文章编号：1003-0077(2022)09-0076-08

融入领域术语词典的司法舆情敏感信息识别

张泽锋[1,2]，毛存礼[1,2]，余正涛[1,2]，黄于欣[1,2]，刘奕洋[1,2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500)

摘 要：司法舆情敏感信息识别主要是从海量网络文本中识别出与司法领域相关的敏感舆情。当前，面向司法舆情敏感信息识别的研究较少，相比通用领域的敏感信息识别任务，司法舆情敏感信息具有描述不规范、冗余信息多以及领域词汇过多等特点，这使得通用模型并不适用该任务。为此，该文提出融入领域术语词典的司法舆情敏感信息识别模型。首先使用双向循环神经网络和多头注意力机制对舆情文本进行编码，得到具有权重信息的文本表示；其次将领域术语词典作为分类的指导知识，与舆情文本表征构建相似矩阵，得到融入领域术语词典的司法敏感文本表征；然后利用卷积神经网络对其进行局部信息编码，再利用多头注意力机制获取具有敏感权重的局部特征；最后实现司法领域敏感信息识别。实验结果表明，相比 Bi-LSTM Attention 基线模型，F_1 值提升了 8%。

关键词：司法舆情；敏感信息；领域术语词典；多头注意力机制

中图分类号：TP391 **文献标识码**：A

Sensitive Judicial Public Opinion Information Recognition with the Domain Terminology Dictionary

ZHANG Zefeng[1,2], MAO Cunli[1,2], YU Zhengtao[1,2], HUANG Yuxin[1,2], LIU Yiyang[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China; 2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: Currently, there are few researches on sensitive information identification for judicial public opinion, which is challenged by nonstandard descriptions, rich redundant information and numerous domain words. To address these issues, we propose a novel recognition model of judicial public opinion sensitive information via integrating the domain terminology dictionary. Firstly, the bi-directional recurrent neural network and multi-head attention mechanism are used to encode the public opinion text. Secondly, the domain terminology dictionary is used as the guiding knowledge for classification, and a similarity matrix is constructed with the public opinion text representation to derive the judicially sensitive text representation. Furthermore, convolutional neural network is used to encode local information, and multi-head attention mechanism is used to derive the weight aware local features. Finally, the identification of sensitive information in the judicial field is employed. The experimental results show that compared with the Bi-LSTM Attention baseline model, the F_1 value increases by 8%.

Keywords: judicial public opinion; sensitive information; domain terminology dictionary; multi-head attention mechanism

0 引言

在社交网络中，用户可以随时随地表达自己的观点[1]，其中针对司法审判部门的相关工作有大量误解和片面的言论，这些信息传播迅速，敏感度高，容易引发网络舆情。本文将涉及司法公职人员或特定案件的负面、不实的网络舆情定义为司法舆情敏感信息。从海量舆情新闻中快速、准确的识别这些信息，对于辅助司法部门开展工作尤为重要。

目前，针对司法领域舆情信息的分析工作已经取得一些进展。赵等人[2]将案件要素作为监督信息

收稿日期：2020-08-19 定稿日期：2020-09-13

基金项目：国家重点研发计划(2018YFC0830105，2018YFC0830101，2018YFC0830100)；云南省自然科学基金(2019FA023)；云南省中青年学术和技术带头人后备人才项目(2019HB006)；云南省高新技术产业专项(201606)

融入新闻文档与案件描述的相关性分析中，实现了案件与舆情的相关性分析。韩等人[3]利用案件要素作为外部知识，来指导涉案舆情文本摘要生成，这些研究都为开展司法舆情敏感信息识别提供了可借鉴的方法。司法舆情敏感信息识别是一种特定领域的分类任务，针对该任务，目前主要的研究方法有两类，第一类是基于人工构建的敏感词库，构造敏感信息识别规则。早期，Chow 等人[4]提出基于参考语料库的推理检测模型，其出发点是：词语贡献度是推理的基础，主要是通过预先构建的网络文本语料库提取词频高的词语表示敏感信息的特征，通过计算未知文本与已知文本的置信度，取置信度在阈值之上的文本为敏感文本。Cheng 等人[5]通过意见挖掘技术和语言处理技术，提出敏感信息特征提取算法，利用半结构化的网络文本构建敏感词典，将词典中的词语进行敏感信息特征匹配程度的计算，同时结合词性、实体识别等方法进行敏感信息的判断。Xu 等人[6]提出基于向量空间模型和余弦相似度的文本敏感信息分类方法，其主要思想是利用敏感词表映射出向量空间，通过余弦距离判断未知文本与已知词语的相似度，然后乘以词频系数突出不同位置的词语的权重，从而计算文档的相似性，进而识别敏感信息。但是这类方法依赖于人工构建的匹配规则，随着深度学习的发展，越来越多的研究者将敏感信息识别转换为基于深度学习框架的文本二分类任务，利用神经网络的文本表征能力和特征提取能力，将文本映射到高维向量空间，识别其是否为敏感信息。Neerbek 等人[7]将敏感信息中的短语结构信息表示为树结构，利用循环神经网络[8]（Recurrent Neural Network，RNN）来预测文档的敏感性，对比以往的基于词典的敏感信息识别任务，性能有了很大的提升。Xu 等人[9-10]将主题聚类融入敏感信息识别中，首先提出了一种基于加权潜在狄利克雷分布[11]（Latent Dirichlet Allocation，LDA）的网络敏感信息主题识别方法，利用 Word2Vec 技术训练敏感词汇的词嵌入表示，通过相似度计算寻找更细粒度的关键字和生成高质量的主题词，然后将得到的主题信息与经过双向循环神经网络[12]（Bi-directional Long Short-Term Memory，Bi-LSTM）表征后的文本特征向量进行融合，最后利用注意力机制[13]进行权重计算，从而实现敏感信息识别。

虽然上述方法在通用领域的敏感信息识别任务上已经取得了较好的效果，但是，针对司法领域的敏感信息识别，并不能将其看为一个简单的二分类任务，需要同时考虑是否涉及司法领域以及是否为敏感信息。如表 1 所示，针对司法领域敏感信息的识别任务会出现司法相关不敏感信息和司法无关敏感信息，其同样可以为司法部门的工作提供建议，因此本文将司法舆情敏感信息识别任务转化为一个四分类任务，需要识别敏感性和领域性。

表 1　分类类别及司法舆情数据举例

类型	内容
司法相关敏感信息	例 1：辽宁.又是辽宁。辽宁的法治，从公开资料来看，讲法讲的挺好吗。想必辽宁高院应该收到了当事人的申诉材料，为什么不给人家解决一下？别做双面人！切实的干点工作吧……；天道好轮回，苍天绕过谁！一手制造陈洪伟律师冤案的办案人员，你们真的不怕报应吗？辽宁律师反腐遭报复含冤入狱何日还清白？一个人的不公平就是全社会的不公平！恳请社会各界正义人士关注辽宁陈洪伟律师冤案。 例 2：酒肉穿肠过，法律算个球。四川高院唐法官是这么想的。【恳请佛山中院开庭审理俞崎案】俞崎案，二审律师何兵教授和李仲伟律师 已经向@佛山市中级人民法院 提交了应当开庭的律师意见书，意见书中对一审认……屏其实是复制件，开庭时即未出示手机上的原件，也没有按法定程序…全文
司法相关不敏感信息	例 3：＃共同战疫 法院在行动＃【volg｜一名书记员的独白】一场“战疫”，打开了 2020 年，这段时间，你的生活发生了什么变化？口罩成了生活的必需品，所有住宅区都围蔽起来……而法院人的工作又发生了什么……吴惠筠用 vlog 告诉你↓↓↓↓@广东省高级人民法院…全文
司法无关敏感信息	例 4：近日，…对…政客歇斯底里的表演作出的这番评价，可谓入木三分。在当前多国疫情趋缓、民众生活逐渐重启之际，全球唯一超级大国竟然还保持着每日超过 2 万病例的增长，令人扼腕叹息……
司法无关不敏感信息	例 5：首家全国证券期货纠纷专业调解组织在沪揭牌，注册资本亿元中证资本市场法律服务中心于月日在上海正式揭牌成立……场的全国性证券期货纠纷公益调解机构，注册资本亿元人民……

针对此任务，本文构建了司法舆情敏感信息识别数据集，表1为数据举例。分析表1中示例可以看出：①司法舆情文本具有描述不规范、冗余信息多等问题，导致难以对其进行有效的表征，如例3中的“【volg |”等特殊符号，以及例2中的“【恳请佛山中院开庭审理俞崎案】……，全文”等转发信息，进行表征后均为冗余特征。②涉及司法领域的敏感信息中包含导致文本敏感的短语，而这些短语没有在通用领域的敏感词典中出现，如“越权越规”“无视法律，法规”等，导致直接利用基于敏感词典的方法不能很好地适应于司法领域。为了获得更好的表征，让模型能够学习到司法敏感信息相关的表示，本文抽取司法领域术语构造领域术语词典，将术语词典作为外部指导融入深度学习框架中，进而增强敏感术语的表征[14]。

根据以上分析，本文首先构建司法领域术语词典，并利用领域术语词典作为外部知识，融入舆情文本的编码过程中，从而提高司法舆情敏感信息分类的准确率。

1. 融入领域术语词典的司法舆情敏感信息识别模型

为了更好地融入领域术语词典，本文构造基于多头注意力的双向循环网络和卷积网络结合的司法舆情敏感信息识别模型（Multi-head Attention in RNN and CNN for Sensitive Information，MARC-SI），如图1所示。具体来讲，模型架构包括四个部分：编码层、领域术语词典融入层、局部特征提取层和分类层。编码层分别将舆情文本和领域术语词典进行编码和特征关注；领域术语词典融入层计算领域术语词典与舆情文本的相似度并将其融入文本表征中；局部特征提取层在领域术语词典融入层的基础上提取权重高的局部特征；分类层对提取的特征进行类别概率的预测。下面将对每层进行详细描述。

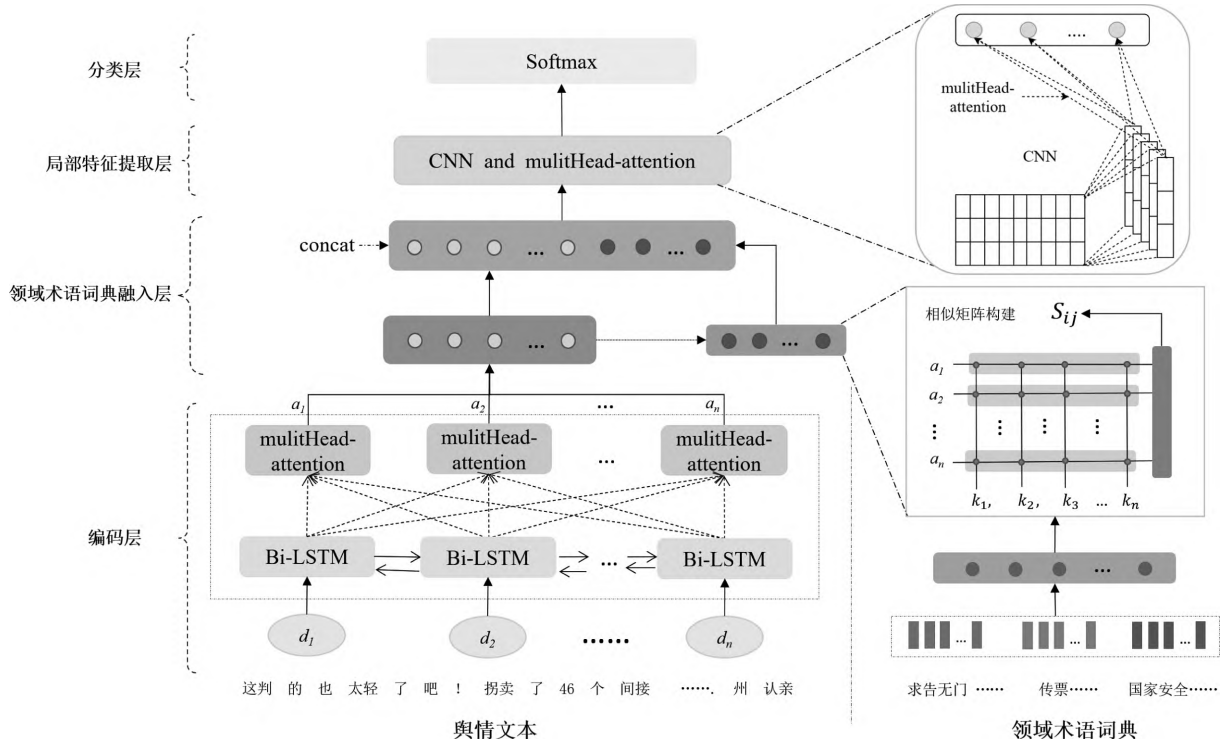

图1 融入领域术语词典的司法敏感信息识别模型示意图

1.1 编码层

要实现领域术语词典融入，需要先对舆情文本和领域术语词典进行编码。

对于舆情文本，设 $D=\{d_1,d_2,\cdots,d_n\}$，其中，$d_i\in V$ 为第 i 个位置的词，V 为词表。通过词嵌入将 D 转为 $\hat{\boldsymbol{D}}^e=\{\hat{d}_i^e\}$，其中 e 为词嵌入的维度，词嵌入由 Word2Vec 开源工具中的 Skip-gram[15] 模型预训练得到。

由于此前向量表征未考虑上下文语义特征，故需要将舆情文本向量表征 $\hat{\boldsymbol{D}}^e$ 输入一个可以理解上

下文的编码机制中。本文采用 Bi-LSTM 作为理解上下文信息的嵌入机制，模拟单词之间的特征交互，并将两个方向的输出进行简单拼接，得到该网络层的输出 $\boldsymbol{H}$，其中每列向量表示舆情文本上下文的表征，如式(1)所示。

$$\boldsymbol{H}=\text{Bi-LSTM}(\hat{\boldsymbol{D}}^{e}) \tag{1}$$

研究发现，重要的信息分布在一段文字的几句话中[16]，为得到更为有效的上下文表征，需要计算词语之间的依赖关系和整句话的权重大小。本文利用注意力机制对上下文表征 $\boldsymbol{H}$ 进行权重的计算，如式(2)所示。

$$\text{att}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d_k})\boldsymbol{V} \tag{2}$$

首先将上下文表征 $\boldsymbol{H}$ 映射得到查询向量(query，$\boldsymbol{Q}$)和键值向量(key，$\boldsymbol{K}$；value，$\boldsymbol{V}$)，输出是 $\boldsymbol{V}$ 的加权求和，计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的点积后除以 $\sqrt{d_k}$，保证内积的结果不会太大。为考虑多角度、多层面的特征，所以使用多头注意力机制将多个自注意力连接起来，过程如式(3)所示。

$$\begin{aligned}&\text{multiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{concat}(\text{head}_1,\cdots,\text{head}_h)\boldsymbol{W}^{O}\\&\text{where head}_i=\text{att}(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{k},\boldsymbol{V}\boldsymbol{W}_i^{V})\end{aligned} \tag{3}$$

其中，concat 表示拼接操作，其中 $\boldsymbol{W}_i^{Q}\in\mathbf{R}^{d_{\text{model}}\times d_k}$，$\boldsymbol{W}_i^{K}\in\mathbf{R}^{d_{\text{model}}\times d_k}$，$\boldsymbol{W}_i^{V}\in\mathbf{R}^{d_{\text{model}}\times d_v}$，$\boldsymbol{W}^{O}\in\mathbf{R}^{hd_v\times d_k}$，$d_k$ 表示隐层特征维度大小。经过多头注意力网络将得到文本内的权重关系，为防止丢失原文本语义，对于输出的结果进行残差连接，如式(4)所示。

$$\boldsymbol{A}^{h}=\text{residualConnect}(\boldsymbol{M}^{d},\boldsymbol{H}) \tag{4}$$

其中，residualConnect 表示残差连接，$\boldsymbol{M}^{d}$ 表示经过多头注意力机制的输出结果，$\boldsymbol{A}^{h}$ 表示残差连接后的结果。

对于领域术语词典，设有 $X=\{x_1,x_2,\cdots,x_n\}$，其中 $x_j\in V$ 表示词典中第 j 个词，首先通过词嵌入将 X 转为 $\hat{X}^{e}=\{\hat{x}_j^{e}\}$，本文使用全连接网络对领域术语词典的文本内容进行编码，过程如式(5)所示。

$$\boldsymbol{K}^{h}=f(\boldsymbol{W}\hat{\boldsymbol{X}}+\boldsymbol{b}) \tag{5}$$

其中 $\boldsymbol{W}$ 表示可训练权重矩阵，b 为偏置向量，$f()$ 表示激活函数，对于术语词典的编码结果本文用 $\boldsymbol{K}^{h}$ 表示。

经过编码层将舆情文本和领域词典进行分别编码，得到 $\boldsymbol{A}^{h}$、$\boldsymbol{K}^{h}$，表示获取到具有上下文权重信息的舆情文本编码矩阵，以及领域术语词典的编码矩阵。

1.2　领域术语词典融入层

该网络层的目的是挖掘领域术语词典中词语信息的隐含知识，对模型的表征学习进行指导，主要是将领域术语词典与文本表征进行相关度计算后，融入文本表征。

首先将领域术语词典的表征 $\boldsymbol{K}^{h}$ 与舆情文本表征 $\boldsymbol{A}^{h}$ 计算相似矩阵，其过程如式(6)所示。

$$S_{ij}=\text{sim}(\boldsymbol{k}_i^{h},\boldsymbol{a}_j^{h}) \tag{6}$$

其中，S_{ij} 表示术语词典表征 $\boldsymbol{K}^{h}$ 中的第 i 个领域词与文本特征 $\boldsymbol{A}^{h}$ 的第 j 个隐向量之间的相似性，k_i^{h} 表示对应词典第 i 个领域词表征向量，a_j^{h} 表示 A^{h} 的第 j 列向量，sim 表示计算 k_i^{h} 与 a_j^{h} 之间相似度的可训练函数，如式(7)所示。

$$\text{sim}(k,a)=\boldsymbol{W}_{(s)}^{\mathrm{T}}(k;a;k\cdot a) \tag{7}$$

其中，$\boldsymbol{W}_{(S)}^{T}\in\mathbf{R}^{6d}$ 是待训练的权重向量，“·”表示元素按位相乘，“;”表示向量在行上进行拼接，k 与 $\boldsymbol{K}^{h}$ 的列向量对应，a 与 $\boldsymbol{A}^{h}$ 的列向量对应。为获取到每一个舆情描述与词汇直接的权重关系，如式(8)所示。

$$\hat{\boldsymbol{S}}=\text{softmax}(S_{ij})\boldsymbol{A}^{h} \tag{8}$$

其中，将 S_{ij} 进行归一化后与词汇嵌入矩阵 A^{h} 进行相乘后得到具有权重信息的相关矩阵 $\hat{\text{S}}$。最终将相似矩阵与原文本编码矩阵进行拼接，得到融入词典信息的文本表征 $\hat{\boldsymbol{A}}$，如式(9)所示。

$$\hat{\boldsymbol{A}}=[\boldsymbol{A}^{h};\hat{\boldsymbol{S}}] \tag{9}$$

1.3　局部特征提取层

对于融入领域术语词典的文本表征，本文利用 CNN 网络对其进行局部特征的编码，由于冗余信息同样包含在局部特征中，为此利用多头注意力机制对敏感信息权重高的局部特征进行关注。

首先，对已经融入词典信息的文本表征 $\hat{\boldsymbol{A}}$ 进行卷积操作，其中卷积操作的核心是利用滑动窗口对文本进行局部特征的编码，获取文本的 N-gram 信息，如式(10)所示。

$$\hat{\boldsymbol{Z}}^{k}=\text{CNN}(\hat{\boldsymbol{A}}) \tag{10}$$

其中，k 表示 CNN 网络的输出通道。然后将 $\hat{\boldsymbol{Z}}^{k}$ 进行多头注意力操作，获取到具有权重信息的特征矩阵。这里与舆情描述编码层的注意力机制不同，前者主要是关注文本表征时权重高的信息，而这里主要对 CNN 网络输出通道进行关注，关注局部

特征中更为明显的 N-gram 信息，过程如式(11)所示。

$$\boldsymbol{O}^k = \text{mulitHead-attention}(\hat{\boldsymbol{Z}}^k) \tag{11}$$

其中，mulitHead-attention 表示多头注意力操作，其中具体细节如式(2)和式(3)所示，$\boldsymbol{O}^k$ 表示多头注意力操作的结果。

1.4 分类层

在分类层中，为得到文本分类概率分布，将在局部特征提取层中得到的 $\boldsymbol{O}^k$，利用 softmax 将其映射到分类空间，如式(12)所示，其中，$P(D)$表示对于输入舆情文本 D 的预测概率分布。

$$P(D) = \text{softmax}(\boldsymbol{O}^k) \tag{12}$$

在训练时，文本采用交叉熵损失来衡量预测敏感信息概率分布和真实概率分布之间的差距，并通过反向传播算法训练和更新模型参数，损失函数如式(13)所示。

$$\text{loss} = -\sum_{x\in T}\sum_{i=1}^{C} P_i^t(x)\log(P_i^p(x)) \tag{13}$$

其中，T 为训练集，x 为训练集中的每一个样本，$P_i^t(x)$表示样本 x 的真实概率分布，$P_i^p(x)$表示样本 x 的预测概率分布，C 表示所有类别。

2 实验与结果分析

2.1 数据集以及领域术语词典的构建

为了验证司法舆情敏感识别模型的有效性，本文构建了司法舆情信息数据集。首先，从 2020 年 3 月 1 日到 2020 年 6 月 1 日对新浪微博、github 等网站的舆情信息进行获取，经过人工筛选和标注后构成共两万条司法舆情文本，包括司法相关敏感信息 4 169 条，其他三类数据 15 685 条，具体统计信息如表 2 所示。

表 2 数据集大小以及数据特点

类别	总计/篇	平均字数/字
司法相关敏感信息	4 169	142
司法相关不敏感信息	8 338	139
司法无关敏感信息	6 154	138
司法无关不敏感信息	1193	146

本文按照 8∶1∶1 的比例构建训练集、验证集和测试集，具体的数据集划分细节信息如图 2 所示。从图 2 可以看出，在训练集、测试集、验证集中，四类数据分布与原始数据保持一致。

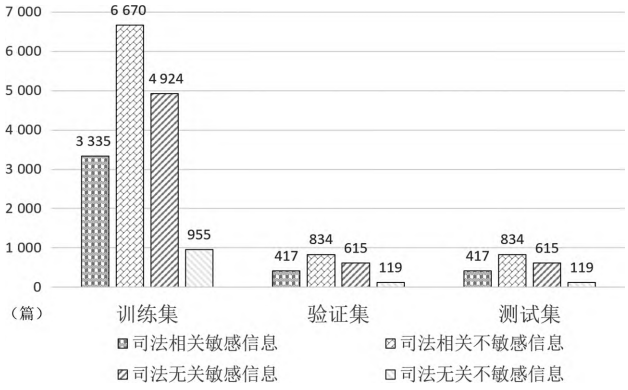

图 2 数据集划分信息

领域术语词典的构建对于司法敏感信息识别任务极为重要。本文构建的领域术语词典包含人工构建的司法敏感词汇、司法相关词汇和公开中文敏感词汇①。其中从裁判文书网②和中国法院网③筛选部分词汇构建司法相关词汇；根据司法舆情数据特点中选取部分词汇构建司法敏感词汇。其中术语的组成有字、词及短语，具体词汇数量及示例如表 3 所示。

表 3 领域术语词典大小及示例

领域术语词典	数量	示例
司法相关词汇	431	行政赔偿、法定期限、人身权利、交通事故
司法敏感词汇	3 470	压迫、罪恶滔天、姑息养奸、鸣冤、官威、阳奉阴违
公开敏感词汇	770	法轮功、禁制

从表 3 示例中可以看出，司法相关词汇与司法敏感词汇在感情表达上不同，前者更倾向于描述法律事实，而后者则倾向于表达情感。由于司法相关的词汇对于实现司法领域分类有着明显的指向性，但不能反映司法舆情的敏感性，因此本文构造包含司法敏感词汇和司法相关词汇组成的司法敏感词汇以及公开敏感词汇的领域术语词典来辅助司法敏感信息识别。

2.2 参数设置

本实验中，设计训练轮次为 20 轮，模型的学习

① https://github.com/57ing/Sensitive-word
② http://wenshu.court.gov.cn/
③ https://www.chinacourt.org/index.shtml

率为 1e-4，设置舆情文本最大截取长度为 300 字，词嵌入维度为 512，Dropout[17]为 0.5，卷积神经网络模型中滤波器（输出通道）个数为 256，滑动窗口大小为（2，3，4），优化算法使用 Adam[18]。

2.3　评价指标

对于本任务来讲，通过计算宏平均和微平均的指标更能评价司法敏感信息分类模型的效果。本文主要采用微平均 F_1 值（Micro-F_1）、宏平均精确率（Macro_Precision，P）、宏平均召回率（Macro_Recall，R）、宏平均 F_1 值（Macro-F_1）作为评价指标[19]，其中计算过程如式（14）～式（17）所示。

$$\text{micro-}F_1=\frac{2\,\overline{\text{TP}}}{2\,\overline{\text{TP}}+\overline{\text{FP}}+\overline{\text{FN}}} \tag{14}$$

$$P=\frac{1}{n}\sum_{1}^{n}\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{15}$$

$$R=\frac{1}{n}\sum_{1}^{n}\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{16}$$

$$\text{macro-}F_1=\frac{2\times P\times R}{P+R} \tag{17}$$

这些指标基于“混淆矩阵[20]”，其中，TP 表示真正例，FP 为假正例，TN 为真反例，FN 为假反例，$\overline{\text{TP}}$、$\overline{\text{FP}}$、$\overline{\text{TN}}$、$\overline{\text{FN}}$分别为混淆矩阵对应元素在所有类别 n 上的平均值。

2.4　基线模型选取

此部分本文选用几类非常典型的文本分类模型作为其线模型，基线模型的介绍如下所示。

CNN[21]模型　主要包括一个卷积层和一个池化层，再通过一个全连接层进行分类。

Bi-LSTM Attention[22]模型　使用双向循环神经网络和一个 Attention 层，再通过一个全连接层进行分类。

RCNN[15]模型　Lai 等人提出的一种结合 RNN 和 CNN 进行分类的神经网络模型，主要包括一个循环神经网络层和一个卷积层，再通过一个全连接层进行分类。

BERT[23]模型　通过 BERT 预训练模型进行文本表征后通过全连接网络进行分类[24]。

Transformer[13]模型　使用 Transformer 模型中的两个编码器编码，再通过一个全连接层进行分类。

FastText[25]模型　将整篇文档的词及 N-gram 向量叠加平均得到文档向量，然后将文档向量进行归一化后做多分类。

SVM　定义在特征空间上的间隔最大的线性分类器，通常用于文本分类任务，模型的文本特征提取和表示方法与文献[26]一致。

2.5　实验结果

2.5.1　基线模型对比实验

为了验证本文提出方法的有效性，与 2.4 节的基线模型进行了对比，实验结果如表 4 所示。

表 4　基线模型对比实验

（单位：%）

Model	Micro-F_1	P	R	$F1$
CNN	84.04	86.41	84.67	85.28
Bi-LSTM Attention	81.25	82.18	83.76	82.88
RCNN	85.38	86.44	86.92	85.27
BERT	83.95	84.95	84.21	84.57
Transformer	75.23	81.76	71.06	74.78
FastText	83.38	86.44	79.92	82.27
SVM	76.34	78.33	76.21	76.78
MARC-SI	**89.28**	**91.16**	**89.66**	**89.14**

从表 4 中可以看出，与基线模型、预训练模型和机器学习模型相比，MARC-SI 都有不错的效果，证明本文中所提出的融入领域术语词典的方法对于司法领域敏感信息识别任务是有效的。从实验结果中可以看出，RCNN、FastText 模型均有不错的效果，表明本文中所选用的模型架构和基于局部特征提取的思想是合理的，而其中 BERT 预训练模型由于其分词结构固定反而不适用于本任务。对于多数任务效果不错的 Transformer 模型，在本任务中效果不佳，可能是由于舆情文本中冗余信息过多，导致其中自注意机制不能有效地进行特征提取。从结果可以看出，本文所提模型 MARC-SI 在司法舆情敏感信息分类中具有明显的优势。

2.5.2　消融实验

为验证 MARC-SI 模型中每一层网络对于整体分类的有效性，本文设计了消融实验，其中（-）编码层是将 Bi-LSTM Attention 层去除，代替为全连接层，（-）领域术语词典融入层是将领域术语词典融入层去除，（-）局部特征提取层是将 CNN 和 Self-Attention 层替换为全连接层，实验结果如表 5 所示。

表5 消融实验 （单位：%）

Model	P	F_1
(-)编码层	82.11	82.02
(-)领域术语词典融入层	88.66	88.03
(-)局部特征提取层	87.66	87.83
MARC-SI	**90.83**	**89.14**

分析表5中的结果，去除编码层的效果比MARC-SI的F_1值低7%左右，说明对于舆情文本和领域术语词典的编码仍然是重要的一块；而融入领域术语词典可以提升整体的模型效果1%左右，说明对于本任务来讲，领域术语词典对于模型的学习是有指导作用的；对于去除局部特征提取网络，其比MARC-SI的F_1值低2%左右，说明融入术语词典后，整体网络还是需要进行特征的提取。从消融实验中可以看出，本文所提出的网络模型架构对于司法敏感信息识别任务是有效的。

2.5.3 不同术语词汇对实验结果的影响

由于领域术语词典的类型对模型影响比较大，为比较领域术语词典对于模型的影响，将不同领域术语词汇分别输入MARC-SI模型进行实验。其中分别融入手动构建的司法敏感词汇和公开敏感术语词汇进行实验，结果如表6所示。

表6 不同术语词汇对实验结果的影响

（单位：%）

Model	P	F_1
(a)司法敏感词汇	87.12	87.23
(b)公开敏感词汇	86.25	86.11
(a+b)所有词汇	90.83	89.14

从表6中的结果可以看出，手动构建的司法敏感词汇比公开的敏感词汇的F_1有1%左右的提升，说明本文中构建的司法领域术语词汇具有很强的领域特征，能够加强敏感术语的表征。而整体的领域术语词典的融入比少量的领域术语效果更佳，表示领域知识的覆盖面对于增强领域术语的表征有很大的影响。

分析表4和表6的实验结果，表明本文所提出的融入领域术语词典的方法比基线模型的方法均有不错的效果提升，反映出领域知识的融入可以增强专业术语的表征。

2.5.4 案例分析

为验证领域术语词汇和冗余信息对于MARC-SI的结果的影响，本文列举表7中所示例1和例2信息，二者均为司法敏感信息，选用基线模型CNN、Bi-LSTM Attention模型与MARC-SI的结果进行对比，表7中“结果”一栏表示相应模型判断的结果是否为司法敏感信息。

表7 案例分析

舆情文本	模型	结果
例1：@福建高院 @福建检察//@漳州古雷征海林建新冤案：还是漳州司法的耻辱，也是李祥等法官终身抹不去的污点//@古雷神：//@斯诺登科_996：这样断案奇才也只有漳州法院才能出现，中国司法的悲哀！发布了头条文章：《没有……罪吗？互不相识的“同案犯”》没有投资款也没有贪污款去向，不影响定罪吗？互不相识的“同案犯”	CNN	是
	Bi-LSTM Attention	否
	MARC-SI	是
例2：醉驾入刑。依照相关法律，应该是双开了。关键是现在的武汉，湖北高院的干部还能公然醉驾违法，胆子的确够大。【#官方回应湖北高院防疫干部醉驾撞车#：已停职，警方已刑事立案】3月11日，武汉救援车队女志愿者开……警。目前，@湖北高院 通报，许某被停职。@武汉武昌交警通报，因醉驾已对许某刑事立案。...全文	CNN	否
	Bi-LSTM Attention	否
	MARC-SI	是

分析表7，由于冗余信息过多使Bi-LSTM Attention模型不能进行有效的识别；同时，虽然CNN模型对局部信息进行提取后可以减少冗余信息的影响，但由于其难以增强敏感术语的表征，如例2中“双开”“公然醉驾违法”等敏感术语的表征，使其效果不如MARC-SI模型。从结果可以看出，本文设计的MARC-SI模型对于描述不规范、具有冗余信息的文本有更好的表征能力，也可以很好地利用司法敏感信息词汇对司法舆情敏感信息进行更有效的分类。

3 总结与展望

本文基于领域知识指导模型自动学习特征的思想，提出融入领域术语词典的司法舆情敏感信息识别模型。该模型通过融入领域术语词典，能够很好地提升司法敏感信息识别的效果，证明融入领域术语词典对于司法舆情敏感信息识别任务的有效性。

下一步可以研究将司法知识、法律文书或司法知识图谱等外部知识融入深度学习模型，进而提升识别效果。

参考文献

[1] Ding Z Y, Jia Y, Zhou B. Survey of data mining for microblogs[J]. Journal of computer research and development, 2014, 51(4): 691-706.

[2] **赵承鼎，郭军军，余正涛，等.基于非对称孪生网络的新闻与案件相关性分析[J].中文信息学报**，2020，34(03)：99-106.

[3] **韩鹏宇，高盛祥，余正涛，等.基于案件要素指导的涉案舆情新闻文本摘要方法[J].中文信息学报**，2020，34(05)：56-63.

[4] Chow R，Golle P, Staddon J. Detecting privacy leaks using corpus-based association rules[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008: 893-901.

[5] Cheng X Y, Kang W, Guo Y. An algorithm of network sensitive information features extracting[C]// Applied Mechanics and Materials. Trans Tech Publications Ltd, 2014, 556: 3558-3561.

[6] Xu G, Yu Z, Qi Q. Efficient sensitive information classification and topic tracking based on Tibetan web pages[J]. IEEE Access, 2018, 6: 55643-55652.

[7] Neerbek J, Assent I, Dolog P. Detecting complex sensitive information via phrase structure in recursive neural networks[C]//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2018: 373-385.

[8] Chung J，Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]//Proceedings of the NIPS Workshop on Deep Learning, 2014.

[9] Xu G, Wu X, Yao H, et al. Research on topic recognition of network sensitive information based on SW-LDA model[J]. IEEE Access, 2019(7): 21527-21538.

[10] Xu G, Yu Z, Chen Z, et al. Sensitive information topics-based sentiment analysis method for big data [J]. IEEE Access, 2019(7): 96177-96190.

[11] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[C]//Proceedings of the Advances in Neural Information Processing Systems 14 Vancouver, British Columbia, Canada, 2001.

[12] Zhang S, Zheng D, Hu X, et al. Bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, 2015: 73-78.

[13] Vaswani A，Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5998-6008.

[14] Yang M, Chang H, Luo W. Discriminative analysis-synthesis dictionary learning for image classification [J].Neurocomputing, 2017, 219: 404-411.

[15] Mikolov T. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, 26: 3111-3119.

[16] Alfonseca E, Pighin D, Garrido G. Heady: News headline abstraction through event pattern clustering [C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia: ACL, 2013: 1243-1253.

[17] Srivastava N, Geoffrey E H，Krizhevsky A，et al. Dropout: A simple way to prevent neural networks from over fitting [J].Journal of Machine Learning Research, 2015, 15(1): 1929-1958.

[18] Kingma D P，Ba J. Adam: A method for stochastic optimization[C]//Proceedings of the International Conference on Learning Representations，2015: 1-15.

[19] Kowsari K, Jafari M K, Heidarysafa M, et al. Text classification algorithms: A survey[J]. Information, 2019, 10(4): 67-75.

[20] Lever J，Krzywinski M, Altman N. Points of significance: Classification evaluation[J]. Nature methods, 2016, 13(8): 603-604.

[21] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of the Conference on Empirical Methods on Natural Language Processing. Doha, Qatar, 2014: 1746-1751.

[22] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 207-212.

[23] Devlin J, Chang M, Lee K, et al. BERT: Pretraining of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, MN, USA, 2019: 4171-4186.

（**下转第** 92 **页**）

宋威(1981—),**通信作者,博士,教授,主要研究领域为机器学习、数据挖掘、模式识别等。**

E-mail:songwei@jiangnan.edu.cn

周俊昊(1994—),**硕士研究生,主要研究领域为自然语言处理、命名实体识别等。**

E-mail:zhoujunhao@stu.jiangnan.edu.cn

(**上接第** 83 **页**)

[24] Sun C,Qiu X, Xu Y, et al. How to fine-tune bert for text classification?[C]//Proceedings of China National Conference on Chinese Computational Linguistics. Springer, Cham, 2019: 194-206.

[25] Joulin A, Grave É, Bojanowski P, et al. Bag of tricks for efficient text classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 2017: 427-431.

[26] He Y X, Sun S T, Niu F F, et al. A deep learning model enhanced with emotion semantics for microblog sentiment analysis[J]. Chinese Journal of Computers, 2017, 40(4): 773-790.

张泽锋(1996—),**硕士研究生,主要研究领域为自然语言处理、信息检索。**

E-mail:zefeng_z@126.com

余正涛(1970—),**通信作者,教授,博士生导师,主要研究领域为自然语言处理、信息检索、机器翻译。**

E-mail:ztyu@hotmail.com

毛存礼(1977—),**博士,副教授,主要研究领域为自然语言处理、信息检索、机器翻译。**

E-mail:maocunli@163.com