9562 ±0.0025**
提升	3.90%	5.47%	3.95%
DPCNN	0.9127 ±0.0030	0.8798 ±0.0071	0.9120 ±0.0030
DPCNN + TAE	**0.9574 ±0.0028**	**0.9421 ±0.0055**	**0.9573 ±0.0030**
提升	4.47%	6.23%	4.53%
BERT	0.9323 ±0.0029	0.9054 ±0.0067	0.9320 ±0.0029
BERT + TAE	**0.9689 ±0.0028**	**0.9582 ±0.0046**	**0.9689 ±0.0028**
提升	3.66%	5.28%	3.69%

(1)在不使用 TAE 方法时,基于深度学习的模型对自然资源政策文本分类已具有较高性能。5 个基线模型的准确率、加权平均 F_1 值均可达到 90% 以上。其中,基于 CNN 的模型性能高于仅使用 RNN 的分类模型。这是因为在政策文本分类任务中,文本序列的重要性不及文本中 n 元语法关键信息的重要性,而后者正是 CNN 所擅长捕获的。在 5 个基线模型中,基于 BERT 的模型取得最好性能,这主要得益于其强大的语言表征能力,通过模型微调可以更好地捕获政策文本中不同类别间的细微差别,即便是模型宏平均 F_1 值也达到 90% 以上。

(2)应用 TAE 方法可以进一步提高深度学习模型的分类性能。TAE 方法在模型的准确率、宏平均 F_1 值和加权平均 F_1 值 3 个总体指标上均明显高于相应的基线模型。其中,模型宏平均 F_1 值的提升尤为突出,比 5 个基线模型分别提升 5.55%、5.85%、5.47%、6.23% 和 5.28%。值得一提的是,即便是 TextCNN、TextRNN、TextRCNN、DPCNN,仅仅增加 TAE 方法,在模型准确率、宏平均 F_1 值和加权平均 F_1 值上也优于 BERT 基线模型。

TAE 方法的有效性主要得益于以下几个方面:1)对于缺乏业务特征的政策文本句子,深度学习网络难以提取到有效特征信息,从而导致分类器得出的预测概率不确定性较高,这种情况下增加标题信息可以使分类结果倾向于标题文本所属类别,而大部分情况下标题文本所属类别都与政策文本一致,因此能提升召回率。2)对于业务特征有多个指向的政策文本句子,深度学习网络也难以学习到有效的区分特征,同样导致预测概率不确定性较高,与 1)类似,大部分情况下增加标题信息会使分类召回率受益。3)对于业务特征明显且指向单一的政策文本句子,深度学习网络通过训练一般都能提取到明确的类别特征,从而预测概率不确定性较低,这种情况下 TAE 会直接输出预测类别,避免了加入标题信息导致的精确率下降。

当然,不同深度学习网络各有优点和局限,对文本特征的提取能力并不相同,预测概率的不确定性各异,因此在增加 TAE 方法后,不同深度学习网络的提升幅度并不相同。另外,宏平均 F_1 值提升更加明显,说明少样本类别在 TAE 方法中受益较大,原因是深度学习网络对少样本特征的提取更具挑战性,从而标题信息可以发挥更大的辅助作用。

4.6　不确定性阈值的影响

为了探索不确定性阈值 ε 对模型性能的影响,基于 BERT + TAE 模型,以 0.1 为间隔,对 0 ~ 1 之间的参数进行了实验。其中当 $\varepsilon = 0$ 时,所有测试

的政策文本都会添加标题信息；当 ε =1 时，所有测试的政策文本都不会添加标题信息（单句子分类）。图 5 显示了不同 ε 对模型准确率的影响。图 6 显示了模型取不同 ε 时分类错误数的变化。

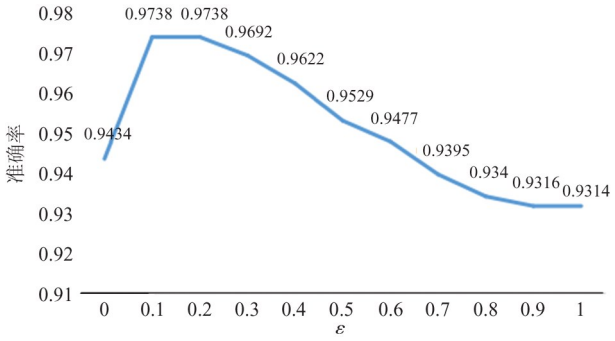

图 5　BERT + TAE 模型取不同 ε 时的准确率变化

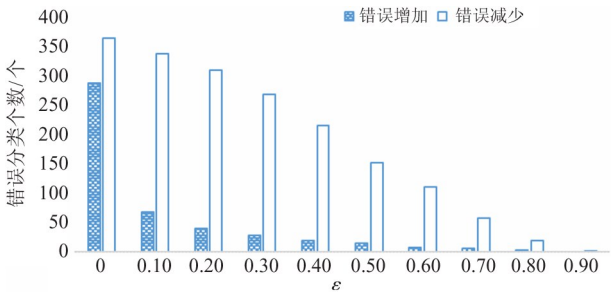

图 6　BERT + TAE 模型取不同 ε 时的错误增减变化

实验结果表明，ε =0 时，模型分类准确率仅比单句子分类时提升 1.2%。这是因为与单句子分类相比，全部增加标题信息后，虽然有 362 个单句子时分类错误的样本被正确分类，但是同时有 286 个单句子时分类正确的样本被错误分类，可见增加标题信息带来的大部分收益（新的正确分类数）被损失（新的错误分类数）所抵消，从而导致分类性能提升有限。这验证了增加标题信息对政策文本分类有利有弊。从图 6 可以看出，随着 ε 由 0 变大，增加标题导致的损失迅速下降，在 ε =0.1 之后，降幅趋于平稳，直到 ε =0.9 时损失为 0。而随着 ε 由 0 增大，增加标题带来的收益只是平稳下降，直到 ε =0.9 时仍有 1 个新的正确分类样本。从而，不同 ε 带来了不同的收益和损失差异，最终带来了模型性能的不同提升。因此，对 ε 进行更精细调参，还可获得更高性能。

不同 ε 直接决定需要增加标题进行延迟决策样本个数（图 7）。当 ε =0 时，6330 个测试样本全部需要延迟决策；分类收益为 362 个，仅占延迟决策样本数的 5.71%。当 ε =0.2 时，需要延迟决策样本数为 799 个，分类收益为 307 个，占延迟决策样本数的 38.42%。在 BERT 模型中，81.90% 的测试样本预测概率的不确定性介于 0 ~ 0.1 之间。这也验证了大部分政策文本句子预测概率的不确定性较低，无需延迟决策，而对少部分不确定性较高的样本进行延迟决策收益占比较高。

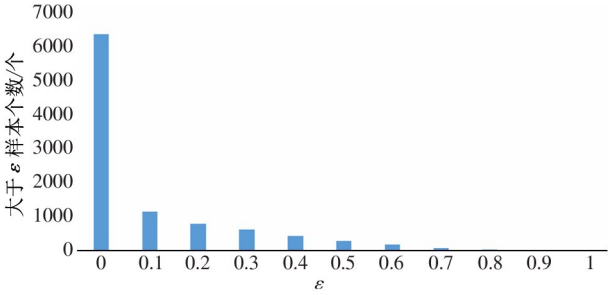

图 7　BERT + TAE 模型取不同 ε 时的延迟决策样本个数

4.7　TAE 对不同类别的影响

TAE 对不同类别样本的影响并不均衡。表 3 记录了分别使用 BERT 模型和 BERT + TAE 模型（ε =0.2）时，8 个业务类别的精确率、召回率和 F_1 值的变化。其中较好结果使用粗体字突出显示。

从表 3 可以看出，使用 TAE 方法后，除测绘地理信息管理类别的精确率降低外，其余类别的性能指标均获得提升。在精确率方面，提升幅度较大的是矿产资源管理和地质环境管理，分别达 8.99% 和 7.36%。测绘信息管理类别精确率下降原因是，真正例 TP 个数虽然增加了 6.23%，但假正例 FP 个数却增加了 75%。在召回率方面，性能提升较大的是地质和海洋管理类别，分别达 14.17% 和 8.29%。总体来看，TAE 方法对地质、矿产资源管理、海洋管理等类别的效果更为显著，F_1 值提升在 6% 以上。说明这些类别的政策文本更符合本文对政策文本特点的基本假设，即大部分政策文本的业务特征明显，部分业务特征不明显的可借助文件标题辅助分类，引入文件标题后带来的噪声有限。

4.8　案例分析

从前文可以看出，TAE方法既有收益也有损

表 3　TAE 方法与基线方法在具体类别上的分类性能对比

	BERT			BERT + TAE					
	精确率	召回率	F_1 值	精确率	提升	召回率	提升	F_1 值	提升
综合管理	0.9408	0.9637	0.9521	**0.9761**	3.53%	**0.9810**	1.73%	**0.9785**	2.64%
土地管理	0.9365	0.9090	0.9225	**0.9796**	4.31%	**0.9707**	6.17%	**0.9751**	5.26%
自然资源确权登记	0.9247	0.8866	0.9053	**0.9684**	4.37%	**0.9485**	6.19%	**0.9583**	5.30%
地质	0.8957	0.7687	0.8273	**0.9173**	2.16%	**0.9104**	14.17%	**0.9139**	8.66%
地质环境管理	0.8882	0.9346	0.9108	**0.9618**	7.36%	**0.9869**	5.23%	**0.9742**	6.34%
矿产资源管理	0.8921	0.8857	0.8889	**0.9820**	8.99%	**0.9635**	7.78%	**0.9727**	8.38%
海洋管理	0.9246	0.8812	0.9024	**0.9776**	5.30%	**0.9641**	8.29%	**0.9708**	6.84%
测绘地理信息管理	**0.9621**	0.9159	0.9385	0.9391	−2.30%	**0.9730**	5.71%	**0.9558**	1.73%

失,不同的 ε 设置即是为了取得收益与损失的最佳平衡。本节给出了实验中的几个具体案例。

(1)因为 TAE 方法而正确分类的案例。一是政策文本中没有明确业务特征的样本。如“建立动态巡查责任追究制度,对巡查工作不到位、报告不及时、制止不得力的要追究有关责任人的责任。”,BERT 模型将其误分类为综合管理。在 BERT + TAE 中(ε =0.2,下同),加入标题《国土资源部关于进一步完善农村宅基地管理制度切实维护农民权益的通知》信息后,被正确分类为土地管理。二是政策文本中有业务特征,但可指向多个业务类别的样本。如“采用招标或拍卖方式的,取得投标或竞买资格者不得少于 3 个。”,BERT 模型将其误分类为矿产资源管理,加入标题《国土资源部关于印发 <招标拍卖挂牌出让国有土地使用权规范 >(试行)和 <协议出让国有土地使用权规范 >(试行)的通知》信息后,BERT + TAE 将其正确分类为土地管理。

(2)因为 TAE 方法而错误分类的案例。一是标题信息的业务特征也不明确的样本。如“土地矿产卫片执法检查工作机构通过内业判别和实地……在与矿产资源规划、探矿权、采矿权数据综合对比分析的基础上,初步判定矿产资源勘查开采疑似违法图斑。”,BERT 模型将其正确分类为矿产资源管理。BERT + TAE 加入标题《土地矿产卫片执法检查工作规范(试行)》信息后,反而被误分类为综合管理。二是标题信息的业务特征明确,但与政策文本类别不一致的样本。如“行政复议应诉机构负责为诉讼代理人办理授权委托书等事宜。”,BERT 模型将其正确分类为综合管理。BERT + TAE 加入标题《关于印发 <国家测绘局行政复议和行政应诉办法 > 的通知》信息后,反而被误分类为测绘地理信息管理。

5　结　论

本文提出的结合 NLP 和领域知识的方法可以有效构建句子级自然资源政策文本分类数据集,提出的基于深度学习的 TAE 政策文本分类方法可以灵活利用政策文本自身特点,进一步提升政策文本分类性能。实验结果表明,5 个基于 CNNs、RNNs、Transformers 的常用深度学习分类模型增加 TAE 方法后,模型的准确率、宏平均 F_1 值、加权平均 F_1 值都获得了显著提升。该方法可在类似政策文本分类中应用,也可应用于政策文本大数据分析中。

参考文献

[1] LI Q, PENG H, LI J, et al. A survey on text classification: from shallow to deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020,31(11):1-21.

[2] CORRINGHAM T, SPOKOYNY D, XIAO E, et al. BERT classification of Paris agreement climate action plans[C]//ICML 2021 Workshop on Tackling Climate Change with Machine Learning. Virtual: ICML, 2021:1-6.

[3] CURTOTTI M, MCCREATH E, BRUCE T, et al. Machine learning for readability of legislative sentences[C]//Proceedings of the 15th International Conference on Artificial Intelligence and Law. San Diego: Association for Computing Machinery, 2015:53-62.

[4] BOELLA G, DI CARO L, HUMPHREYS L, et al. NLP challenges for eunomos, a tool to build and manage legal knowledge[C]// The 8th International Conference on Language Resources and Evaluation. Istanbul: LREC, 2012:3672-3678.

[5] SONG J, LEE J K, CHOI J, et al. Deep learning-based extraction of predicate-argument structure (PAS) in building design rule sentences[J]. Journal of Computational Design and Engineering, 2020,7(5):563-576.

[6] SLEIMI A, SANNIER N, SABETZADEH M, et al. An automated framework for the extraction of semantic legal metadata from legal texts[J]. Empirical Software Engineering, 2021,26(3):1-50.

[7] 车万翔, 郭江, 崔一鸣. 自然语言处理:基于预训练模型的方法[M]. 北京:电子工业出版社,2021.

[8] MAAT E, WINKELS R. Automated classification of norms in sources of law[M]. Berlin, Heidelberg: Springer, 2010: 170-191.

[9] 王金甲,陈浩,刘青玉. 大数据下的深度学习研究[J]. 高技术通讯, 2017,27(1):27-37.

[10] LIU W, ZHOU P, ZHAO Z, et al. Fastbert: a self-distilling bert with adaptive inference time[C]// The 58th Annual Meeting of the Association for Computational Linguistics. Seattle: Association for Computational Linguistics, 2020:6035-6044.

[11] PAPALOUKAS C, CHALKIDIS I, ATHINAIOS K, et al. Multi-granular legal topic classification on Greek legislation[C]// Proceedings of the Natural Legal Language Processing Workshop 2021. Punta Cana: NLLP, 2021:63-75.

[12] 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.

[13] PRENDERGAST M D. Automated extraction and classification of slot machine requirements from gaming regulations[C]// 2021 IEEE International Systems Conference (SysCon). Vancouver: IEEE, 2021:1-6.

[14] ESPEJO-GARCIA B, MARTINEZ-GUANTER J, PÉREZ-RUIZ M, et al. Machine learning for automatic rule classification of agricultural regulations: a case study in Spain [J]. Computers and Electronics in Agriculture, 2018, 150:343-352.

[15] CHALKIDIS I, ANDROUTSOPOULOS I, MICHOS A. Obligation and prohibition extraction using hierarchical rnns[C]// The 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018:254-259.

[16] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//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] 龚增辉, 胡建敏. 一种基于机器学习的政策文本分类方法:CN 112668329A[P]. 2021-04-16

[18] YAO Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences, 2010,180(3):341-353.

[19] ZHANG Y, ZHANG Z, MIAO D, et al. Three-way enhanced convolutional neural networks for sentence-level sentiment classification[J]. Information Sciences, 2019, 477:55-64.

[20] LIANG D, YI B. Two-stage three-way enhanced technique for ensemble learning in inclusive policy text classification[J]. Information Sciences, 2021,547:271-288.

[21] 自然资源部. 政策法规库[EB/OL]. [2022-02-16]. http://f.mnr.gov.cn.

[22] GUO B, HAN S, HAN X, et al. Label confusion learning to enhance text classification models[C]// The 35th AAAI Conference on Artificial Intelligence. Virtual: AAAI Press, 2021:4335-4340.

[23] KIM Y. Convolutional neural networks for sentence classification[C]//2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014:1746-1751.

[24] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[C]// International Joint Conference on Artificial Intelligence. New York: AAAI Press, 2016:2873-2879.

[25] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[J]. Journal of Database Management, 2016,32(4):65-82.

[26] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]// The 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017:562-570.

[27] LI S, ZHAO Z, HU R, et al. Analogical reasoning on Chinese morphological and semantic relations[C]// The 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018:138-143.

Research on classification of natural resources policy text based on deep learning

HU Rongbo[*,**,***], GUO Cheng[*,***], WANG Jinhao[*,***], FANG Jinyun[*]

([*]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**]Information Center of Ministry of Natural Resources, Beijing 100036)

([***]University of Chinese Academy of Sciences, Beijing 100190)

Abstract

Policy text classification is a comprehensive technology involving natural language processing(NLP), machine learning, policy analysis and other fields, which can be applied to policy management, research, information service, etc. Firstly, aiming at the problem that there are few public datasets in the field of policy text at present, a semi-automatic method of combining domain knowledge and NLP to construct policy text classification dataset is proposed, and a sentence-level natural resource policy text classification dataset is constructed. Secondly, taking advantage of the characteristics of policy texts, a deep learning-based title adaptive enhancement policy text classification method is proposed, which is applied to the existing mainstream deep learning models. Finally, extensive experiments on the natural resource policy text classification dataset show that after adding this method, the accuracy of five commonly used deep learning classification models is improved by more than 3%, and the macro-average F_1 score is improved by more than 5%.

Key words: policy text, text classification, deep learning, natural resources, delay decision, dataset construction