中文引用格式:王欣,干嵌锐,许雅玺,等. 基于字词向量融合的民航智慧监管短文本分类[J]. 中国安全科学学报,2024,34(2):37-44.

英文引用格式:WANG Xin,GAN Zurui,XU Yaxi,et al. Short text classification of civil aviation intelligent supervision based on character-word fusion [J]. China Safety Science Journal,2024,34(2):37-44.

基于字词向量融合的民航智慧监管短文本分类*

王 欣[1]教授, 干嵌锐[1], 许雅玺**[2]副教授, 史 珂[3]高级工程师, 郑 涛[1]副教授

(1 中国民用航空飞行学院 计算机学院, 四川 广汉 618307;2 中国民用航空飞行学院 经济与管理学院,四川 广汉 618307;3 中国民用航空飞行学院 民航监察员培训学院,四川 广汉 618307)

中图分类号:X949 **文献标志码**:A **DOI**: 10.16265/j.cnki.issn1003-3033.2024.02.0121

基金项目:国家自然科学基金资助(U2033213);中央高校基本科研业务费专项资金资助(J2022-048, J2019-045)。

【摘 要】 为解决民航监管事项所产生的检查记录仅依靠人工进行分类分析导致效率低的问题,提出一种基于数据增强与字词向量融合的双通道特征提取的短文本分类模型,探讨民航监管事项的分类,包括与人、设备设施环境、制度程序和机构职责等相关问题。为解决类别不平衡问题,采用数据增强算法在原始文本上进行变换,生成新的样本,使各个类别的样本数量更加均衡。将字向量和词向量按字融合拼接,得到具有词特征信息的字向量。将字词融合的向量分别送入到文本卷积神经网络(TextCNN)和双向长短期记忆(BiLSTM)模型中进行不同维度的特征提取,从局部的角度和全局的角度分别提取特征,并在民航监管事项检查记录数据集上进行试验。结果表明:该模型准确率为0.983 7,F_1 值为0.983 6。与一些字嵌入模型和词嵌入模型相对比,准确率提升0.4%。和一些常用的单通道模型相比,准确率提升3%,验证了双通道模型提取的特征具有全面性和有效性。

【关键词】 字词向量融合; 民航监管; 短文本; 文本卷积神经网络(TextCNN); 双向长短期记忆(BiLSTM)

Short text classification of civil aviation intelligent supervision based on character-word fusion

WANG Xin[1], GAN Zurui[1], XU Yaxi[2], SHI Ke[3], ZHENG Tao[1]

(1 School of Computer, Civil Aviation Flight University of China, Guanghan Sichuan 618307, China; 2 School of Economics and Management, Civil Aviation Flight University of China, Guanghan Sichuan 618307, China; 3 Institute of Civil Aviation Supervisor Training, Civil Aviation Flight University of China, Guanghan Sichuan 618307, China)

Abstract: In order to address the inefficiencies in manually classifying and analyzing inspection records about civil aviation supervision, a dual-channel feature extraction short text classification model was proposed. The model combined data augmentation techniques and character-word vector fusion. The model aimed to tackle classification issues related to people, equipment and facilities, institutional procedures and institutional responsibilities in civil aviation supervised matters. In order to tackle the issue of class

* 文章编号:1003-3033(2024)02-0037-08; 收稿日期:2023-08-14; 修稿日期:2023-11-20

** 通信作者:许雅玺(1976—),女,四川成都人,硕士,副教授,硕士生导师,主要从事决策分析与优化、数据挖掘等方面的研究。E-mail:31858255@qq.com。

imbalance, data augmentation algorithms were employed to generate new samples by transforming the original texts, thereby balancing the sample sizes across different categories. The word vectors and character vectors were fused by combining them at the character level, resulting in character vectors that retain word-level features. These fused character vectors were then fed into TextCNN and BiLSTM for feature extraction at different dimensions. By extracting features from both local and global perspectives, this dual-channel approach aimed to capture comprehensive and effective information from the inspection records dataset in civil aviation regulatory matters. Experimental results on the civil aviation regulatory matter inspection record dataset demonstrate that the proposed model achieves an accuracy of 0.983 7 and an F_1 score of 0.983 6. Compared with some existing word embedding models and character embedding models, the accuracy is improved by 0.4%. Furthermore, when compared with commonly used single-channel models, the accuracy is increased by 3%, which validates the effectiveness and comprehensiveness of the features extracted by the dual-channel model.

Keywords: character-word vector fusion; civil aviation supervision; short text; text convolutional neural networks(TextCNN); bi-directional long short-term memory(BiLSTM)

0 引 言

面对航空运输量快速增长的挑战,高效的安全监管是保障民航运行安全的重中之重[1]。依靠传统的监管方式不能满足民航业安全监管的需求,必须积极开发和采用先进的安全监管技术和手段[2]。民航局正在智慧民航的框架下大力推进智慧监管建设,深入应用大数据、人工智能等新一代信息技术,促进监管效能的全面提升,使民航局能够及时掌握全行业、航空公司和相关机构的安全运行情况,提升安全管理水平[3]。民航监管事项检查记录是民航监管执法检查中针对监管事项所发现的问题而记录的文本信息,文本长度一般在10~70字之间。对监管事项检查记录文本进行分类是民航监管数据分析的基础任务,伴随着数据的海量增长,运用自然语言处理的文本分类技术,自动分类监管事项检查记录文本。对推动民航监管的智慧化、精准化,提升行业安全管理水平具有重要意义。

常用的文本分类方法主要分为基于统计机器学习的方法和基于深度学习[4]的文本分类方法2类。在基于统计学习的方法中,对文本的表征能力有限,而深度学习通过多个层数和多个神经元来处理问题,让每个神经元处理简单的任务,同时,通过增加层数来挖掘数据更深的涵义,大大提高处理复杂问题的能力[5]。因此,很多学者基于深度学习的文本分类开展了研究,如尚麟宇等[6]为更加充分分析铁路安全事件,提出一种基于宽度学习系统的铁路安全事件文本分类模型,提高了分类的准确率,能够更好地解决实际的铁路安全问题;辛苗苗等[7]为更加高效地提取出文本的核心内容,从字、词、句子3个方面构建向量,利用Word2Vec构建字向量和词向量,并结合双向长短期记忆(Bi-directional Long Short-Term Memory, BiLSTM)提取字向量和词向量的上下文信息,利用FastText模型提取句向量特征,通过试验验证此方法提高了文本的分类效率。针对常用的深度学习模型在梅花信息文本数据集中分类效果较差的问题,付红萍等[8]提出一种基于知识融合增强表征(Enhanced Representation from kNowledge IntEgration, ERNIE)模型和文本递归卷积神经网络(Text Recurrent Convolutional Neural Networks, TextRCNN)组成的分类模型,通过ERNIE预训练模型对文本进行编码,增强了模型的特征提取能力,TextRCNN利用卷积操作对文本进行自动特征抽取,在文本分类任务上取得了较好的效果。目前已有的研究中,大多数领域文本分类方法主要依赖于基于字符或词语的向量表示。然而,词向量表征可能会忽略单个字符所携带的语义信息,而字符级别的向量则可能无法充分捕捉到词汇组合的意义。此外,这些方法通常采用单通道的特征提取模型,限制了它们在同时捕获文本的全局和局部特征方面的能力。

民航监管事项检查记录是典型的短文本,同时又具有类别极度不平衡的特点。因此,笔者拟提出一种结合数据增强和深度学习的短文本分类方法。在不平衡数据集上应用简单数据增强(Easy Data Augmentation, EDA),再分别基于双向编码器的变压器表示(Bidirectional Encoder Representations from Transformers, BERT)进行字向量嵌入和Word2Vec

进行词向量嵌入,将字词融合向量分别送入文本卷积神经网络(Text Convolutional Neural Networks, TextCNN)和 BiLSTM 中进行局部和全局的多维度特征提取,并融合 2 个特征提取模块的输出结果进行分类预测,以期提高民航监管效能。

1　民航智慧监管短文本分类模型

1.1　模型框架

融合 BERT 与 Word2Vec 的 TextCNN 与 BiLSTM(Word2Vec based on TextCNN and BiLSTM with Attention, BERT and BWCLA)模型的结构如图 1 所示。在预处理阶段,预处理数据集,包括文本数据增强、清洗、分词、分字等操作。在嵌入层,将分字后得到的文本序列输入到 BERT 预训练模型中,以获得字嵌入向量。同时,将分词的序列输入到 Word2Vec 中进行训练,以得到民航领域的预训练模型。将分词文本输入该预训练模型,以获得词嵌入向量。将字向量和词向量逐字拼接融合,得到具有词含义特征的字向量。将融合的向量分别送到 TextCNN 模块和 BiLSTM 模块中进一步提取特征。在输出层融合 TextCNN 和 BiLSTM 的特征进行分类预测。

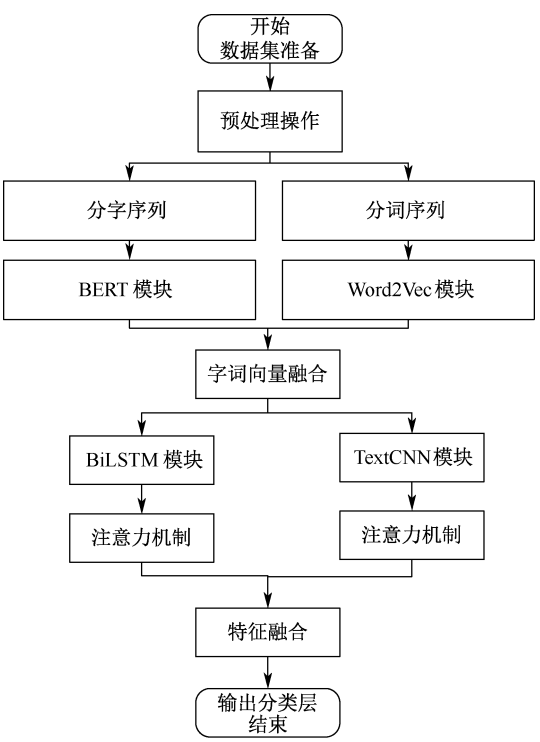

图 1　BWCLA 模型结构

Fig. 1　Model structure of BWCLA

1.2　数据增强

文中所面对的民航监管事项检查记录数据集是类别极度不平衡的文本数据集。如果直接在数据集上应用文本分类算法,很难获得精度上的提升。故采用 EDA[9] 通过以下 4 种方法进行文本数据集的数据增强:

1) 同义词替换。在句子中选取若干个词,并随机选择这些词的同义词进行替换。

2) 随机插入。随机从句子中选择一个词,并求出这个词的近义词,随机插入到句子的某个位置。

3) 随机交换。随机从句子中选出 2 个词并交换位置。

4) 随机删除。以某个概率随机删除句子中的词。

EDA 可以快速生成和原文本相似的句子,扩大样本数据量,实现数据集类别平衡,同时通过在数据集中增加合理的噪声,从而提升模型的泛化性能。

1.3　字词向量融合嵌入层

文本表示作为文本研究的基础任务,其准确度影响着众多下游任务的结果,目前主流的方法是通过向量嵌入进行文本的表示。英文语句以词为基本表示单位,通过空格切分而后进行词级别的向量嵌入可满足大部分任务的需求。不同于英文语句,中文语句以字为基本表示单位,每个字都代表一定的语义信息。使用字向量可以更准确地表达每个字的含义和上下文,将多个汉字组合成词会失去一部分信息。除此以外,字向量对于特定领域或任务的专业术语更友好,民航中存在大量的专业术语,通用的分词方法难以准确切分出专业术语,使用字向量可以更好地处理这些专业术语。但是基于字级别的向量嵌入容易丢失部分上下文语义信息,基于词级别的向量可以充分利用词之间的特征并且可以保留在分类任务中比较重要的词序信息。故将字向量与词向量结合起来,充分融合它们的优点。

使用 BERT[10] 预训练模型对文本进行字嵌入。BERT 采用 Transformer 架构,通过大规模的有标注文本数据进行训练。在预训练的过程中,BERT 采用掩码语言模型任务(Masked Language Model, MLM),随机掩盖输入序列中的一些标记,模型在训练时需要预测这些掩盖的标记。同时,BERT 还引入了下一句预测任务(Next Sentence Prediction, NSP),让模型学习相邻句子之间的关系。预训练完成后,BERT 会将参数保存下来,供后续的微调阶段使用。

为将字向量特征和词向量特征更好地融合起来,逐字将字向量与上下文可能组成词的词向量相融合,使字向量也拥有词的特征信息,增强字向量的

表达能力。为充分描述字与词之间的关系,引入一个矩阵,行为词的编号,列为字的编号,用于描述字是否属于词。若第 j 个词包含第 i 个字,那么矩阵相应位置为 1,否则为 0。矩阵中的元素大部分是 0,属于稀疏矩阵。为加快模型的收敛速度以及减少内存的消耗,将稀疏矩阵改进为三元组表。三元组表是稀疏矩阵的一种压缩存储方式。三元组表中只保存值为 1 的元素,提高了模型的收敛速度。字向量和词向量的融合模型结构如图 2 所示。

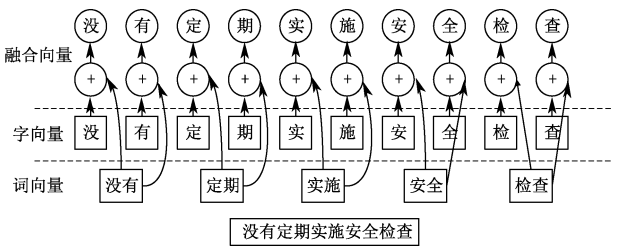

图 2 字向量和词向量的融合模型结构

Fig. 2 Fusion model structure of character vector and word vector

字向量和词向量融合如下式:

$$\boldsymbol{X} = M(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \tag{1}$$

式中:$\boldsymbol{Z}_1$ 为当前位置的字向量;$\boldsymbol{Z}_2$ 为当前位置的上下文组成词的词向量;$M(\,)$ 为拼接融合特征的函数,用于拼接字向量和词向量。

1.4 BiLSTM

输入融合向量矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$,其中,$\boldsymbol{x}$ 为输入的字词融合向量。正向长短期记忆网络(Long Short-Term Memory,LSTM)按顺序处理输入序列。正向 LSTM 通过输入当前字向量以及前一个时间步的隐藏状态,来计算当前时间步的隐藏状态,得到正向的隐状态 $\boldsymbol{h}_t = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T)$。反向 LSTM 则按反序处理输入序列。反向 LSTM 通过输入当前字向量表示以及后一个时间步的隐藏状态,来计算当前时间步的隐藏状态,得到反向的隐状态 $\boldsymbol{h}_t' = (\boldsymbol{h}_1', \boldsymbol{h}_2', \cdots, \boldsymbol{h}_T')$。正向 LSTM 能够捕捉到当前字与前面的字之间的依赖关系,反向 LSTM 能够捕捉到当前字与后面的字之间的依赖关系。每个 LSTM 单元内的门控机制可以控制信息的流动和记忆的更新,从而有效处理长距离的依赖关系。将正向 LSTM 和反向 LSTM 得到的隐状态合并起来,形成句子的特征表示。综合文本正向和反向的特征信息,能够全面捕捉到句子的语义特征和结构特征。

单向的 LSTM 只能获取和前文依赖相关的特征,具有一定的局限性,而 BiLSTM 是一种 LSTM 的变体,可以从正向和反向 2 个方向来获取文本的特征,能够进一步获取上下文语义的依赖关系,生成具有上下文语义特征的向量。

1.5 TextCNN

卷积神经网络(Convolutional Neural Networks,CNN)[12]通过设置多个不同大小的卷积核并行,有效提取局部关键信息,具有较强的特征提取能力。

使用 3 种不同的卷积核,尺寸分别为 3、4、5。输入融合后的向量矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$,3 种卷积核分别对融合向量进行卷积计算,提取出不同尺寸的局部特征 $[\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3]$。针对不同卷积核提取的特征,在池化层对每个局部特征进行最大池化操作,保留每个特征图中最显著的特征 $\boldsymbol{u}$。将所有池化后的特征 k 输入到全连接层进行拼接,得到局部特征向量 $[\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3]$,作为 CNN 的输出结果。TextCNN 模型结构如图 3 所示。

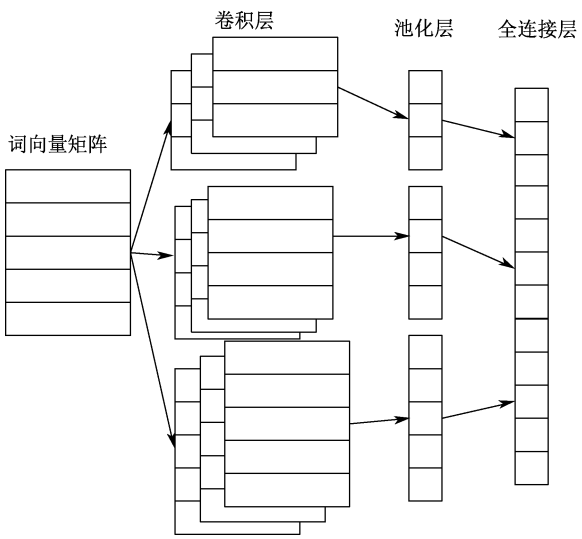

图 3 TextCNN 模型结构

Fig. 3 Model structure of TextCNN

输入融合向量矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$。卷积核滤波器的计算公式为

$$\boldsymbol{C}_j = g(\boldsymbol{x}_j \circ p + b) \tag{2}$$

式中:$\boldsymbol{C}_j$ 为卷积层特征向量;$g(\,)$ 为卷积层激活函数;$\circ$ 表示卷积计算;p 为卷积核;b 为偏置项。

经过卷积后的向量最大池化保留特征 $\boldsymbol{u}$ 作为全连接层的输入的计算公式为

$$\boldsymbol{u} = \max(\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_n) \tag{3}$$

通过全连接层将所有池化后的特征值拼接到特征向量,并通过 TextCNN 结构,提取文本局部特征,即

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n] \tag{4}$$

1.6　注意力机制

注意力机制[13]的重点就是让网络关注到它更需要关注的地方。利用注意力机制分别为 BiLSTM 和 TextCNN 模型的输出赋予不同的权重,提高重要词汇对分类结果的影响。

对于 BiLSTM 或 TextCNN 的特征提取结果,将特征向量乘以 3 个相应的权重矩阵得到查询向量 $\boldsymbol{Q}$,键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。下式以 TextCNN 结构的输出特征 $\boldsymbol{U}$ 为例。

$$\boldsymbol{Q} = \boldsymbol{U} \cdot \boldsymbol{W}_Q \tag{5}$$

$$\boldsymbol{K} = \boldsymbol{U} \cdot \boldsymbol{W}_K \tag{6}$$

$$\boldsymbol{V} = \boldsymbol{U} \cdot \boldsymbol{W}_V \tag{7}$$

式中 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$和 $\boldsymbol{W}_V$为可以学习的权重矩阵。

将得到的查询、键和值向量切分成 m 个注意力头。那么 $\boldsymbol{Q} = [\boldsymbol{Q}_1, \boldsymbol{Q}_2, \cdots, \boldsymbol{Q}_m]$, $\boldsymbol{K} = [\boldsymbol{K}_1, \boldsymbol{K}_2, \cdots, \boldsymbol{K}_m]$, $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_m]$,其中,$m$ 为头的数量。

对每个头计算注意力权重。

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{S(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})}{\sqrt{d_{\mathrm{k}}}}\boldsymbol{V} \tag{8}$$

$$\boldsymbol{H}_i = A(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i) \tag{9}$$

式中:A 为注意力机理函数;$S(\ )$为 softmax 函数,对每行进行归一化处理;d_{k} 为每个头中的查询或键向量的维度;$\boldsymbol{H}_i$ 为第 i 个注意力头的输出。

将每个头的向量进行拼接,得到多头注意力的输出。

$$MH(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = M(\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_m)\boldsymbol{W}_o \tag{10}$$

式中 $\boldsymbol{W}_o$ 为可训练的权重矩阵。

将注意力权重与对应的值向量进行加权求和,作为 BiLSTM 或 TextCNN 模块的输出特征。

1.7　特征融合层与输出层

特征融合层的任务是分别将 TextCNN 和 BiLSTM 提取得到的特征融合拼接起来。再将拼接好的特征输入到输出层中预测分类,输出分类结果矩阵。

2　试验结果与分析

试验基于 Python3.8+PyTorch 深度学习框架进行,CPU 为 Intel(R) Core i9-10900K,GPU 为 NVIDIA GeForce RTX 3090,显存 24 G,内存 96 G。

2.1　民航监管事项检查记录数据集

试验使用的民航监管事项检查记录数据集,是民航监管执法检查中针对监管事项所发现的问题而记录的文本信息。通过筛选、去除无用信息,最终获得 5 720 条数据,共计 4 个类别,包括与人、设备设施环境、制度程序和机构职责有关的问题。数据集类别数量极度不均衡。数据集监管事项记录(部分)见表 1。不同文本长度饼状图如图 4 所示。

表 1　监管事项检查记录数据集(部分)

Tab.1　Part of the dataset of supervision item records

监管事项检查记录的文本	类别
客舱机组普遍存在对客舱乘务员上方、服务间、卫生间内氧气面罩的数量分布不熟悉的情况	与人有关的问题
航材库房面积小,不能满足现工作量的需求,收料区已堆满,部分大件航材还在机库中摆放	与设备设施环境有关的问题
公司未建立具体的不定期抽查机制,未开展检查	与制度程序有关的问题
机组未按照最新组织机构调整除冰雪专门协调机构	机构职责的问题
安全检查记录填写不规范	与制度程序有关的问题
机场未设置安保控制中心	机构职责的问题
目前培训计划中缺少救护车司机培训、相关岗位员工培训	与制度程序有关的问题

2.2　数据增强与数据预处理

民航智慧监管数据集是短文本数据集,其类别分布存在不均衡性。为了解决这个问题,首先,对数据集进行去重处理,然后,使用 EDA 算法平衡各个类别的数量。数据增强后的具体数值见表 2。将训练集、测试集和验证集按照 8∶1∶1的比例划分。

数据预处理[14]主要包括中文分词、停用词删除等。

表 2　数据增强后的类别及其数量

Tab.2　Category and quantity after data augmentation

类别名称	数量
与人有关	3 913
与设备设施环境有关	3 904
与制度程序有关	3 896
与机构职责有关	3 921

2.3　评价指标

准确率 ACC 表示预测正确的样本占总样本的

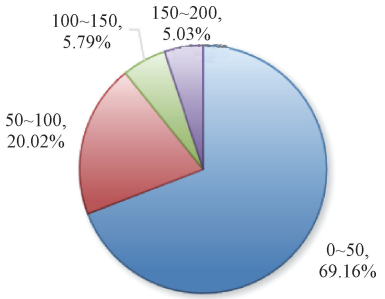

图 4　不同文本长度

Fig. 4　Different text length

比例,精确率 P 表示实际类别且预测类别都为正的样本占所有预测类别为正的样本比重,召回率 R 表示实际类别且预测类别都为正的样本占所有实际类别为正的样本比例,F_1 值为准确率和召回率的加权调和平均值[15]。

准确率公式为:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{11}$$

精确率公式为:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

召回率公式为:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

F_1 公式为:

$$F_1 = \frac{2PR}{P + R} \tag{14}$$

式中:TP 为判断为正且实际为正;TN 为判断为负且实际为负的;FN 为判断为负且实际为正;FP 为判断为正且实际为负。

2.4　试验参数

试验 BERT 预训练模型采用中文的 Bert-Chinese-wwm,损失函数采用交叉熵损失函数,Epoch 为 20,BILSTM 的隐藏层为 384。BWCLA 模型一些参数见表 3。

表 3　BWCLA 模型参数

Tab. 3　BWCLA model parameters

参数	名称	值
BERT	预训练模型	Bert-Chinese-wwm
D_1	字向量维度	768
D_2	词向量纬度	768
max_len	文本最大长度	70
hidden_size	隐藏层纬度	384
kernel_size	卷积核大小	[3,4,5]
n_filters	滤波器通道	128

续表 3

参数	名称	值
lr	学习率	0.000 1
batch_size	批量梯度下降数	64
epoch	训练批次	20
dropout	丢弃率	0.4

2.5　试验结果

为验证提出的 BWCLA 模型的优越性,以民航监管事项检查记录作为试验数据集,与一些经典的分类模型进行对比分析。

设计 1—5 组试验。在嵌入层,选用并对比 3 种不同的嵌入模型:BERT 模型、Word2Vec 模型以及字词融合的向量。在特征提取层,选用并对比 3 种特征提取模型:BiLSTM 模型、CNN 模型,以及 BiLSTM 和 CNN 并联结构的模型。

1 组:Word2Vec + BiLSTM 和 Word2Vec + CNN。采用 Word2Vec 模型进行文本词向量嵌入,再将得到的词向量分别输入到 BILSTM 或 CNN 进行训练分类。

2 组:BERT+BiLSTM 和 BERT+CNN。基于 BERT 预训练模型得到文本的基于字向量表示,再将字向量分别输入到 BILSTM 或 CNN 进行训练分类。

3 组:BERT+BILSTM+CNN。采用 BERT 进行字级别的向量嵌入,将向量分别输入到 BILSTM 和 CNN,再将 2 个通道的结果融合。

4 组:BERT + Word2Vec + BiLSTM。词嵌入层采用提出的字词向量融合,再将向量输入到 BiLSTM 进行训练分类。

5 组:BWCLA。通过 Word2Vec 与 BERT 嵌入得到融合字词特征的文本向量表示,融合字词向量分别输入到 BILSTM 和 TextCNN 模块,融合 2 个通道的特征。

监管事项检查记录数据集在不同模型下对比的试验结果见表 4。

表 4　不同模型对比试验的结果

Tab. 4　Comparison of experimental results of different models

模型	嵌入层	ACC	P	R	F_1
Word2Vec+BiLSTM	词向量	0.909 7	0.912 3	0.909 7	0.908 2
Word2Vec+CNN	词向量	0.853 3	0.854 2	0.852 7	0.852 7
BERT+BiLSTM	字向量	0.958 4	0.959 7	0.958 4	0.958 3
BERT+CNN	字向量	0.952 2	0.952 1	0.952 2	0.952 2

续表 4

模型	嵌入层	A	P	R	F_1
BERT+BiLSTM+CNN	字向量	0.979 8	0.979 8	0.980 2	0.979 9
BERT+Word2Vec+BiLSTM	字词融合的向量	0.972 2	0.972 4	0.972 2	0.972 1
BWCLA	字词融合的向量	0.983 7	0.983 9	0.983 7	0.983 6

从表 4 可以看出,字词融合向量相比于单一的字向量或词向量,能取得更好的结果。BiLSTM 和 CNN 并联的双通道模型相比于单一的模型,也能取得更好的效果。试验 1 和 2 的结果相比,相较于 Word2Vec 模型,BERT 模型在训练向量方面表现出更为优异的效果。试验 2 基于 BERT 的模型 BERT+BiLSTM 和 BERT+CNN 的 F_1 值相比于试验 1 有一定的提高。主要原因是通过 Word2Vec 训练的词向量是静态的,是上下文无关的。和 Word2Vec 相比,BERT 更能深度提取出文本的含义。通过对比试验 2 和 3,同样用 BERT 来训练词向量,文本分类结果都能达到不错的效果。而多通道模型基于将提取全局特征的 BiLSTM 和提取局部特征的 CNN 进行并列,对比单一的深度学习,特征提取效果更好,能更好地提取文本涵义,F_1 值提高 2.77%。通过试验 2 和 4 的对比,试验 4 采用字词向量融合嵌入,能较好地提高语义表征能力,F_1 值提高 1.38%。BWCLA 模型各个指标都取得了比其他模型更好的效果,验证了文中模型的优越性。

为进一步直观地展示 BWCLA 模型在民航监管事项文本分类任务的优越性,分析每个模型的训练过程,各模型训练过程验证集的准确率变化曲线如图 5 所示。

从图 5 可以看出,由于民航监管事项短文本的特性,以 BERT 为嵌入层的模型准确率均大于 Word2Vec 的准确率,以 TextCNN 为基线的模型基本上在 10 次迭代训练后达到收敛状态。以 BILSTM 为基线的模型在训练的过程中有动荡的趋势,大致需要 14 次迭代训练才能收敛。使用双通道模型的验证集准确率变化趋势最为优异,在 5 次迭代训练后达到收敛状态,且准确率最终收敛在 97%~98%,高于对比试验的其他模型。而提出的字词向量融合的双通道模型最终收敛到 98.38%,优于只用 BERT 模型进行字向量的嵌入 97.96%。通过试验验证了所提出的模型在性能上具有更优异的表现。

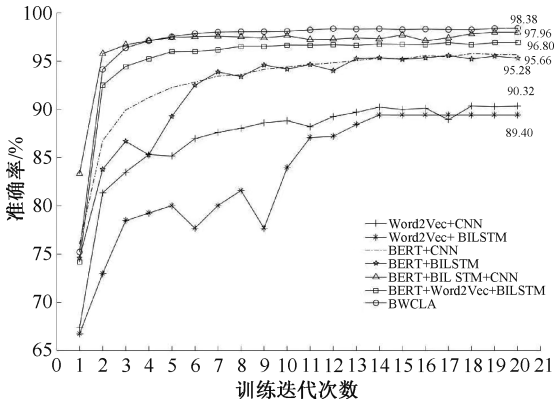

图 5　各模型在验证集上准确率变化曲线

Fig. 5　Accuracy curve of each model on the validation set

3　结　论

1) BWCLA 模型为处理短文本的数据量不均衡问题提供了参考方法,在智慧监管的短文本分类中取得不错的效果,各个评价指标均保持较高的水平。

2) 文中采用字词融合向量作为文本表示方法,并设计一种双通道模型以同时提取全局特征和局部特征。试验结果表明:与仅依赖于字向量或词向量的单通道模型相比,字词向量融合的双通道模型在特征提取方面展现出明显的优势。因此,字词融合向量在结合字级和词级信息后,能够更有效地表征文本数据,进而提高模型的整体性能。

3) 在未来的研究中,针对文中所用数据集,可尝试更好的数据增强方法和更加复杂的模型,进一步提取深层次的特征,提升领域数据集文本分类的准确率。

4) 目前,文中方法在短文本的领域数据集上表现出良好的效果。然而,为验证该方法在更广泛的短文本数据集上的通用性,需要进一步在更多常见的短文本数据集中进行验证。未来,计划进一步验证该模型在长文本中的适用性,并探究其在文本分类领域中的通用性。

参 考 文 献

[1]　吴剑青. 民航监管的数字化转型解决方案建议[J]. 民航管理, 2021(1): 25-27.
WU Jianqing. Suggestions on solutions for digital transformation of civil aviation supervision [J]. Civil Aviation

Management, 2021(1): 25-27.
[2] 张恒, 杨骁勇. 智慧监管怎么管[J]. 大飞机, 2022(1):18-22.
[3] 冯文刚. 基于深度长短记忆模型的民航安保事件分析[J]. 中国安全科学学报,2021,31(9): 1-7.
FENG Wen'gang. Research on civil aviation security event analysis based on deep LSTM model[J]. China Safety Science Journal,2021,31(9): 1-7.
[4] 杨秀璋, 宋籍文, 武帅, 等. 一种融合 Bert 预训练和 BiLSTM 的场景迁移情感分析研究[J]. 计算机时代, 2022(8):69-74,79.
YANG Xiuzhang, SONG Jiwen, WU Shuai, et al. Research on sentiment analysis of scene migration based on Bert pre-training and BiLSTM[J]. Computer Era, 2022(8):69-74,79.
[5] 苗将, 张仰森, 李剑龙. 基于 BERT 的中文新闻标题分类[J]. 计算机工程与设计, 2022, 43(8): 2 311-2 316.
MIAO Jiang, ZHANG Yangsen, LI Jianlong. Classification of Chinese news headlines based on BERT[J]. Computer Engineering and Design, 2022, 43(8): 2 311-2 316.
[6] 尚麟宇,尹明,肖畅,等. 基于 BLS 的铁路安全事件文本分类研究[J]. 中国安全科学学报,2022,32(6): 103-108.
SHANG Linyu,YIN Ming,XIAO Chang,et al. Research on text classification of railway safety incidents based on BLS[J]. China Safety Science Journal,2022,32(6): 103-108.
[7] 辛苗苗, 马丽, 胡博发. 融合多粒度信息的文本分类研究[J]. 计算机工程与应用,2023,59(9):104-111.
XIN Miaomiao, MA Li, HU Bofa. Research on text classification by fusing multi-granularity information[J]. Computer Engineering and Applications,2023,59(9):104-111.
[8] 付红萍, 陈恺之, 陈志泊. 基于 ERNIE-RCNN 梅花研究信息文本分类方法[J]. 东北农业大学学报, 2022, 53(5):20-31.
FU Hongping, CHEN Kaizhi, CHEN Zhibo. Research on plum blossom research information text classification based on ERNIE-RCNN[J]. Journal of Northeast Agricultural University, 2022, 53(5): 20-31.
[9] WEI J, ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks[EB/OL]. (2019-01-31). https://arxiv.org/pdf/1901.11196.pdf.
[10] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019:4 171-4 186.
[11] TANG Huanling, ZHU Hui, WEI Hongmin, et al. Representation of semantic word embeddings based on SLDA and Word2vec model[J]. Chinese Journal of Electronics, 2023, 32(3): 647-654.
[12] 鲍彤, 罗瑞, 郭婷, 等. 基于 BERT 字向量和 TextCNN 的农业问句分类模型分析[J]. 南方农业学报,2022,53(7): 2 068-2 076.
BAO Tong, LUO Rui, GUO Ting, et al. Agricultural question classification model based on BERT word vector and TextCNN[J]. Journal of Southern Agriculture,2022,53(7):2 068-2 076.
[13] 赵程栋, 庄继晖, 程晓鸣, 等. 基于特征注意力机制的 RNN-Bi-LSTM 船舶轨迹预测[J]. 广东海洋大学学报, 2022, 42(5): 102-109.
ZHAO Chengdong, ZHUANG Jihui, CHENG Xiaoming, et al. Ship trajectory prediction of RNN-Bi-LSTM based on characteristic attention mechanism[J]. Journal of Guangdong Ocean University, 2022, 42(5): 102-109.
[14] 王晓明. 基于深度学习的中文文本分类的关键技术研究[D]. 成都:电子科技大学, 2020.
WANG Xiaoming. Research on key technologies of chinese text classification based on deep learning [D]. Chengdu: University of Electronic Science and Technology of China, 2020.
[15] 刘凯洋. 结合 Bert 字向量和卷积神经网络的新闻文本分类方法[J]. 电脑知识与技术, 2020, 16(1):187-188.
LIU Kaiyang. A Chinese news text classification method of combining Bert character vector and convolutional neural networks[J]. Computer Knowledge and Technology, 2020, 16(1):187-188.

作者简介： 王 欣 (1973—),男,四川绵阳人,博士,教授,硕士生导师,主要从事机器学习、数据挖掘、自然语言处理方面的研究。E-mail:cafucwx@163.com。