计算机科学
COMPUTER SCIENCE

基于BERT和CNN的药物不良反应个例报道文献分类方法

孟祥福, 任全莹, 杨东燊, 李可千, 姚克宇, 朱彦

引用本文

孟祥福, 任全莹, 杨东燊, 李可千, 姚克宇, 朱彦. 基于BERT和CNN的药物不良反应个例报道文献分类方法[J]. 计算机科学, 2024, 51(6A): 230400049-6.

MENG Xiangfu, REN Quanying, YANG Dongshen, LI Keqian, YAO Keyu, ZHU Yan. Literature Classification of Individual Reports of Adverse Drug Reactions Based on BERT and CNN [J]. Computer Science, 2024, 51(6A): 230400049-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400049

基于 BERT 和 CNN 的药物不良反应个例报道文献分类方法

孟祥福[1]　任全莹[1]　杨东燊[1]　李可千[2]　姚克宇[3]　朱　彦[3]
1 辽宁工程技术大学电子与信息工程学院　辽宁 葫芦岛 125105
2 长春中医药大学医药信息学院　长春 130117
3 中国中医科学院中医药信息研究所　北京 100700
(marxi@126.com)

摘　要　在临床上,药物不良反应导致的死亡和用药不当造成的住院及门诊费急剧升高,成为临床安全合理用药面临的主要问题之一。目前对药物不良反应的回顾性分析和文献分析多以公开发表的文献资料为依据。学术文献作为重要的数据来源之一,如何自动批量地对其进行数据处理尤为重要。针对医药文本独特的表述方式,基于 BERT 及其组合模型进行文本分类技术比对实验,建立对药物不良反应个例报道文献数据进行高效快速分类的方法,进而分辨出药物不良反应的类型,有效预警药害事件。实验结果表明,使用 BERT 模型的分类准确率达到 99.75%,其可以准确高效地对药物不良反应个例报道文献进行分类,在辅助医疗、构建医学文本结构化数据等方面均具有重要的价值和意义,进而能够更好地维护公众健康。

关键词:药物不良反应;个例文献报道;医学文本分类;深度学习;BERT
中图分类号　TP391.1

Literature Classification of Individual Reports of Adverse Drug Reactions Based on BERT and CNN

MENG Xiangfu[1],REN Quanying[1],YANG Dongshen[1],LI Keqian[2],YAO Keyu[3] and ZHU Yan[3]
1 College of Electronic and Information Engineering,Liaoning Technical University,Huludao,Liaoning 125105,China
2 School of Medical Information,Changchun University of Chinese Medicine,Changchun 130117,China
3 Information Institute of Traditional Chinese Medicine,Chinese Academy of Traditional Chinese Medicine,Beijing 100700,China

Abstract　Clinically,the death caused by adverse drug reactions and the sharp increase in hospitalization and outpatient expenses caused by improper drug use have become one of the main problems faced by clinical safe and rational drug use. At present,the research of adverse drug reactions retrospective analysis and literature analysis is mostly based on published literature information. Academic literature is one of the important sources of data,and how to automatically process data in batches is particularly important. According to the unique expression of traditional Chinese medicine text,based on BERT and its combination algorithm,through the comparison experiment of text classification technology,an efficient and fast classification method for the literature data of adverse drug reactions case reports is established,and then the types of adverse drug reactions are distinguished. Experimental results show that the classification accuracy of BERT algorithm reaches 99.75%,which can accurately and efficiently classify the reported literature of adverse drug reactions,and has important value and significance for auxiliary medical treatment and constructing structured data of medical texts.

Keywords　Adverse drug reactions,Individual case literature report,Medical text classification,Deep learning,BERT

1　引言

在临床上,药物不良反应导致的死亡和用药不当造成的住院及门诊费急剧升高,成为临床安全合理用药面临的主要问题之一。从生物医学文献中提取出有价值的药物不良反应信息,从而有效预警药害事件,可以为临床安全合理用药提供技术参考,进而更好地维护公众健康[1]。

《个例药品不良反应收集和报告指导原则》要求:持有人应建立面向医生、药师、患者等群体的有效信息途径,主动收集临床使用、临床研究、市场项目、学术文献等涉及的不良反应信息。目前,药品不良反应的回顾分析和文献分析等研究的开展大部分都是基于公开发表的文献信息[2-6]。《关于发布个例药品不良反应收集和报告指导原则的通告》规定:学术文献是高质量的药品不良反应信息来源之一,持有人应定期对文献进行检索,并报告文献中涉及的个例不良反应。持有人应制定文献检索规程,对文献检索的频率、时间范围、文献来源、文献类型和检索策略等进行规定。有关不良反应的文献类型主要包括:个案报道、病例系列、不良反应综述等[7]。药

基金项目:国家自然科学基金(82174534);中央级公益性科研院所基本科研业务费专项资金(ZZ13-YQ-126,ZZ150314)
This work was supported by the National Natural Science Foundation of China(82174534) and Fundamental Research Funds for the Central Public Welfare Research Institutes(ZZ13-YQ-126,ZZ150314).
通信作者:朱彦(zhuyan166@126.com)

物不良反应个例报道每年发表的文献众多，如药监局不良反应数据库[8]（该数据库更新至2006年，共计7366条），如何自动批量地进行数据处理显得尤为重要，这需要对数据进行文本分类。医学文本分类对于辅助医疗、构建医学文本结构化数据以及快速进行药物不良反应个例与非个例数据分类，均具有重要的价值和意义。

文本分类是指在给定分类体系下，根据文本内容自动确定文本类别的过程。20世纪90年代，机器学习技术逐渐成熟，出现了很多经典的文本分类算法，如决策树[9]、朴素贝叶斯[10]、支持向量机[11]、最大熵[12]、最近邻[13]等。由于机器学习前期需要繁杂的人工特征提取过程，而人工选取特征难免存在疏漏和错误，因此自2012年深度学习方法被提出便逐渐被应用到文本分类中，同时被应用在信息检索、文档自动分类、文本过滤等多个领域。深度学习能够自动提取文本潜在特征，因此比机器学习方法更具优势。对于中文文本分类而言，一般流程可分为5步：预处理、索引、抽取统计特征、分类器以及评价，如图1所示。

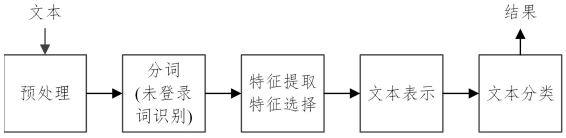

图1 文本分类处理流程

Fig. 1 Text classification process procedure

基于上述分析，本文首先介绍文本分类及其使用的方法；然后阐述基于深度学习的医学文本分类，并使用BERT及其组合模型对药物不良反应个例报道文献进行文本分类，对比各模型实验效果并做详细分析；最后进行总结与展望。

2 相关工作

目前，文本分类的实现方法大致可分为两类：一类基于统计机器学习，另一类则基于深度学习。由于大多数经典的基于机器学习的模型存在严重的依赖数据、依赖学习模型类型、不能执行非特定多任务等的局限性，相对而言，深度学习覆盖范围广、适应性好、学习能力强、有较强的可移植性且数据驱动上限高，因此深度学习算法在文本分类中经常被使用。常见的模型有：卷积神经网络（Convolutional Neural Network，CNN），递归神经网络（Recurrent Neural Network，RNN），预训练语言模型（Bidirectional Encoder Representations from Transformer，BERT），增强语义表示模型（Enhanced Language Representation with Informative Entities，ERNIE）等。

近年来，使用深度学习对基础分类算法进行改进成为了提升实验效果的有效途径之一。Zhng等[14]尝试用特征学习的方法来改进朴素贝叶斯文本分类模型，提出了一种双层贝叶斯模型——随机森林朴素贝叶斯（Random Forest Naive Bayes，RFNB）。Wang等[15]采用遗传算法优化的支持向量机对故障文本建立分类模型，其分类预测精确性较高，具有良好的使用价值。Jie等[16]通过引入Simhash和相邻文本的平均汉明距离，提出了一种改进的基于Simhash的KNN文本分类算法，解决了传统KNN文本分类算法中数据不平衡及计算量大的问题。Alsaleh等[17]采用了一种基于遗传算法的卷积神经网络对阿拉伯语进行文本分类。

文本分类可以被用于分类临床记录，以辅助识别患者所患疾病[18]。例如Turner等[19]评估了多种传统分类器（包括神经网络、随机森林、朴素贝叶斯、支持向量机等）在系统性红斑狼疮患者识别中的性能，其中具有统一医学语言系统（Unified Medical Language System，UMLS）概念唯一标识符（CUIs）的浅神经网络和同时具有CUIs和词袋模型（Bag-of-Words，BoW）的随机森林表现最优。Zhang等[20]提出一种基于迁移学习和集成学习的临床试验筛选标准短文本分类技术，基于BERT方法构建模型，并在CHIP2019评测三测试集上达到了0.811的F1值。Zhou等[21]基于深度学习对生物医学文本进行研究，以中国医院科技量值研究中累积的神经病学科、消化病学科、肿瘤学科的SCI论文为数据来源，分别训练并测试多种模型并评估其性能。Ye等[22]基于CNN和LSTM等算法对中医病历资料进行病史信息字段自动分类与抽取，旨在解决在医学病历混杂的文本信息中自动抽取所有病史信息与分类的问题。实验结果发现，LSTM的病史信息分类抽取F1值为0.8810，具有良好的分类效果。

Huang等[23]使用基于长短期项记忆（Long Short Term Memroy，LSTM）和门控递归单元（Gated Recurrent Unit，GRU）计算节点的双向递归神经网络提取文本特征，然后使用softmax对文本特征进行分类。该方法不需要人工设计特征，因此具有很好的可移植性。Luo等[24]采用的模型是先将经词嵌入处理的不完全数据输入栈式降噪自编码器中进行去噪训练，接着再将其输出传入BERT预训练模型中进行精化，以进一步改进词的特征向量表示。Chen等[25]以今日头条新闻公开数据集和THUCNews新闻数据集为实验对象，使用BERT和ERNIE模型通过领域预训练并结合TextCNN模型生成高阶文本特征向量并进行特征融合，实现语义增强，进而提升短文本分类效果。

以上各研究均在对文本分类算法进行改进，以得到更好的实验效果。本研究则是针对医药文本独有的表述，建立对药物不良反应个例报道文献数据进行高效快速分类的方法，进而分辨出药物不良反应的类型。

3 问题定义

药物不良反应个例报道文献数量众多，蕴含着大量有价值的药物不良反应信息。中医药文本有其独特的表述，如腹痛—肚子疼需要进行症状术语的对齐；导致药物不良反应的药品及中药方面的术语较杂，如土三七、复方丹参、牛黄解毒片等；且不良反应产生的症状、疾病较多，如胺碘酮片与血脂康胶囊联用致肝功能异常，宫内节育器致月经量增多，牛黄解毒片致肝小静脉闭塞症等。

针对上述情况，传统人工标识并进行分类的方法存在耗时久、易错率高等显著问题，故如何通过计算机自动化、批量化地进行文本分类，以高效快速地对药物不良反应个例报道与非个例报道进行区分成为文献报道数据处理的首要问题。

给定一个药物不良反应个例报道文献的项集合 $D=\{x_1,\cdots,x_n\}$，其中每一个元素是一个待分类项；类别集合 $C=\{y_1,y_2\}$，其中每一个元素是一个类别，即是个例、非个例；确定映射函数 $Y=f(x)$，使得任意一个 x_i 有且仅有一个 y_i 成立，其中 f 叫做分类器，分类算法的任务就是构造分类器 f。本文的目的是建立一种快速有效的方法，分辨出药物不良反应的类型 Y。

4 基于 BERT 的药物不良反应个例报道文献分类方法

使用 BERT 以及 BERT 的算法改进，完成对药物不良反应数据类型（是否为个例）的判定。

4.1 BERT 预训练语言模型

BERT 本质上是一种基于 Transformer 架构且能够进行双向深度编码的神经网络语言模型[26-28]，关键在于其利用自注意力机制的原理，并引入掩码语言模型（Masked Language Model，MLM）和下一句预测（Next Sentence Prediction，NSP）两种策略对不同层的上下文联合处理来进行双向深度预训练，以此缓解单向性约束问题。BERT 模型结构如图 2 所示。

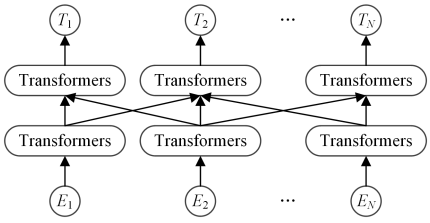

图 2 BERT 预训练语言模型

Fig. 2 BERT pre-training language model

为融合字左右两侧的上下文，使其根据上下文很好地表征字词的多义性，BERT 采用双向 Transformer 作为解码模块。Transformer 由编码器（Encoder）和解码器（Decoder）两部分组成，编码器用于对输入的文本数据进行编码表示，解码器用于生成与输入端相对应的预测序列。它将输入信号或句子对转换为隐藏向量序列[29]。此外，BERT 使用 MLM，可以预测一个顺序，因此双向上下文被认为是训练单词表示法[30]。

首先，对数据进行预处理，将其分成单字，并在数据前后各添加【CLS】符号和【SEP】，完成 Masked LM，如图 3 所示，即：

原数据：【CLS】牛黄解毒片致肝小静脉闭塞症【SEP】

处理后输入的数据：【CLS】牛黄【MASK】【MASK】片致肝小静【MASK】闭塞症

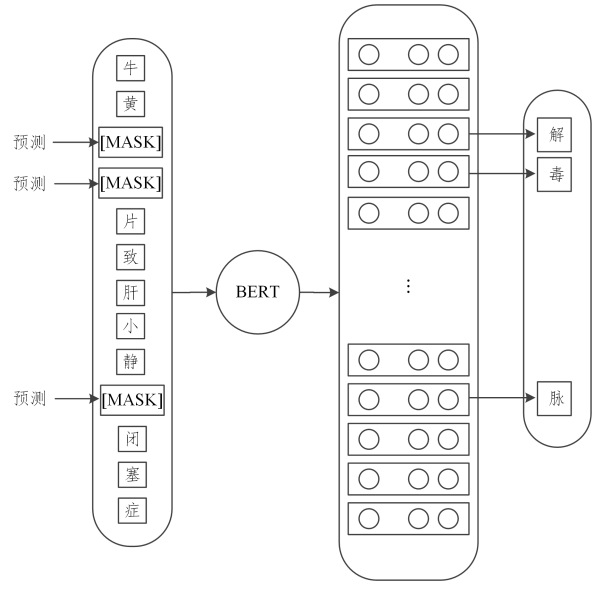

图 3 基于 BERT 的中医药文本分类流程

Fig. 3 BERT-based traditional Chinese medicine（TCM）text classification process procedure

把处理后的数据放入模型中，每个字符对应的字向量均由标记词嵌入、片段词嵌入、位置词嵌入 3 个向量组合，把向量输入模型中，判断是否是个例。BERT-classify 模型结构如图 4 所示。

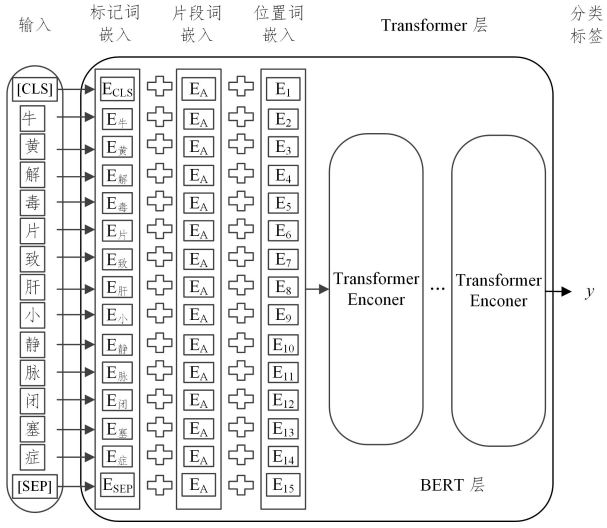

图 4 BERT-classify 的模型结构

Fig. 4 Structure of BERT-classify model

4.2 BERT＋CNN 模型

CNN 是一种前馈神经网络，是具有高效学习能力的深度学习算法之一。

标准的 CNN 由卷积层、池化层和全连接层 3 部分组成，其通过局部感知和权值共享的方法减少了训练过程中的参数数量，使训练的效率和准确率得到提升。CNN 模型结构如图 5 所示。

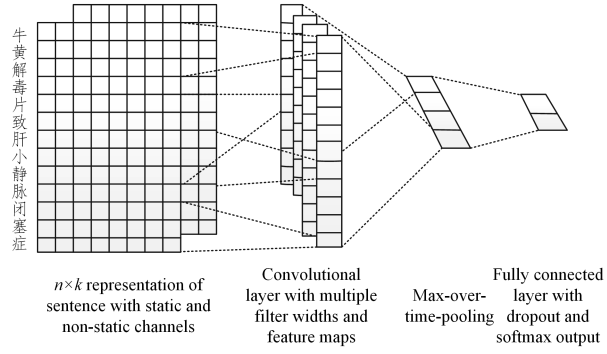

图 5 CNN 结构模型

Fig. 5 Architecture of CNN

使用 BERT 模型，先将标题集合 $A=\{A_1,A_2,\cdots,A_n\}$，例如（[牛黄解毒片致肝小静脉闭塞症]），变成词向量，输入卷积层中。每个卷积层由一个滤波器层、一个非线性层和一个空间采样层组成。第一层卷积层只提取一些低级的特征，更多层的网络能从低级特征中迭代提取更复杂的特征。将标题信息通过池化层进行特征提取后，输入到全连接层，把局部特征结合变成全局特征，使用 softmax 作为分类器，最后输出对应标题是否为个例。

4.3 BERT＋RNN 模型

RNN 是一类神经网络，由输入层、隐藏层及输出层 3 部分组成。在自然语言处理中，每个字的前后有语义联系，则使用 RNN 效果会更好。因为其他的神经网络的隐藏层仅能从输入层接收信息，而 RNN 通过对序列内的时间关系建模来

处理序列数据，一个时间序列中的时间点对应一个普通前馈神经网络，并在每个时间点加入一个记忆单元记录其输出值，将输出值传递给下一时刻。RNN结构模型如图6所示，其数学表达式如下所示，引自Wang等[31]基于互信息理论与递归神经网络的短期风速预测模型。

$$h_t=\sigma(UX_t+Wh_{t-1}) \tag{1}$$

$$O_t=f(h_t) \tag{2}$$

其中，h为状态值；$\sigma(\cdot)$为Tanh激活函数；U为t时刻输入的权重；W为$t-1$时刻状态的权重；X和O为神经网络的输入值和输出值；$f(\cdot)$为生成神经网络输出值的函数。

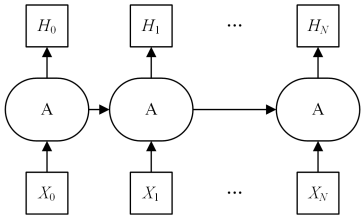

图6 RNN结构模型

Fig. 6 Architecture of RNN

4.4 BERT+DPCNN模型

DPCNN是一个广泛而有效的基于词级的深层文本分类卷积神经网络，它可以通过不断加深网络来提取长距离文本依赖[31]。DPCNN主要由一个区域嵌入层(文本区域嵌入层)和两个卷积块组成。标题集合$A=\{A_1,A_2,\cdots,A_n\}$进入DPCNN网络会经过一个包含3个不同卷积特征提取器的region embedding层，并经过两层的等长卷积来为接下来的特征抽取提供更宽的感受野，最后输入分类类型以判断是否为个例。DPCNN采用预激活的方法。图7给出了DPCNN模型结构[32]。

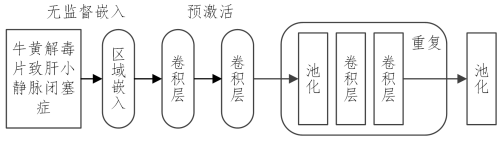

图7 DPCNN模型框架

Fig. 7 Architecture of DPCNN

5 实验

5.1 数据来源

本文所用数据的来源为CNKI下载的题录数据，人工对其进行标注产生用于文本分类的数据集。该数据收集整理了目前医学文献上出现的不良反应个例共2768条，其中收集了不良反应个例1529条，不良反应非个例1239条。实验数据格式如表1所列。

表1 实验数据

Table 1 Experimental data

标题	关键字	摘要
复方丹参注射液致药物性肝炎1例分析	复方丹参注射液；药物性肝炎	丹参注射液是临床上广泛应用于心、脑血管疾病和肝脏疾病的药物。在临床治疗中发现1例患者两次静点复方丹参注射液出现肝脏损害表现，经检查确诊为药物性肝炎。
大学生水痘误诊1例报告	水痘；误诊	1病历摘要：女，19岁，大学二年级，集体住宿，冬末春初发病。首次因颜面部及颈部红色斑丘疹2d，伴瘙痒、轻度畏寒，于他处就诊，诊断：湿疹，给予防风通圣丸口服。[第一段]
复方丹参液治疗早期脑出血31例临床分析	复方丹参液；活血化瘀法；早期脑出血	采用复方丹参液静滴治疗早期脑出血31例，痊愈率为51.6%，显效率为87%，有效率达100%。
牛黄解毒片致肝小静脉闭塞症	牛黄解毒片，静脉	患者女，37岁。因纳差18年，反复双下肢浮肿12年，腹胀2个月，于2006年3月7日来我院就诊。1988年因受惊吓及失恋刺激后出现闭经等内分泌紊乱症状，营养不良，无肝病史及其他病史。2001年因解大便困难开始自服牛黄解毒片4片，4次/d，约2年。2004年出现乏力，未引起重视。
顺铂联合艾迪注射液治疗胸腹腔积液疗效观察	胸腔积液/药物疗法；腹水/药物疗法；顺铂/治疗应用	2004-02/2007-07我院采用顺铂联合艾迪注射液腔内注射治疗恶性胸腹腔积液33例，取得了较好疗效，现报告如下。[第一段]

5.2 数据筛选

从2768条数据中筛选出含有标题、摘要和关键字的信息2364条，并标注是否为个例的真实标签。

5.3 实验设置

将筛选的2364条数据顺序随机打乱，按5:2:3的比例划分训练集、验证集和测试集。

本文模型实验的所有参数为：训练迭代次数$epoch=3$；BERT模型$batch_size=32$，其他模型$batch_size=128$；$pad\_size=32$；学习率$learning_rate=5\times10^{-5}$；隐藏层$hidden_size=768$；CNN，RCNN，RNN模型的卷积核数量$num_filters=256$，DPCNN模型的卷积核数量$num_filters=250$；$dropout=0.1$。为了降低模型过拟合的风险，训练遵循早停(Early stopping)原则，设置检测参数$detect_imp=1000$，即若模型在持续1000个batch的训练中所得结果没有明显的提升，就提前结束训练。

5.4 评价指标

分类结果是否准确一般由评价指标来评估。对于分类结果而言，使用的分类算法不同会导致最终所得的分类结果存在巨大差异。本实验选用了常见的分类算法评估模型，有以下4个评价指标，即准确率(Accuracy)、精准率(Precision)、召回率(Recall)和F1值。具体计算式如下：

1)准确率(Accuracy)

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}\times100\% \tag{3}$$

2)精准率(Precision)

$$Precision=\frac{TP}{TP+FP}\times100\% \tag{4}$$

3)召回率(Recall)

$$Recall=\frac{TP}{TP+FN}\times100\% \tag{5}$$

4)F-measure

$$F\text{-}measure=2\times\frac{Precision\times Recall}{Precision+Recall}\times100\% \tag{6}$$

本实验采用了二分类模型，即只有0和1两类，结合预测结果和实际结果得到以下4种判别结果的数量：正样本被模型预测为正的样本为正确肯定(True Positive，TP)；正样本被模型预测为负的样本为错误肯定(False Positive，FP)；负样本被

模型预测为正的样本为错误否定(False Negative,FN);负样本被模型预测为负的样本为正确否定(True Negative,TN)。

5.5 实验结果分析

本实验一共分为 4 组:实验一为标题和标签,实验二为标题、关键字和标签,实验三为标题、摘要和标签,实验四为标题、关键词、摘要和标签。将 4 组实验分别用以下 5 个模型进行预测,即 BERT,BERT_CNN,BERT_DPCNN,BERT_RCNN 和 BERT_RNN,最后得到各模型下每组数据的准确率、召回率和 F1 值等评价指标。本实验结果的详细数据如表 2 所列。

表 2 实验结果对比

Table 2 Experimental results comparison

(%)

模型	一(标题)			二(标题、关键字)			三(标题、摘要)			四(标题、关键字、摘要)		
	准确率	召回率	F1 值	准确率	召回率	F1 值	准确率	召回率	F1 值	准确率	召回率	F1 值
BERT	98.89	99.67	99.28	92.95	86.35	89.53	97.74	98.05	97.89	**99.58**	**99.68**	**99.63**
BERT_CNN	97.81	98.88	98.34	89.42	80.89	84.94	98.73	98.40	98.56	99.15	99.67	99.41
BERT_DPCNN	97.66	99.62	98.63	90.83	83.06	86.77	97.18	95.61	96.39	98.18	97.45	97.81
BERT_RCNN	97.97	97.82	97.89	89.99	81.91	85.76	97.04	96.77	96.90	97.88	97.13	97.50
BERT_RNN	97.81	99.25	98.52	92.67	86.07	89.25	97.18	97.39	97.28	97.60	95.94	96.76

通过实验结果可知,实验一中,各模型差距较小,其中,BERT 模型以准确率 98.89%、召回率 99.67%以及 F1 值 99.28%的数据优于其他模型;实验二中,各模型差距明显,其中 BERT 模型和 BERT_RNN 模型的效果较好,准确率都在 92%以上,召回率均在 86%以上,F1 值都在 89%以上。相对来说,BERT 效果最佳,而 BERT_CNN 效果最差,各项数据均低于 90%,其中召回率只有 80.89%;实验三中,BERT_CNN 模型(准确率 98.73%、召回率 98.40%、F1 值 98.56%)效果最好,整体达到了 98%以上,超越 BERT 模型(准确率 97.74%、召回率 98.05%、F1 值 97.89%)不到一个百分点;实验四中,BERT 模型预测最准确,召回率和 F1 值也是最高的。

基于实验数据的选择,实验四含有标题、关键字、摘要以及标签的实验组最佳。实验四的 5 个模型准确率、召回率以及 F1 值均高于 95%,其中 BERT 模型的准确率、召回率以及 F1 值更是均高于 99.58%。基于模型的角度,4 个实验中最佳模型是 BERT 模型,其实验四的准确率、召回率以及 F1 值都在 99.58%以上,而其他模型相对逊色。

表 3 为实验四(标题、关键字、摘要以及标签)5 种模型的时间对比。BERT_DPCNN 模型所用训练时间最短,仅为 11.32 s,准确率居中;BERT_RNN 模型所用运算时间最长为 0.23 s,准确率也最低为 97.60%,其训练时间较长。BERT 模型、BERT_CNN 模型和 BERT_RCNN 模型运算时间相同,均为 0.19 s,其中,BERT_RCNN 模型的训练时间居中,准确率也较低;BERT_CNN 模型的训练时间在仅次于 BERT_DPCNN 模型的较短时间下,准确率却居于次高,达到 99.15%;而 BERT 模型虽训练时间最长,但其准确率却高达 99.58%。结合各项评价指标分析,BERT_CNN 模型为本实验中效果最佳的模型。

实验表明,在区分医学的个例与非个例问题中,将深度学习模型用于文本分类时,由于模型选择不同,实验结果有所差异,以上 5 个模型在 4 组数据的对比情况下,基础的 BERT 模型在实验四的精确数据下的正确率与召回率都有显著的提升。综合时间和各项指标分析,BERT_CNN 模型在本实验四中预测效果最佳。BERT 和 CNN 在特征提取方面具有互补的优势,即 BERT 能够学习到丰富的词级和句子级特征表示,而 CNN 可以通过卷积和池化操作捕捉文本的局部特征,将它们结合起来,能够同时利用它们在特征提取方面的优势,从而提升模型对文本特征的表达能力。BERT 是一个大型模型,具有大量的参数,而 RNN 和 DPCNN 也相对较复杂。相比之下,BERT 与 CNN 的组合可以减少整体模型的参数量,降低计算复杂度,提高模型的训练和推理效率。

表 3 实验模型的时间对比

Table 3 Time comparison of experimental models

模型	准确率/%	训练时间/s	运算时间/s
BERT	99.58	36.10	0.19
BERT_CNN	99.15	11.39	0.19
BERT_DPCNN	98.18	11.32	0.18
BERT_RCNN	97.88	12.39	0.19
BERT_RNN	97.60	14.24	0.23

基于实验四的结果,本文继续对个例的不良反应种类进行研究。在区分个例是否为药物不良反应的问题中,对同一数据分别采用两种方法(A 模型为使用 BERT_CNN 模型进行不良反应种类分类,B 模型为对数据进行药品名称识别),4 种组合方法即 A 模型识别、B 模型识别、先 A 后 B 模型识别、先 B 后 A 模型识别,其中先 B 后 A 模型识别效果最佳。具体实验结果如表 4 所列。

表 4 个例的不良反应种类结果

Table 4 Recognition accuracy results

(%)

BERT	准确率	召回率
A	99.25	99.75
B	69.67	69.52
A-B	99.25	99.75
B-A	99.75	99.25

总的来说,在个例的不良反应种类判别实验中,先 B 后 A 模型识别方法的准确率和召回率最优。

结束语 与普通文本数据不同,中医药文本有其独特的表述。针对该问题,本文基于 BERT 及其组合模型,通过文本分类技术对比实验,建立对药物不良反应个例报道文献数据分类的高效快速方法,进而分辨出药物不良反应的类型。

实验结果表明,在区分医学的个例与非个例问题中,将深度学习模型用于文本分类时,由于模型选择不同,实验结果有所差异。其中,BERT 模型的分类准确率高达 99.75%,可对药物不良反应做出有效分类。

此项工作是构建药物不良反应文献数据库前期工作中的一部分,将有效提高药物不良反应文献采集、分析的效率。后续将继续对题录和全文数据进行基于 NLP 技术的知识抽取、知识库构建和知识发现工作。

参 考 文 献

[1] WU M Z. Extraction and Analysis of Relationships between

Drugs and Diseases from Medical Literature[D]. Shenyang: China Medical University, 2010.
[2] YANG Y, BAI Y N, ZHANG Y Y, et al. Bibliometric research on 323 cases of drug fever induced by antimicrobial[J]. Practical Pharmacy and Clinical Remedies, 2021, 24(8): 722-726.
[3] LI Y L, LV M, ZHANG M. Analysis of 11 cases of levothyroxine sodium-induced liver injury in literature[J]. Drug Evaluation Research, 2021, 44(7): 1508-1512.
[4] LI Y, XU X L, LIU Y, et al. Sirolimus-induced adverse reactions in children: a literature analysis [J]. Clinical Medication Journal, 2021, 19(9): 74-78.
[5] CHEN J Y, MA J, TONG G L, et al. Literature analysis of lung injury induced by oxaliplatin [J]. Chinese Journal of Hospital Pharmacy, 2021, 41(10): 1059-1063.
[6] CHEN J S. 《Chinese Journal of New Drugs》1998—2002 Comprehensive analysis of adverse drug reaction reports[J]. Chinese Journal of New Drugs, 2003, 12(12): 1046-1047.
[7] Circular of the State Drug Administration on Issuing the guiding principles for the collection and reporting of individual adverse drug reactions [EB/OL]. [2018-12-19]. https://www.cdr-adr.org.cn/drug_1/zcfg_1/zcfg_zdyz/202009/t20200924_47831.html.
[8] Adverse Drug Reactions Database[Z]. 2019-07-29.
[9] CANETE-SIFUENTES L, MONROY R, MEDINA-PEREZ M A. A Review and Experimental Comparison of Multivariate Decision Trees[J]. IEEE Access, 2021, PP(99): 1.
[10] KE B, KHANDELWAL M, ASTERIS P G, et al. Rock-burst occurrence prediction based on optimized Nave Bayes models[J]. IEEE Access, 2021, PP(99): 1.
[11] ABDELSALAM M M, ZAHRAN M A. A novel approach of Diabetic Retinopathy early detection based on Multifractal Geometry Analysis for OCTA Macular Images using Support Vector Machine[J]. IEEE Access, 2021, PP(99): 1.
[12] JALAL A, AHMED A, RAFIQUE A A, et al. Scene Semantic Recognition Based on Modified Fuzzy C-Mean and Maximum Entropy Using Object-to-Object Relations [J]. IEEE Access, 2021, 9(99): 27758-27772.
[13] CHEN K L, LU G, LI C W, et al. Obstructed Nearest Neighbor Query Under Uncertainty in the Internet of Things Environment[J]. IEEE Access, 2021, PP(99): 1-1.
[14] ZHANG W J, JIANG L X, ZHANG H, et al. A Two-Layer Bayes Model: Random Forest Naive Bayes [J]. Journal of Computer Research and Development, 2021, 58(9): 2040-2051.
[15] WANG S W, XU F, CHAO H T, et al. Research on fault classification of textile hot rolling mill based on support vector machine [J]. Modern Manufacturing Engineering, 2021(6): 116-121.
[16] JIE L, JIN T, PAN K, et al. An improved KNN text classification algorithm based on Simhash[C]// 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing(ICCICC), 2017.
[17] ALSALEH D, LARABI-MARIE-SAINTE S. Arabic Text Classification using Convolutional Neural Network and Genetic Algorithms[J]. IEEE Access, 2021, PP(99): 1-1.
[18] WU Z Y, BAI K L, YANG L R, et al. Review on Text Mining of Electronic Medical Record [J]. Journal of Computer Research and Development, 2021, 58(3): 513-527.
[19] TURNER C A, JACOBS A D, MARQUES C K, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus[J]. Bmc Medical Informatics & Decision Making, 2017, 17(1): 126.
[20] ZHANG B, SUN Y, LI M Y, et al. Medical Text Classification Based on Transfer Learning and Deep Learning[J]. Journal of Shanxi University(Natural Science Edition), 2020, 43(4): 947-954.
[21] ZHOU Y C, CUI Z F, FAN S P, et al. Deep-learning-based biomedical text classification[J]. Chinese Journal of Medical Library and Information Science, 2019, 28(11): 1-10
[22] YE H, ZHOU Y R, CAO D, et al. Intelligent Classification of Medical History in Chinese Medical Records Based on Deep Learning[J]. China Digital Medicine, 2019, 14(3): 41-43.
[23] HUANG L, DU C S. Application of recurrent neural networks in text classification[J]. Journal of Beijing University of Chemical Technology(Natural Science Edition), 2017, 44(1): 98-104.
[24] LUO J, CHEN L F. Sentiment classification of incomplete data based on bidirectional encoder representations from transformers [J]. Journal of Computer Applications.
[25] CHEN J, MA J, LI X F. A Short Text Classification Method Based on the Fused Text Features of Pre-trained Models[J]. Data Analysis and Knowledge Discovery, 2021, 5(9): 21-30.
[26] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv. 1810.04805, 2018.
[27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[J]. arXiv. 1706.03762, 2017.
[28] ZHANG H, JIA T Y, LUO F, et al. Study on Predicting Psychological Traits of Online Text by BERT[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(8): 10.
[29] YANG P, DONG W Y. Chinese Named Entity Recognition Method Based on BERT Embedding[J]. Computer Engineering, 2020, 46(4): 40-45, 52.
[30] SHAO Y F, LI H R, GU J H, et al. Extraction of causal relations based on SBEL and BERT model[J]. Database The Journal of Biological Databases and Curation, 2021. DOI: 10.1093/database/baab005.
[31] WANG Y, CHEN Y R, HAN Z L, et al. Short-Term Wind Speed Forecasting Model Based on Mutual Information and Recursive Neural Network[J]. Journal of Shanghai Jiao Tong University, 2021, 55(9): 1080-1086.
[32] ZHOU S Y, FU R, LI J. Zeus at HASOC 2020: Hate speech detection based on ALBERT-DPCNN [J]. FIRE (Working Notes), 2020: 195-201.

MENG Xiangfu, born in 1981, Ph. D, professor, Ph. D supervisor, is a senior member of CCF(No. 26143S). His main research interests include spatial big data query and analysis, artificial intelligence.

ZHU Yan, born in 1983, Ph. D, associate professor, postgraduate supervisor. His main research interests include data mining and knowledge organization of traditional Chinese medicine.