doi:10.6041/j.issn.1000-1298.2025.10.055

主题特征与领域预训练模型融合的农业问句分类方法

熊蜀峰[1]　沈佳龙[1]　许恒瑞[1]　王丙坤[2]

(1.河南农业大学信息与管理科学学院,郑州 450046;2.郑州财经学院信息工程学院,郑州 450049)

摘要:问句分类的准确性对问答系统至关重要,尤其农业领域具有多样化的问句类别和复杂的语义内容情况下,有效分类用户提出的问句是开发智能农业问答系统的关键任务。针对如何在农业领域中对用户提出的问句进行有效分类的问题,本文提出了一种结合领域 BERT(Bidirectional encoder representational from transformers)预训练与主题模型的农业问句分类方法 TAB(Topic model with agriculture BERT for agricultural text classification),以提高分类性能。为了应对农业领域数据的专业性和复杂性,开发了专用的预训练语言模型 AgCBERT(Agriculture BERT),用于文本特征向量表示。此外,采用 LDA(Latent dirichlet allocation)模型表示主题特征向量,作为文本向量的补充。TAB 方法最后将上述 2 类特征进行线性组合后输入全连接网络进行分类预测。在真实农业问句数据集上的实验结果显示,本文方法的 F1 值达到了 72.72%,显著优于常见的文本分类模型。结果表明,融合 AgCBERT 与主题概率模型的 TAB 方法在农业问句分类中表现卓越。

关键词:农业领域;问句分类;预训练语言模型;主题模型;BERT

中图分类号:S43;TP391.1　　**文献标识码**:A　　**文章编号**:1000-1298(2025)10-0606-08　　**OSID**:

Agricultural Question Classification by Fusing Topic Features with Domain Pre-trained Models

XIONG Shufeng[1]　SHEN Jialong[1]　XU Hengrui[1]　WANG Bingkun[2]

(1. *College of Information and Management Science, Henan Agricultural University, Zhengzhou* 450046, *China*
2. *School of Information Engineering, Zhengzhou College of Finance and Economics, Zhengzhou* 450049, *China*)

Abstract: The accuracy of question classification was crucial for question-answering systems, particularly in the agricultural domain where question categories were diverse and semantic content was complex. Effectively classifying user-submitted questions was a key task in the development of intelligent agricultural question-answering systems. To address the issue of effectively classifying user-submitted questions in the agricultural domain, an agricultural question classification method (TAB, topic model with agriculture BERT (AgCBERT) for agricultural text classification) was proposed, combining domain-specific BERT pre-training with topic modeling to improve classification performance. To manage the professional and complex nature of agricultural data, a specialized pre-trained language model, AgCBERT, was developed for text feature vector representation. Additionally, the LDA model was employed to represent topic feature vectors as a supplement to the text vectors. The TAB method linearly combined these two types of features and input them into a fully connected network for classification prediction. The efficacy of this approach was underscored by experimental results obtained from a real-world agricultural question dataset. The method achieved F1 score of 72.72%, which was substantially higher than those achieved by commonly used text classification models, indicating a significant improvement in classification performance. The results demonstrated that the TAB method, which integrated AgCBERT with a topic probability model, excelled in agricultural question classification. This integration allowed for a more nuanced understanding and processing of complex question types, leading to more accurate classifications.

Key words: agricultural domain; question classification; pre-trained language models; topic model; BERT

收稿日期:2024-06-29　修回日期:2024-07-19

基金项目:教育部人文社科项目(19YJCZH198)

作者简介:熊蜀峰(1980—),男,教授,博士,主要从事文本分析在农业领域应用研究,E-mail:xsf@whu.edu.cn

通信作者:王丙坤(1977—),男,教授,博士,主要从事文本分析技术研究,E-mail:wangbingkun@zzife.edu.cn

0　引言

中国作为农业大国，推动农业科技发展是实现农业生产现代化的重要途径。随着科技水平的不断提升，农业的组织形式和生产方式逐渐变化，农民对农业技术服务的需求也在不断增长。当前，许多农技问答平台可以为用户提供技术指导，但这些平台需要准确分类用户提出的问句，才能有效匹配答案或推荐专家。然而，许多农民缺乏专业知识，难以准确分类他们的提问，这直接影响了技术指导的准确性，降低了问答系统的整体效果。在全自动问答系统中，智能系统需要首先理解问句，再生成答案，因此，问句分类是问句理解的关键步骤之一。有效分类用户问句是实现智能问答系统的关键技术，直接影响答案检索和提取的准确性。

问答系统中有关问句分类的研究已经取得了一些成果，从最初的机器学习方法，到目前主流的深度学习方法都取得了很好的分类性能。在机器学习领域，支持向量机（Support vector machines，SVM）[1]、K 最近邻（K-nearest neighbor，KNN）[2]、隐马尔可夫模型（Hidden Markov model，HMM）[3]和朴素贝叶斯模型（Naive Bayes model，NBM）[4]等方法已经证明了分类的有效性。然而，由于这些模型依赖于人工定义的特征，并且还存在与数据稀疏性相关的问题，它们的泛化效果并不理想。相比之下，卷积神经网络（Convolutional neural network，CNN）[5]、循环神经网络（Recurrent neural network，RNN）[6]、长短期记忆（Long short-term memory，LSTM）[7]和门循环单元（Gated recurrent unit，GRU）[8]等深度学习模型的分类结果较为理想。不过，它们也存在一定局限性。例如，CNN 未考虑文本的语序，不能获取上下文信息；RNN 虽然能够记忆上下文信息，但它长期依赖学习特征，容易出现梯度弥散的问题，并且 RNN 对局部的关键信息提取不够敏感。为了更好地提取文本关键信息，注意力机制[9]被广泛应用于文本分类任务，其通过模仿人的注意力分配机制来计算不同词向量的权重，使关键词语获得更高的权重，从而获得高质量的文本特征信息。在此基础上，研究者又提出了 BERT 预训练语言模型[10]，显著提高了文本分类的性能。然而，农业领域是一个庞大的类别体系，通用的预训练模型仍无法很好地表示复杂的农业文本语义信息。此外，农业问句大多表现为短文本形式。其语义的稀疏性也是模型设计时需要考虑的因素之一。

综上，为解决农业领域问句分类任务中数据专业、复杂且文本短小的问题，本文提出 TAB 方法，通过引入 2 个模块来解决：农业领域专用预训练模型和基于主题模型的主题特征向量表示。由于农业领域的数据具有特定的语义信息，依赖于通用语料库的标准 BERT 模型在这方面的问句分类任务中表现并不理想。为克服这一问题，在采集大量农业文本的基础上训练一个专用的农业领域语言模型 AgCBERT 来替代标准的 BERT 模型。此外，为克服短文本语义信息固有的数据稀疏性，同时引入主题概率模型 LDA[11]来表示文本的主题特征，从而更好地捕捉和应用农业领域的专业语义信息，并与 AgCBERT 输出的文本语义表示进行融合，最终通过全连接网络对农业领域的问句进行分类。

1　问句分类与预训练语言模型

1.1　问句分类

问句分类是自然语言处理中的重要应用，也是问答系统中的关键步骤。其核心是将问句映射到预先给定的主题类别，这一过程包括文本特征提取和分类模型训练。问答系统中的问句分类是一项经典的自然语言处理任务。许多学者对此进行了研究，并提出了许多解决方案。从传统的 SVM[12]、结合词频和反文档的 TF - IDF（Term frequency-inverse document frequency）算法[13]、利用空间聚类的 KNN 算法、基于条件独立假设的 NBM 等机器学习方法，到 CNN、RNN、LSTM、GRU 等深度学习方法，再到 ELMo（Embeddings from language models）、OpenAI GPT（OpenAI generative pre-trained transformer）、BERT 等预训练神经网络语言模型，都取得了长足进步。这些深度学习方法具有自动提取特征的能力，可以使用大量数据进行训练，从而提高问句分类的性能。

针对特定的农业领域，少数研究者开展了将语言模型用于农业问答系统的相关研究，但仍处于起步阶段。段青玲等[14]研究了基于支持向量机的文本分类。魏芳芳等[15]和赵燕等[16]利用农业专业词典，基于规则和机器学习相结合的方法构建了一个问句分类模型。赵明等[17]建立了基于 Word2vec 和 LSTM 的分类模型，还提出了基于 BiGRU（Bidirectional gated recurrent unit）的番茄病虫害问答系统的问句分类模型[18]。金宁等[19]根据词的重要性对文本词向量进行加权，建立了一个结合 BiGRU 和多尺度卷积神经网络的农业问句分类模型。冯帅等[20]提出了一种基于深度卷积神经网络的水稻知识文本分类方法，该方法改进了残差模块的设计，并使用 CapsNet（Capsule network）进行分类。王郝日钦等[21]使用专门的农业词库对文本进

行向量编码,引入注意力机制强化关键词特征,构建了水稻问答系统中问句的自动分类模型。随着BERT等预训练语言模型的出现,一些学者利用BERT训练粒度更细的动态词向量,在特定领域的文本分类任务中取得了较好的效果[22-25]。然而,将BERT模型直接应用于农业领域仍然面临领域适应问题。

1.2 预训练语言模型

自Transformer模型问世以来,基于Transformer的预训练语言模型(Pre-trained language model, PLM)的研究急剧增多[26]。这些模型通过自监督学习在大型文本语料库上预训练,学习大量与自然语言处理(Natural language processing, NLP)任务相关的常识性知识[27]。根据所使用的训练语料库类型,PLM大致可分为通用预训练语言模型和特定领域预训练语言模型。

通用预训练语言模型的第一类典型模型是基于自编码方法的语言模型,如BERT及其变体RoBERTa(A robustly optimized BERT approach)[28]、alBERT(A lite BERT)[29]和ELECTRA(Efficiently learning an encoder that classifies token replacements accurately)[30]等,主要用于自然语言理解。另一类方法是自回归预训练技术,如GPT-1和ELMo[31],主要用于生成式自然语言处理任务。OpenAI的GPT-1在跨领域的下游任务(包括一些零样本任务)中表现出色,GPT-2[32]和GPT-3[33]等后续版本表现更加出色。此外,针对中文的预训练模型如百度的ERNIE(Enhanced representation through knowledge integration)[34]和哈尔滨工业大学与科大讯飞联合实验室的BERT-wwm(BERT with whole word masking)[35],也取得了良好效果。

特定领域的预训练语言模型针对特定领域(如金融或医疗)进行训练,以便更好地理解该领域使用的语言。这些模型由大量数据建立而成,提高了特定领域自然语言处理任务的准确性。在生物医学领域,BioBERT(Biomedical language representation model based on BERT)[36]利用同行业的文章和摘要进行预训练,以改进生物医学任务,它通过提供见解和解决方案帮助用户理解复杂的生物医学文献。同样,SCIBERT(Scientific BERT)[37]结合生物医学和计算机科学的论文来评估与生物医学行业相关的NLP任务。通过学习专业领域术语,它能够提供更准确的结果。在金融领域,FinBERT(Financial BERT)[38]利用金融新闻和文章进行预训练任务,以评估捕捉金融领域特有的语言知识和语义信息的NLP任务。对于金融从业人员来说,该工具可以详细分析和深入了解金融数据。

2 方法和设置

2.1 数据获取与处理

本文采用的数据包括两部分,分别用于AgCBERT模型预训练和问句分类模型训练。

预训练AgCBERT语言模型所使用的文本语料库来自不同的农业主题网站,分为4种类型,包括农业技术文章(TA)、在线农业问答库(QA)、农业新闻(NEWS)和农业科技专家在线知识库(KB)。将收集到的原始数据去除HTML标记,转换成纯文本文档。在下句预测(NSP)任务中,为了确保句子上下文之间的语义相关性,使用分段的独立段落作为原始训练样本。训练语料的统计结果如表1所示。该语料库由4 103 994个样本(段落)组成。在NSP训练中,使用句号、问号和感叹号等标点符号将段落分割成句子。为防止连续标点符号或无意义文本的影响,删除了少于5个字的句子。在训练过程中,按比例98∶1∶1将数据集分为训练集、测试集和验证集。

表1 语言模型预训练语料的统计结果

Tab.1 Statistical results of language model pre-training corpus

来源	样本数量	类别
农业网(http://www.agronet.com.cn)	49 844	TA
会农网(https://www.huinong.net)	265 564	TA,QA
农村网(http://www.nongcun5.com)	77 430	NEWS,QA
农业站(http://www.nongyezhan.com)	13 320	QA,TA
农业百科(https://www.nongyebaike.cn/)	20 137	KB,TA
五谷(http://nst.wugu.com.cn)	47 507	NEWS
农业科技网络书屋(http://zjzx.cnki.net/)	39 638	KB
中国农技推广信息服务平台(http://njtg.nercita.org.cn)	3 590 554	QA,TA

问句分类所使用的数据来自中国农技推广信息服务平台网站(http://njtg.nercita.org.cn)的问答板块,获取到的原始数据经过去重等预处理后,共有15个类别的农业问句,其中包含120 969个句子,按比例8∶1∶1将数据集分为训练集、测试集和验证集,训练集共96 775个句子,测试集共12 097个句子,验证集共12 097个句子,各类别数量如表2所示。

2.2 TAB方法

TAB方法的核心模块由农业领域预训练语言模型和主题特征表示组成,以及二者融合后的分类决策层。农业领域预训练语言模型AgCBERT目的是精细地进行农业领域文本编码。在此基础上,使用LDA模型提取问句的主题特征向量,与AgCBERT得到的文本向量线性结合,学习全局语义

表示，得到新的向量表示，从而提高模型的表示能力。本文提出的 TAB 模型结构如图 1 所示。

表 2　文本分类数据集各类别数量

Tab. 2　Text classification dataset

类别	训练集	测试集	验证集
品种选择	7 954	994	994
病虫害和杂草	7 947	993	994
土壤和肥料	7 922	991	990
栽培管理	7 918	990	990
市场营销	11 839	1 480	1 480
收获与加工	7 934	991	992
储存和保鲜	7 774	972	971
农业机械	4 762	596	595
其他种植	7 916	989	990
良种培育	2 659	332	332
动物疾病	7 801	975	975
饲料营养	2 107	263	264
农业管理	7 950	994	994
屠宰与加工	718	90	89
其他农业	3 574	447	447
总计	96 775	12 097	12 097

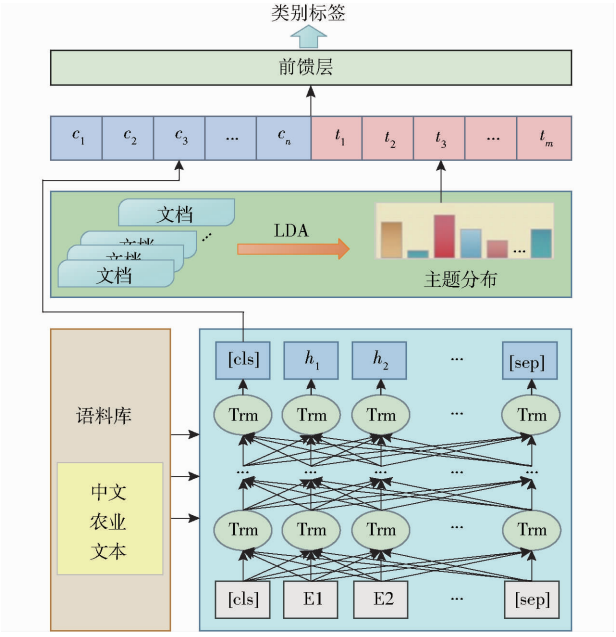


图 1　TAB 模型结构图

Fig. 1　TAB model diagram

2.2.1　AgCBERT

AgCBERT 是将 BERT 模型创新性地应用于中文农业文本的一种尝试，旨在学习农业文本的深层次表征。这将极大地促进中文农业语料库的构建，并有助于应对不断增长的农业文本量带来的挑战。模型沿用原始 BERT 模型的架构，其基础结构基于双向 Transformer 编码器，模型通过屏蔽语言模型 MLM(Masked language model)和下句预测 NSP(Next sentence prediction)2 个无监督任务进行预训练。在 MLM 任务中，输入序列中的部分字符会按照模型大小的比例被随机屏蔽，其目的是根据上下文来预测被屏蔽的单词，从而学习深度双向表征。相反，在 NSP 任务中，BERT 模型会处理句子对，并判断给定的 2 个句子是否连续，从而捕捉这些句子之间的关系。在对该问句分类任务进行 BERT 微调时，只需要修改超参数即可。最终训练好的 AgCBERT 采用了一个有 12 层、768 个隐藏单元和 12 个自注意头的 BERT 架构，并使用约 1.02×10^8 个参数进行预训练。由于 AgCBERT 与 BERT 兼容，因此无需重建词汇表，之后加入检查点来提高训练效率，减少训练过程中无关因素的影响。这种方法避免了从零开始进行模型的预训练，减少了训练时间，并提高了模型对指定领域[39]的适应性。该操作的主要目的并不是采用更复杂的表示方法，而是为研究人员提供一个可靠的参考点。预训练的过程如图 2 所示。AgCBERT 模型在中文农业文本语料库上进行训练。训练的轮数设置为 30，学习率为 4×10^{-5}。

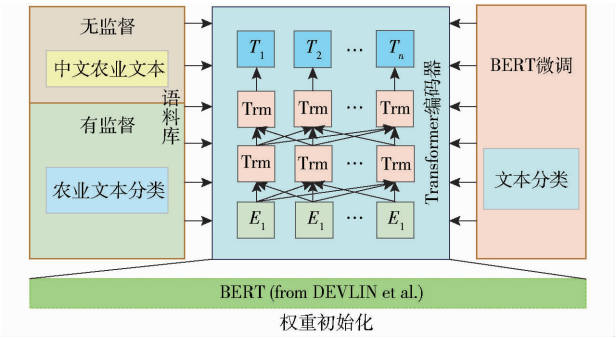


图 2　AgCBERT 的预训练和微调

Fig. 2　Overview of pre-training and fine-tuning of AgCBERT

2.2.2　LDA

LDA 是一种经典的主题建模技术，常用于文本主题分析相关任务。作为一种无监督学习方法，LDA 利用词袋模型将文档转换为词频向量，便于分析和建模。LDA 假设每篇文档包含多个主题，每个词都与一个或多个主题相关，通过选择主题和关联词的迭代过程生成文档内容。LDA 的核心任务是从文档集中发现基本主题，并识别与这些主题相关的词，进而将主题的概率分布分配给每个文档。这使得 LDA 能够对文本进行分类，揭示数据中的隐藏模式和结构，成为自然语言处理和文本分析的基础工具。

LDA 的输入是一组文档，被表示为词袋模型，即每个文档由一组词构成，可以计算得出每个词的出现次数。LDA 的输出是一组主题，用来描述一个文档中与该主题相关的内容，每个主题都是一个用于描述与该主题相关的词的概率分布。同时，每个文档都有一个主题概率分布，用于描述该文档感兴

趣的主题。因此,LDA 的输出主题由一组词组成,计算出每个词的出现次数,用于共同描述文档的内容。通过研究每个主题的概率分布,可以确定每个文档感兴趣的主题。在本实验中,LDA 的输入是一组 Documents 和 m 个预先设定的主题,输出是一个 m 维的文档-主题分布向量,每一维表示该文档包含每个主题的概率。这个向量反映了文档与主题之间的关系,可以用于进一步分析和应用。具体公式为

$$T = \mathrm{LDA}(D) \tag{1}$$

式中 D——输入的文档

T——输出的主题分布

LDA( )——将输入文档集合 D 经过 LDA 建模与推断过程,得到每个文档的主题概率分布

每个文档最后分别有一个 m 维的主题分布向量 $\boldsymbol{z}$。$\boldsymbol{z}$ 的表达式为

$$\boldsymbol{z} = \sum_{i=1}^{m} t_i \tag{2}$$

式中 t_i——文档在第 i 个主题上的概率

输入到 LDA 的文档最终会得到一个文档数量乘以 m 维向量的矩阵。

2.2.3 特征融合方法 TAB

在 TAB 方法中,AgCBERT 是专门为农业领域设计的预训练语言模型。因此,其输入和输出格式都与 BERT 模型一致。在文本数据预处理阶段,需要在文本序列的开头添加特殊标记“[CLS]”,同时在句子分隔符处插入标记“[SEP]”,以表示句子的连接性,这有助于表示文本或句子之间的关系。AgCBERT 模型的输出是一个向量表示,它包含了与输入的每个词相对应的完整语义信息,计算式为

$$\boldsymbol{E} = \sum_{i=1}^{n} \boldsymbol{E}_i \tag{3}$$

$$\boldsymbol{h}_i = \mathrm{AgCBERT}(\boldsymbol{E}_i) \tag{4}$$

式中 $\boldsymbol{E}_i$——输入句子中的第 i 个词向量

$\boldsymbol{h}_i$——输出句子中的第 i 个词向量

n——

输出的一组 $\boldsymbol{h}_i$ 组成一个代表句子的 i 维向量 $\boldsymbol{c}$,输入 AgCBERT 模型的句子集合,最后得到一组 i 维向量 $\boldsymbol{o}$,$\boldsymbol{o}$ 的表达方式为

$$\boldsymbol{o} = \sum_{i=1}^{n} \boldsymbol{c}_i \tag{5}$$

最后,将 AgCBERT 模型输出的 y 与 LDA 模型输出的 $\boldsymbol{z}$ 进行线性组合,作为 TAB 模型的最终输出。其中,LDA 输出的主题向量是一个 m 维的向量,它表示了文本中涉及的 m 个主题的维度,而 BERT 输出的向量是一个 768 维的向量,表示了文本中所有词的语义信息。将这 2 个特征向量拼接起来,可以得到一个新的特征向量,维数为 $m+768$。将这个新的特征向量作为文本的特征,用于问句分类任务。最简单的方式是直接将原始特征值拼接用作最终的文本表示,输入到分类器进行预测。

由于 LDA 输出的主题权重是一个密集向量,而 AgCBERT 输出的嵌入向量是一个稀疏向量,因此,更好的做法是在拼接 2 个向量时考虑两类向量的差异性。在本文提出的 TAB 模型中,通过超参数 λ 将主题特征向量与文本向量进行融合,公式为

$$f = \mathrm{concatnate}(\boldsymbol{o}, \lambda \boldsymbol{z}) \tag{6}$$

式中 f——输入文本的综合语义信息

concatnate( )——向量拼接操作,将两个向量合并成一个新的向量

加权融合后的特征 f 被输入到全连接层,进一步学习不同特征之间的权重关系,从而更好地捕捉文本的语义信息,提高问句分类的准确性。最终通过 softmax 层进行分类预测

$$y = \mathrm{softmax}(\mathrm{feedforward}(f)) \tag{7}$$

式中 y——输入文本在各分类标签上的概率分布

softmax( )——归一化函数,将前馈网络的输出映射为多类别的概率分布

feedforward( )——前馈神经网络(全连接层),用于学习特征之间的非线性关系

在实验中,连接主题特征向量和文本向量的超参数 λ 值设置为 5,LDA 预定义的主题数 m 设置为 20,经过多次超参数的调整优化,将 require_improvement 设置为 1 000,num_epochs 设置为 100,batch_size 设置为 128,pad_size 设置为 32,learning_rate 设置为 5×10^{-5},hidden_size 设置为 768。

2.3 对比模型

为了验证本文提出的模型在问句分类任务上的可行性和准确性,使用了一些先进的文本分类模型与 TAB 模型进行对比实验。

(1)BERT:使用 BERT 预训练模型作为输入层,并输入经过处理的问句序列以获得相应的输出。在 BERT 模型的输出层后面增加了一个分类层,用于将模型的输出映射到预定义的类别。

(2)BERT-TEXTCNN:使用 BERT 作为特征提取器,TextCNN 作为分类器。在处理过程中,使用 BERT 预训练模型对输入文本进行编码,然后使用 TextCNN 模型对生成的嵌入结果进行分类。

(3)BERT-CNN:使用 BERT 预训练模型对输入文本进行编码以获得嵌入,然后由 CNN 模型获得嵌入并预测输出标签。

(4)BERT－RNN：使用 BERT 预训练模型对输入文本进行编码以获得嵌入，然后由 RNN 模型获得嵌入并预测输出标签。

(5)Topic－BERT：使用 BERT 预训练通用模型代替农业领域的 AgCBERT 预训练语言模型，然后结合 LDA 提取的主题特征向量进行分类预测。

2.4　评价指标

本文使用精确率、召回率、F1 值和正确率来衡量模型的性能。其中，精确率(Precision)是指分类器预测为正例的样本中，真正为正例的比例，即针对预测结果，真正正确的样本数量。召回率(Recall)是指所有真正为正例的样本中，被分类器预测为正例的比例，即针对实际情况，正例被正确预测出来的样本数量。F1 值是综合考虑精确率和召回率的指标，是精确率和召回率的调和平均数，用于评估分类器的综合性能。正确率(Accuracy)是指正确分类的样本数占总样本数的比例。此外，由于每个任务都有多个类别，样本量不平衡，因此，使用宏观平均法来衡量模型在每个任务上的总体性能。

3　实验结果

首先比较了所提出的 TAB 模型与其他农业问句分类模型的性能。结果表明，TAB 模型的关键指标超越了所有其他方法，其精确率、召回率、F1 值和正确率分别达到了 72.95%、73.41%、72.72% 和 73.41%。AgCBERT 与 LDA 相结合的各类别实验结果如表 3 所示。不同模型对比实验结果如表 4 所示。

表 3　问句分类实验结果

Tab.3　Results of question classification　%

类别	精确率	召回率	F1 值
品种选择	70.51	76.96	73.59
病虫害和杂草	81.55	93.05	86.92
土壤和肥料	88.09	79.11	83.36
栽培管理	65.42	68.99	67.16
市场营销	79.34	91.35	84.92
收获与加工	75.40	61.55	67.78
储存和保鲜	86.46	82.10	84.22
农业机械	82.53	80.03	81.26
其他种植	45.32	44.59	44.95
良种培育	61.56	64.16	62.83
动物疾病	75.43	90.67	82.35
饲料营养	65.99	49.43	56.52
农业管理	69.43	66.50	67.93
屠宰与加工	59.77	57.78	58.76
其他农业	52.45	23.94	32.87
正确率			73.41
宏平均值	70.62	68.68	69.03
加权平均值	72.95	73.41	72.72

从表 3 可以看出，模型在各类别上的表现存在一定差异，部分类别的分类效果较好，如病虫害和杂草、土壤和肥料和储存和保鲜等类别，优于整体分类效果；而在其他类别如其他种植和其他农业等类别的分类效果较差，远低于整体分类效果，仍有较大的提升空间。

表 4　模型对比实验结果

Tab.4　Results of model comparison experiments

模型	精确率/%	召回率/%	F1 值/%	正确率/%	浮点运算次数	参数量
BERT	68.45	69.38	68.50	69.38	$1.821\,9\times10^{9}$	1.023×10^{8}
BERT－TEXTCNN	69.15	70.03	69.29	70.03	$1.826\,4\times10^{9}$	1.031×10^{8}
BERT－CNN	69.34	69.66	69.26	69.66	$1.828\,4\times10^{9}$	1.025×10^{8}
BERT－RNN	68.68	69.60	68.87	69.60	$2.123\,9\times10^{9}$	1.118×10^{8}
Topic－BERT	72.04	72.59	71.96	72.59	$1.828\,6\times10^{9}$	1.055×10^{9}
本文模型(TAB)	72.95	73.41	72.72	73.41	$1.828\,6\times10^{9}$	1.055×10^{8}

注：不同版本的 BERT 参数量有差异，本文所有模型使用的 BERT 均为 bert-base-chinese，其参数量约为 1.023×10^{8}。

从表 4 可以看出，TAB 模型的整体分类性能优于其他常见的文本分类模型。无论是精确度、召回率、F1 值还是正确率，TAB 这种组合模型都优于其他几种常见的文本分类模型。相比之下，单纯使用 BERT 模型的效果最差，精确率、召回率、F1 值和正确率仅为 68.45%、69.38%、68.50% 和 69.38%，BERT－TEXTCNN 模型分别为 69.15%、70.03%、69.29% 和 70.03%，BERT－CNN 模型分别为 69.34%、69.66%、69.26% 和 69.66%，BERT－RNN 的实验结果分别为 68.68%、69.60%、68.87% 和 69.60%。表 4 还展示了各模型的浮点运算次数(Floating point operations，FLOPs)和参数量。

消融实验结果如表 5 所示，在不使用农业领域专用预训练模型 AgCBERT 模块时，模型的精确率、召回率、F1 值和正确率分别下降 0.91、0.82、0.76、0.82 个百分点；在不使用基于主题模型的主题特征向量表示的 Topic 模块时，模型的精确率、召回率、F1 值和正确率分别下降 3.53、2.92、3.16、2.92 个百分点。模型的性能均有明显下降，从而验证了本文提出的 2 个模块对提高模型性能的有效性。

表 5 消融实验结果

Tab. 5 Ablation experiment results %

模型	精确率	召回率	F1 值	正确率
TAB	72.95	73.41	72.72	73.41
TAB - AgCBERT	72.04	72.59	71.96	72.59
TAB - Topic	69.42	70.49	69.56	70.49

注:TAB - AgCBERT 模型中不使用农业领域专用预训练模型 AgcBERT 模块;TAB - Topic 模型中不使用基于主题模型的主题特征向量表示的 Topic 模块。

实验结果表明,将主题信息纳入 BERT 模型这一方法可以提高模型在特定领域分类任务(如农业问句分类)中的性能,TAB 模型可以有效地按农业领域的主题对用户提出的问句进行分类,这对于实现智能问答至关重要。通过收集并利用涵盖各个领域的中文农业语料库来进行农业领域预训练,证明了预训练模型在农业领域出色的文本表示能力。AgCBERT 对处理中文农业信息也具有重要意义,它可以提高农业问句分类任务的性能,促进农业文本分析技术的发展。

4 结束语

总体而言,本文为问答系统提出了一种融合主题信息与领域预训练向量的创新性问句分类模型。该模型使用线性组合方法将主题信息与预训练文本向量进行融合获取更丰富的特征表示信息。在一个包含 15 个类别的 12 097 个农业问答样本的数据集上进行实验验证,结果表明,所提出的模型不仅优于传统的 BERT 模型,而且在与其他最先进的模型的比较中表现出色,具有竞争力。然而,问句分类只是问答系统的部分内容,随着问答系统的广泛应用,模型的轻量化设计逐渐成为重要内容。

参 考 文 献

[1] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273 - 297.

[2] COVER T M, HART P E. Nearest neighbor pattern classification[J]. IEEE Trans. Inf. Theory, 1953, 13(1): 21 - 27.

[3] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989(2): 77.

[4] MCCALLUM A, NIGAM K. A comparison of event models for naive bayes text classification[C]// AAAI - 98 Workshop on Learning for Text Categorization, 1998: 41 - 48.

[5] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278 - 2324.

[6] ELMAN J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179 - 211.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.

[8] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv Preprint, arXiv: 1412. 3555, 2014.

[9] GALASSI A, LIPPI M, TORRONI P. Attention in natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4291 - 4308.

[10] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint, arXiv: 1810. 04805, 2018.

[11] SMITH D A, MCMANIS C. Classification of text to subject using LDA[C]// Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015). IEEE, 2015: 131 - 135.

[12] WANG Z Q, SUN X, ZHANG D X, et al. An optimal SVM-based text classification algorithm[C]// 2006 International Conference on Machine Learning and Cybernetics. IEEE, 2006: 1378 - 1381.

[13] MOHAMMED M, OMAR N. Question classification based on Bloom's taxonomy cognitive domain using modified TF - IDF and word2vec[J]. PloS One, 2020, 15(3): e0230442.

[14] 段青玲,魏芳芳,张磊,等. 基于 Web 数据的农业网络信息自动采集与分类系统[J]. 农业工程学报, 2016,32(12): 172 - 178.
DUAN Qingling, WEI Fangfang, ZHANG Lei, et al. Automatic acquisition and classification system for agricultural network information based on Web data[J]. Transactions of the CSAE, 2016, 32(12): 172 - 178. (in Chinese)

[15] 魏芳芳,段青玲,肖晓琰,等. 基于支持向量机的中文农业文本分类技术研究[J]. 农业机械学报, 2015,46(增刊): 174 - 179.
WEI Fangfang, DUAN Qingling, XIAO Xiaoyan, et al. Classification technique of Chinese agricultural text information based on SVM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(Supp.): 174 - 179. (in Chinese)

[16] 赵燕,李晓辉,周云成,等. 基于朴素贝叶斯的农业文本分类方法研究[J]. 节水灌溉, 2018(2):98 - 102.
ZHAO Yan, LI Xiaohui, ZHOU Yuncheng, et al. Astudy on agricultural text classification method based on naive Bayesian [J]. Water Saving Irrigation, 2018(2): 98 - 102. (in Chinese)

[17] 赵明,杜会芳,董翠翠,等. 基于 word2vec 和 LSTM 的饮食健康文本分类研究[J]. 农业机械学报, 2017,48(10):202 - 208.
ZHAO Ming, DU Huifang, DONG Cuicui, et al. Diet health text classification based on word2vec and LSTM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(10): 202 - 208. (in Chinese)

[18] 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.
ZHAO Ming, DONG Cuicui, DONG Qiaoxue, et al. Question classification of tomato pests and diseases question answering system based on BIGRU[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(5): 271 - 276. (in Chinese)

[19] 金宁,赵春江,吴华瑞,等. 基于 BiGRU_MulCNN 的农业问答问句分类技术研究[J]. 农业机械学报, 2020,51(5): 199 - 206.
JIN Ning, ZHAO Chunjiang, WU Huarui, et al. Classification technology of agricultural questions based on BiGRU_MulCNN [J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(5): 199 - 206. (in Chinese)

[20] 冯帅,许童羽,周云成,等. 基于深度卷积神经网络的水稻知识文本分类方法[J]. 农业机械学报, 2021,52(3):257 - 264.
FENG Shuai, XU Tongyu, ZHOU Yuncheng, et al. Rice knowledge text classification based on deep convolution neural network[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(3): 257 - 264. (in Chinese)

[21] 王郝日钦,吴华瑞,冯帅,等. 基于 Attention_DenseCNN 的水稻问答系统问句分类[J]. 农业机械学报, 2021, 52(7): 237 - 243.
WANG Haoriqin, WU Huarui, FENG Shuai, et al. Classification technology of rice questions in question answer system based on Attention_DenseCNN[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(7): 237 - 243. (in Chinese)

[22] 李悦晨,钱玲飞,马静. 基于 BERT - RCNN 模型的微博谣言早期检测研究[J]. 情报理论与实践, 2021, 44(7): 173
LI Yuechen, QIAN Lingfei, MA Jing. Earlydetection of micro blog rumors based on BERT - RCNN model[J]. Information Studies: Theory & Application, 2021, 44(7): 173. (in Chinese)

[23] 刘忠宝,王宇飞,张志剑. 基于深度学习模型的摘要结构功能识别方法研究[J]. 情报科学, 2021, 39(3):107 - 112.
LIU Zhongbao, WANG Yufei, ZHANG Zhijian. Research on therecognition method of abstract structure function based on deep learning model[J]. Information Science, 2021, 39(3): 107 - 112. (in Chinese)

[24] 王思丽,杨恒,祝忠明,等. 基于 BERT 的领域本体分类关系自动识别研究[J]. 情报科学, 2021, 39(7):75 - 82.
WANG Sili, YANG Heng, ZHU Zhongming, et al. Research onautomatic identification of domain ontology classification relations based on BERT[J]. Information Science, 2021,39(7): 75 - 82. (in Chinese)

[25] PENGBO Y, WEIMIN P, HAIJUN Z, et al. Identifying clickbait with BERT - BiGA model[J]. Data Analysis and Knowledge Discovery, 2021, 5(6): 126 - 134.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998 - 6008.

[27] 岳增营,叶霞,刘睿珩. 基于语言模型的预训练技术研究综述[J]. 中文信息学报. 2021, 35(9): 15 - 29
YUE Zengying, YE Xia, LIU Ruiheng. A survey of language model based pre-training technology[J]. Journal of Chinese Information Processing, 2021, 35(9): 15 - 29. (in Chinese)

[28] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach[J]. arXiv Preprint, arXiv: 1907. 11692, 2019.

[29] LAN Z, CHEN M, GOODMAN S, et al. Albert: a lite bert for self-supervised learning of language representations[J]. arXiv Preprint, arXiv: 1909. 11942, 2019.

[30] CLARK K, LUONG M T, LE Q V, et al. Electra: pre-training text encoders as discriminators rather than generators[J]. arXiv Preprint, arXiv: 2003. 10555, 2020.

[31] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 2227 - 2237.

[32] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[33] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877 - 1901.

[34] SUN Y, WANG S, LI Y, et al. Ernie: enhanced representation through knowledge integration[J]. arXiv Preprint, arXiv: 1904. 09223, 2019.

[35] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504 - 3514.

[36] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36(4): 1234 - 1240.

[37] BELTAGY I, LO K, COHAN A. SciBERT: a pretrained language model for scientific text[J]. arXiv Preprint, arXiv: 1903. 10676, 2019.

[38] LIU Z, HUANG D, HUANG K, et al. Finbert: a pre-trained financial language representation model for financial text mining [C]// Proceedings of the Twenty-ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021: 4513 - 4519.

[39] REZAYI S, LIU Z, WU Z, et al. AgriBERT: knowledge-infused agricultural language models for matching food and nutrition [C]// IJCAI, 2022: 5150 - 5156.