

黄双斌, 王梅嘉, 高浏洋. 基于 BERT_DPCNN 的牛类疾病问句分类研究[J]. 智能计算机与应用, 2025, 15(3): 140–144.
DOI:10.20169/j.issn.2095-2163.250320

基于 BERT_DPCNN 的牛类疾病问句分类研究

黄双斌, 王梅嘉, 高浏洋

(陕西科技大学 电子信息与人工智能学院, 西安 710021)

摘要: 为帮助养殖户实现牛类养殖的精准、科学化管理, 基于 BERT、TextCNN、TextRNN 模型, 研究牛类疾病的问句分类方法, 为构建面向牛类疾病的问答系统提供技术支撑。通过设计爬虫获取惠农网、百度贴吧等原始数据, 并对数据进行预处理, 获取了包含 5 056 条数据的数据集, 将数据进一步划分为定义、预防、病因、症状、治疗和诊断共 6 类, 以构建牛类疾病分类语料库。实验表明, BERT 模型在 6 类精度有 4 类不弱于其他模型, 在不同大小的数据集上预训练 BERT 模型在加权 $F1$ 值上均优于 TextCNN 和 TextRNN 模型, 与 BERT 其他变种模型进行了实验对比, BERT_DPCNN 模型比 BERT 模型加权 $F1$ 值高 0.3%, 考虑问答系统问句分类精确度要求高, 选取 BERT_DPCNN 模型作为问句分类模型。

关键词: 自然语言处理; BERT; 牛类疾病; 问答系统; 问句分类

中图分类号: TP391

文献标志码: A

文章编号: 2095-2163(2025)03-0140-05

Classification of bovine disease questions based on BERT_DPCNN

HUANG Shuangbin, WANG Meijia, GAO Liuyang

(School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China)

Abstract: In order to help farmers achieve accurate and scientific management of cattle breeding, this paper studies the classification method of questions for cattle diseases based on BERT, TextCNN and TextRNN models, and provides technical support for the construction of a question-answering system for cattle diseases. The crawler is designed to obtain the original data of Huinong network and Baidu Post Bar, and the data is preprocessed to obtain a dataset containing 5 056 pieces of data, and the data is further divided into 6 categories: definition, prevention, etiology, symptoms, treatment and diagnosis, so as to build a classification corpus of bovine diseases. Experiments show that BERT model is no weaker than other models in four of the six categories of accuracy. On data sets of different sizes, pre-trained BERT model is superior to TextCNN and TextRNN model in weighted $F1$ value. Experimental comparison is conducted with other BERT variant models. The weighted $F1$ value of BERT_DPCNN model is 0.3% higher than that of BERT model. Considering the high accuracy requirement of question classification in question answering system, BERT_DPCNN model is selected as the question classification model.

Key words: natural language processing; Bidirectional Encoder Representations from Transformers; cattle diseases; question answering system; question classification

0 引言

随着物质生活水平的提高, 居民对肉类的消费进入转型升级阶段, 牛肉由于富含营养备受人们青睐。然而国内肉牛养殖方式基本以个体户为主分散在郊区、农村等区域经营^[1], 农户缺乏科学饲养技术, 肉牛疫病时有发生, 降低了牛肉的品质与产量。

问答系统作为信息检索的高级形式, 以自然语言的方式回答用户提出的问题, 帮助其快速获取所需信息。因此, 构建面向牛类疾病的问答系统, 提供关于常见疾病的诊断、预防、以及治理等技术支持, 能够帮助养殖户实现精准科学养殖, 提升牛肉的产量与品质。典型的问答系统由问句分类、答案检索、答案抽取三个模块构成。其中, 问句分类的精度, 直接影

基金项目: 陕西省自然科学基金研究计划(2022JQ-175); 陕西省教育厅科学研究计划项目(22JK0303)。

作者简介: 黄双斌(1999—), 男, 硕士研究生, 主要研究方向: 自然语言处理, 知识图谱; 高浏洋(1998—), 男, 硕士研究生, 主要研究方向: 图像处理目标检测。

通信作者: 王梅嘉(1990—), 女, 博士, 讲师, 主要研究方向: 自然语言处理, 面向多智能体的软件工程。Email: 4672@sust.edu.cn。

收稿日期: 2023-09-03

哈尔滨工业大学主办 ◆ 专题设计与应用

响另外两个模块的准确率^[2]。目前,已有的文本分类方法分为传统的机器学习方法和深度学习方法,不同于传统机器学习方法提取浅层特征,深度学习方法能够提取文本的复杂特征,且具有更强的文本表示能力,而基于深度学习的文本分类主要过程为数据预处理、文本表示、深度学习模型和分类器。文本表示是文本分类的基础,近年来,词向量^[3]成为了文本表示的主流。Word2Vec 是谷歌发布的词向量训练工具,能够从大规模未经标注的语料中高效地生成词的向量形式,并提供了 CBOW 和 Skip-gram 这 2 种模型^[4]。Kim^[5]首先提出基于卷积神经网络(Convolutional Neural Network, CNN)的文本分类模型。Johnson 等学者^[6]提出了深度金字塔卷积神经网络(Deep Pyramid Convolutional Neural Networks, DPCNN)。该模型复杂度低,分类效果比浅层的卷积神经网络更为出色。Jordan^[7]提出循环神经网络(Recurrent Neural Network, RNN)。RNN 主要处理序列数据,但对于过长文本表现不佳。Schmidhuber^[8]提出了长短期记忆网络(Long Short-Term Memory network, LSTM)。LSTM 能有效解决 RNN 中的长期记忆依赖和梯度消失等问题,但单一的模型无法取得较好的分类效果。王海涛等学者^[9]利用 CNN 获取文本的局部特征,同时利用 LSTM 获取文本的上下文关联语义,再将特征融合进行重用。赵博轩等学者^[10]提出 CNN 和门控递归单元(Gated Recursive Unit, GRU)混合的文本分类模型,使用拼音作为特征表示,显著提高了分类准确率。这种方法也运用在了医学^[11]、机械^[12]、语言^[13]等领域的文本分类研究中。而在农业领域,杨国峰等学者^[14]提出了经预训练的 BERT(Bidirectional Encoder Representations from Transformers)应用在作物病害问句分类上。赵明等学者^[15]提出了双向门控循环单元神经网络(Bi-Directional Gated Recursive Unit, Bi-GRU)的番茄病虫害问句分类模型。该模型利用 Word2Vec 将句子中的词转换为具有语法、语义信息的词向量。综上所述,目前文本分类研究已经取得了一定的进展, Bert、CNN、RNN 等相关模型在已有的数据集上均表现出良好的效果。然而现如今在农业领域,关于牛类疾病的相关研究仍然较少,如何获取关于牛类疾病的数据集以及如何选择合适的问句分类模型仍然是值得研究的重要问题。

本文以牛类疾病问句分类为研究对象,通过爬取数据集,基于 Bert、TextCNN、TextRNN 等神经网络模型分别进行实验,研究适合于牛类疾病的问句分

类方法,为构建特定领域问答系统提供技术基础,帮助养殖户解决难以获取科学指导技术的问题。

1 材料与方法

1.1 数据收集及预处理

为获取牛类疾病领域的相关数据集和语料,本文于 2023 年 2 月至 2023 年 3 月,设计爬虫工具爬取惠农网农技学堂版块中专家问答、牛病吧和养牛吧等有关牛类疾病的问句,累计数据 8 000 条。

由于从网站上爬取的原始数据存在缺失、无意义数据、特殊符号、长空格和重复等问题,首先需要对数据进行预处理,这样才可以使模型进行有效识别并提取信息。

预处理数据后,获得 5 056 条有效数据。通过分析爬取的所有数据,并咨询相关领域专家,进一步将预处理后的数据分为 6 类:定义、预防、病因、症状、治疗和诊断,将数据按照 3 : 1 : 1 的比例划分为训练集、验证集和测试集。其中,每个集合包含类别数据见表 1,每个类别下的部分问句见表 2。

表 1 类别数量信息
Table 1 Class quantity information

数据集	定义	预防	病因	治疗	症状	诊断
训练集	343	335	646	278	714	717
验证集	114	112	215	93	238	239
测试集	115	111	215	92	239	240

表 2 不同类别的问句
Table 2 Different types of questions

标签名称	具体说明
定义	新生牛犊腹泻是什么?
预防	如何防治结节病?
病因	牛突然拉肚子是什么原因?
治疗	牛不吃不喝怎么办?
症状	牛结节病的症状有哪些?
诊断	牛身上起了个大包是结节病吗?

1.2 BERT_DPCNN 模型

BERT 是 2018 年谷歌 AI 研究院提出的一种自然语言处理预训练模型,由双层双向 Transformer^[16]编码器构成,其模型结构如图 1 所示。相较于 RNN、LSTM, BERT 可以做到并发执行,能解决远距离依赖问题。

BERT 模型以问句作为输入,输入被送到词嵌入层,包含位置向量、段向量和词向量。其中,位置向量就是每个字在句中的位置;段向量就是句子之

间的关系,是首句、还是末句;词向量则是以[CLS]为开头、句子间[SEP]分隔的向量。输出则是[CLS]表示的问句的语义表示,用于文本分类。

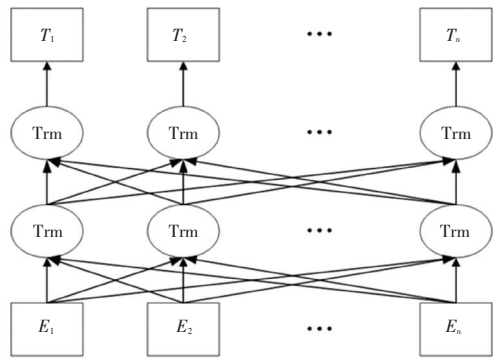


图 1 BERT 模型图

Fig. 1 BERT model diagram

为了增强对文本语义的表达能力,BERT 模型提出了 2 个预训练任务。一个是遮蔽语言建模 (Masked Language Model),对句中的部分词汇进行随机遮蔽,然后预测被遮蔽的词;另一个是下一句预测 (Next Sentence Prediction),让模型理解句子与句子之间的关系,从而更好地完成文本分类任务。

DPCNN 模型通过不断加深网络,可以抽取长距离的文本依赖关系,DPCNN 模型结构如图 2 所示。

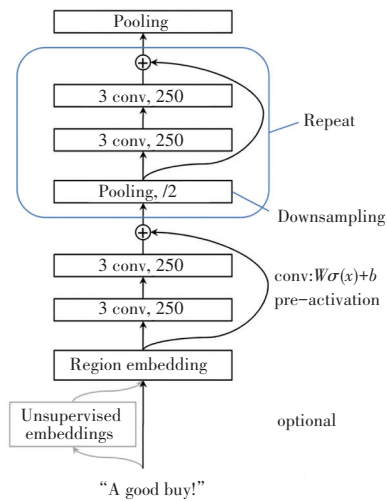


图 2 DPCNN 模型结构图

Fig. 2 DPCNN model diagram

DPCNN 模型分为 embedding 层,2 层等长卷积块,残差连接后进入 Repeat 结构、即重复最大池化层,使每个卷积核的维度减半,再接 2 层等长卷积块和残差连接,直到序列长度等于 1 后接入池化层。DPCNN 主要特点为减采样的设计,使每一个块进行步长为 2 的池化操作,如此一来计算量将减半且准确率不受影响;残差连接的设计使网络可以扩展为

深层网络进行参数的训练。

BERT_DPCNN 模型通过 BERT 和 DPCNN 模型的结合更细致地提取出长问句中的语义依赖关系,可以提高问句分类的准确率。

1.3 预训练词向量 TextCNN 模型

TextCNN 模型是将 CNN 应用到文本分类任务中,具有并行处理、模型参数少和训练速度快等优点,是一种非常优秀的文本分类模型。TextCNN 模型结构如图 3 所示。

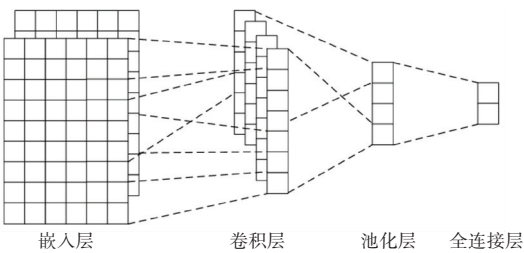


图 3 TextCNN 模型图

Fig. 3 TextCNN model diagram

模型结构分为嵌入层、卷积层、池化层和全连接层。在嵌入层中,对文本进行分词,通过预训练的词向量将自然语言数值化,以方便计算机处理;在卷积层中,对输入的词向量矩阵与卷积核进行卷积运算操作,用于特征的提取;在池化层中,进行最大池化操作降维,然后送到全连接层;最后,全连接层使用 dropout 来防止神经网络过拟合,并且输出各个类别的概率。

1.4 TextRNN 模型

TextRNN 模型能够捕获更长的序列信息,在该模型中使用双层 LSTM 结构,具体结构如图 4 所示。

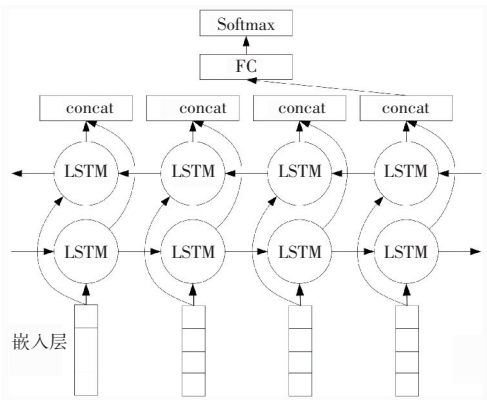


图 4 TextRNN 模型图

Fig. 4 TextRNN model diagram

图 4 中,最下面为嵌入层,使用预训练词向量对输入文本进行向量化,再通过前后向 LSTM 层,接下来对输出进行拼接,通过全连接层后则续接 Softmax

函数输出各个类别的概率。

目前,除了经典的 Bert、CNN、RNN 模型外,BERT_DPCNN、BERT_CNN、BERT_RNN 等组合模型也已陆续应用到文本分类中,BERT 变种模型是将 BERT 模型输出的[CLS]稍加改变、使其符合下一个模型的输入,最后得到输出结果。

2 问句分类实验

2.1 问句分析

对经过预处理的问句进行长度统计(包含标点),最短的问句长度为 2,最长的为 300,具体数据见表 3。

表 3 问句长度信息								
Table 3 Question length information								
长度	20	40	60	80	100	150	200	300
数量	2 893	1 371	469	180	71	58	8	7

由表 3 可知,84%的问句长度都在 40 字以内,对 40 字以上问句长度类别情况进行统计,在定义、预防和症状三类中长问句较为少见。分析可知,长句主要存在于病因、治疗和诊断三类中。

2.2 实验环境

程序运行平台的 CPU 为 Intel(R) Core(TM) i5-8300H CPU,GPU 为 GeForce GTX 1050,操作系统为 Win10,Python 版本为 3.7,PyTorch 版本为 1.7,内存容量为 16 GB。

2.3 实验模型及参数

本文以 TextCNN、TextRNN、Bert 作为实验模型。对于 TextCNN 模型,使用了腾讯中文预训练词向量,分词工具为 jieba。对于 TextRNN 模型,LSTM 层数设置为 2。对于 Bert 模型,预训练模型使用的是中文 base 模型。

为了得到针对于每个模型的最优参数,以获得每个模型在当前实现环境下的最优分类结果,选取部分数据进行初始调参。由表 2 可知,输入长度需在 100~130 之间,以尽可能获取问句的特征,从而减少训练时间。经过多轮微调得到的模型超参数见表 4。

表 4 各模型参数				
Table 4 Each model parameter				
模型名称	批处理	输入长度	迭代次数	学习率
BERT	8	128	3	1e-5
TextCNN	32	115	30	5e-4
TextRNN	32	128	30	2e-3

2.4 评价指标

为了有效且全面地评估模型性能,实验采用的评价指标主要包括精确率(Precision)、召回率(Recall)、 F_1 值($H-mean$ 值)和加权平均(weighted avg)。其中,精确率指分类器预测为正样本且预测正确的样本占全部预测为正样本的比例;召回率指分类器预测为正样本且预测正确的样本占有真实正样本的比例; F_1 值表示综合精确率和召回率的一个评估指标(精确率和召回率的调和平均值);加权平均表示每个类别的 P 、 R 和 F_1 的加权平均,权重为各类别占样本数的比例,其中 F_1 的加权平均公式为:

$$F_{weight_avg} = \frac{\sum_{i=1}^n F_{1i} rate_n}{n} \quad (1)$$

其中, F_{1i} 表示各类别的 F_1 值, $rate_n$ 表示权重。本文中涉及的所有 P 、 R 和 F_1 值均为加权平均。

2.5 结果分析对比

为确保实验数据的准确性,数据集划分采用 5 次随机划分,再对 5 次实验结果取平均值;分别以 BERT 模型、TextCNN 模型、TextRNN 模型训练数据集,得到各类别的 F_1 值和加权 F_1 值见表 5。

表 5 模型在各类别权重 F_1 值						
Table 5 Weight F_1 value of the model in various categories						
模型	定义	预防	病因	症状	治疗	诊断
BERT	99.56	99.56	97.89	97.87	97.87	97.32
TextCNN	99.12	99.56	98.14	98.40	97.25	97.30
TextRNN	97.39	98.68	97.42	96.30	97.03	95.18

由表 5 可知,BERT 模型在定义、预防、治疗和诊断四类中取得了最优的结果,TextCNN 则在预防、病因和症状中取得了最优的结果,而 TextRNN 却均逊色于 BERT 和 TextCNN 模型。

在保持各模型参数不变的情况下,改变数据集大小(500、2 500、5 056)进行实验,测试结果见表 6。

由表 6 可知,数据集的规模对 BERT、TextCNN 和 TextRNN 模型的分类结果均有较大影响,且在不同的训练集规模下 BERT 模型都比 TextCNN 和 TextRNN 模型效果好,在 5 000 条数据下 BERT 在权重 F_1 值上取得了 98.12%的最优效果,TextRNN 模型在更小数据集规模下表现远不如 BERT 和 TextCNN 模型。研究表明问答系统问句分类对正确率有着较高的要求,所以训练模型时应提高问句数据集数量以及质量。

表 6 不同大小数据集下各模型结果

Table 6 Results of each model in different datasets							%		
模型	P			R			F ₁		
	500	2 500	5 000	500	2 500	5 000	500	2 500	5 000
BERT	96.20	96.77	98.15	96.04	96.64	98.12	96.05	96.65	98.12
TextCNN	94.52	96.11	98.03	94.06	96.05	98.02	93.99	96.05	98.02
TextRNN	43.34	90.75	96.85	46.53	90.32	96.83	38.51	90.35	96.83

为了进一步选取最优模型,本文对经典 Bert 模型以及 Bert 模型的改进模型进行了对比实验,结果见表 7。

表 7 BERT 及变种模型实验结果

Table 7 Experimental results of BERT and variant models				%
模型	P	R	F ₁	
BERT	98.15	98.12	98.12	
BERT_CNN	98.44	98.42	98.42	
BERT_RNN	97.75	97.73	97.73	
BERT_RCNN	98.24	98.22	98.22	
BERT_DPCNN	98.45	98.42	98.42	

由表 7 可知,在当前实验环境下,BERT_DPCNN 和 BERT_CNN 模型在加权 F_1 值上表现最好,在精确率(P)中,BERT_DPCNN 表现最好,在召回率(R)中 BERT 的 CNN 和 DPCNN 变种模型表现最好,除了 BERT_RNN,其他变种模型均有较少提升。考虑到问答系统问句分类精确度要求较高,研究后则选择 BERT_DPCNN 模型作为最终的问句分类模型。

3 结束语

本文通过抓取面向牛类疾病领域的数据,基于 Bert、TextCNN、TextRNN 模型,研究牛类疾病问句分类。实验表明,在相同实验环境下,BERT 模型在 6 类问句中有 4 类不弱于其他模型,加权 F_1 值比 TextCNN 模型高 0.1%,比 TextRNN 高 1.29%。在自制数据集不同规模下 BERT 模型的精确率、召回率和 F_1 值均高于 TextCNN 和 TextRNN 模型。实验结果还表明,问句规模和分类模型的选择对牛疾病问答系统问句分类效果有较大影响。在 BERT 与 BERT 变种模型对比中,BERT_DPCNN 模型在精确率、召回率和加权 F_1 值中结果均要优于其他 BERT 变种模型,由于问答系统问句分类精度直接影响后 2 个模块的精确度,所以选择 BERT_DPCNN 作为最终的问答系统问句分类模型。在数据标注期间,部

分数据存在多标签情况,未来将进一步研究问句多标签分类,全面提升问句分类精度,扩大问句类别范围,满足更多的用户对牛类疾病的知识需求。

参考文献

[1] 布炳鹁·阿布拉,张立昌,李明. 中国肉牛养殖业现状与发展策略研究[J]. 黑龙江粮食,2022(8):91-93.

[2] 郑幸福,刘挺,秦兵,等. 自动问答综述[J]. 中文信息学报,2002,16(6):46-52.

[3] DAVIS C A,FONSECA F T. Assessing the certainty of locations produced by an address geocoding system[J]. Geoinformatica,2007,11(1):103-129.

[4] MIKOLOV T,KARAFIAT M,BURGET L,et al. Recurrent eural network based language model[C]// Proceedings of the 11th Annual Conference of the International Speech Communication Association. Chiba,Japan:dblp,2010;1045-1048.

[5] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv,1408.5882,2014.

[6] JOHNSON R,ZHANG Tong. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. ACL,2017;562-570.

[7] JORDAN M I. A parallel distributed processing approach[J]. Advances in Psychology,1997,121:471-495.

[8] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural Networks,2015,61:85-117.

[9] 王海涛,宋文,王辉. 一种基于 LSTM 和 CNN 混合模型的文本分类方法[J]. 小型微型计算机系统,2020,41(6):1163-1168.

[10] 赵博轩,房宁,赵群飞,等. 利用拼音特征的深度学习文本分类模型[J]. 高技术通讯,2017,27(7):596-603.

[11] 温俊,彭丽,郭珊,等. 基于神经网络文本分类方法的甲状腺结节恶性风险评估[J]. 中国数字医学,2023,18(3):73-79.

[12] 董丽丽,魏胜辉. 一种面向机械领域文本分类器的设计[J]. 微电子学与计算机,2012,29(4):142-145.

[13] 才智杰,道吉扎西. 藏文文本分类特征基元选择[J]. 中文信息学报,2023,37(1):64-70.

[14] 杨国峰,杨勇. 基于 BERT 的常见作物病害问答系统问句分类[J]. 计算机应用,2020,40(6):1580-1586.

[15] 赵明,董翠翠,董乔雪,等. 基于 BIGRU 的番茄病虫害问答系统问句分类研究[J]. 农业机械学报,2018,49(5):271-276.

[16] 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. Long Beach, USA: NIPS Foundation,2017;6000-6010.