计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100059

ZHA_TGCN:面向低资源壮文的主题分类方法

赵卓洋[1] 秦董洪[1,4] 白凤波[1,4] 梁贤烨[1] 徐 晨[1] 郑月华[1] 梁宇锋[1] 蓝 盛[2,4] 周国平[3]
1 广西民族大学人工智能学院 南宁 530006
2 广西民族大学文学院 南宁 530006
3 广西民族大学预科教育学院 南宁 530006
4 语言计算与智能广西高校工程研究中心 南宁 530006
(zhuoyangzhao@stu.gxmzu.edu.cn)

摘 要 传统图卷积网络方法在数据有限的条件下能够有效建模图结构,但由于依赖稀疏的独热编码,其捕捉词与词之间上下文关系的能力存在局限性。这一问题在低资源语言环境中尤为突出。以壮文文本主题分类任务为例,该任务不仅面临数据稀缺的困境,还需应对复杂语言结构的挑战。针对这些挑战,提出了一种适用于低资源环境的壮文主题分类方法——ZHA_TGCN。该方法利用壮文预训练模型 ZHA_BERT 提取文本特征,并将文本特征与壮文声调特征相结合,输入 BiGRU 以学习深层语义表示,将学习到的表示向量作为文档节点的特征提供给 GCN,通过在 GCN 中执行标签传播来学习训练数据和未标记测试数据的特征表示。最后,利用 Softmax 层输出分类结果。实验结果表明,提出的方法在低资源壮文主题分类任务中的准确率为 82.12%,精确率为 90.08%,召回率为 92.46%,F1 值为 90.18%,验证了该方法的有效性。

关键词:低资源语言;壮文;主题分类;预训练模型;图卷积网络
中图分类号 TP391

ZHA_TGCN:A Topic Classification Method for Low-resource Sawcuengh Language

ZHAO Zhuoyang[1],QIN Donghong[1,4],BAI Fengbo[1,4],LIANG Xianye[1],XU Chen[1],ZHENG Yuehua[1],LIANG Yufeng[1],LAN Sheng[2,4] and ZHOU Guoping[3]
1 College of Artificial Intelligence,Guangxi Minzu University,Nanning 530006,China
2 College of Chinese Language and Literature,Guangxi Minzu University,Nanning 530006,China
3 College of Preparatory Education,GuangxiMinzu University,Nanning 530006,China
4 University Engineering Research Center of Computational Linguistics and Intelligence,Nanning 530006,China

Abstract Traditional graph convolutional network methods can effectively model graph structures under data-limited conditions. However,due to their reliance on sparse one-hot encoding,they face limitations in capturing the contextual relationships between words. This issue is particularly pronounced in low-resource language environments. Taking the Sawcuengh language text topic classification task as an example,this task faces not only data scarcity but also the challenge of complex linguistic structures. To address these challenges,this paper proposes a Sawcuengh language topic classification method suitable for low-resource settings—ZHA_TGCN. This method leverages the Sawcuengh pre-trained model,ZHA_BERT,to extract textual features,and combines these features with Sawcuengh tone features. These combined features are then input into a BiGRU to learn deep semantic representations. The learned representation vectors are used as node features for the GCN,which propagates labels to learn the feature representations of both the training data and the unlabeled test data. Finally,a Softmax layer is used to output the classification results. Experimental results show that the proposed method achieves an accuracy of 82.12%,precision of 90.08%,recall of 92.46%,and an F1 score of 90.18% in the low-resource Sawcuengh language topic classification task,demonstrating the effectiveness of the method.

Keywords Low-resource language,Sawcuengh language,Subjet classification,Pre-trained model,Graph convolutional network

基金项目:广西壮族自治区中央引导地方科技发展资金项目(桂科 ZY24212045);广西科技基地和人才专项(桂科 AD23026054);广西科技基地和人才专项(桂科 AD22035200)
This work was supported by the Central Guidance on Local Science and Technology Development Fund of Guangxi Zhuang Autonomous Region (GUIKEZY24212045),Guangxi Science and Technology Base and Talent Project(GUIKEAD23026054) and Guangxi Science and Technology Base and Talent Project(GUIKEAD22035200).
通信作者:白凤波(baif@gxun.edu.cn)

1 引言

主题分类作为自然语言处理领域的一项基石性任务，在新闻分类[1]、情感分析[2]、垃圾邮件检测[3]等多个应用场景中发挥着不可或缺的作用，推动了企业服务质量的优化，并为政府机构及学术界的决策过程提供了理论与数据的支撑。然而，当前主题分类研究在高资源语言与低资源语言之间表现出发展不均衡现象。一方面，高资源语言（例如英语和中文）在该领域的研究进展迅猛，相关理论框架与实用技术已日趋成熟。另一方面，针对低资源语言（如藏文和壮文）的研究仍显薄弱，存在显著的知识与研究缺口，亟需更为系统和深入的探索，以期平衡并推动整个主题分类领域的发展。

现代壮文（英文：Sawcuengh，也常被写成 Zhuang 或 Chuang，在 ISO 639-2/639-3 Code 代码中为 zha）[1)] 作为中国少数民族语言之一，属于声调语言范畴，有别于古壮文[4]，如图 1 所示。自 1957 年起，现代壮文采用以拉丁字母为基础的书写系统，并于 1982 年进一步修订，形成现今通用的标准壮文（第二套拉丁壮文）。壮文承载着丰富的民族文化和历史信息，具有重要的文化传承和技术应用价值。然而，由于壮文资源稀缺，尤其是文本标注数据的匮乏，为该领域的文本主题分类研究带来了诸多复杂且艰巨的挑战。

近年来，随着深度学习的发展，多种深度学习模型被广泛应用于文本表示学习，显著提升了文本主题分类的性能。在早期研究中，word2vec[5] 和 GloVe[6] 等词嵌入模型通过对大规模语料的训练，将词语映射到低维稠密向量空间，为文本表示奠定了基础。随后，以卷积神经网络（Convolutional Neural Network，CNN）[7] 为代表的深度学习模型凭借其出色的局部特征提取能力，在文本分类任务中展现出良好性能。循环神经网络（Recurrent Neural Network，RNN）[8] 及其变体，如双向长短期记忆网络（Bidirectional Long Short Term Memory，BiLSTM）[9] 和门控循环单元（Gated Recurrent Unit，GRU）[10] 的引入，进一步增强了模型捕获序列数据中长程依赖关系的能力。近期，以 BERT[11] 为代表的预训练语言模型通过设计双向编码器和自注意力机制，在海量文本上进行预训练，在文本表示学习领域取得了突破性进展，成为当前文本主题分类研究的重要方向。

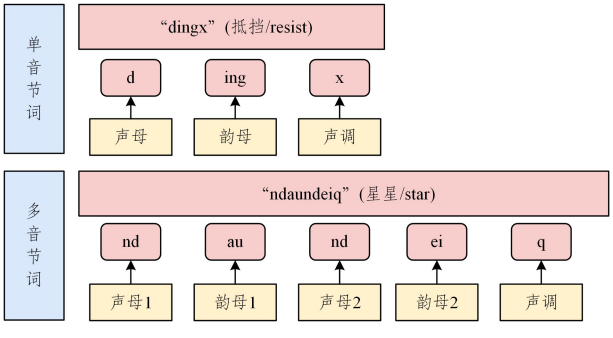


图 1 壮文拉丁字示例

Fig. 1 Example of Sawcuengh latin script

上述模型在捕获局部连续词序列的语义和句法信息方面表现出色，但在处理非连续和远距离语义文本中全局词的共现[12]现象时存在局限性。与此同时，图神经网络（Graph Neural Network，GNN）[13] 的兴起为文本主题分类提供了全新的建模方式。GNN 通过构建文本中的知识图谱，模拟文档内容之间的复杂关系，将文本主题分类问题转换为节点分类问题，从而有效地处理文本等非结构化数据，并在图嵌入中保留文本的全局结构信息。其中，图卷积网络（Graph Convolutional Network，GCN）[14] 作为 GNN 的经典方法，利用节点特征和邻接关系进行卷积操作，有效地捕获节点及其邻居之间的特征信息，广泛应用于文本分类任务。例如 TextGCN[15] 和 TensorGCN[16] 等方法，利用 GCN 建模文本与单词的图结构，成功地将文档和词节点嵌入统一的图表示中，展现出卓越的分类效果。然而，这些模型在处理低资源语言时面临两个主要挑战：一是节点特征初始化通常采用独热编码（One-hot Encoding），导致特征表示过于稀疏；二是现有模型缺乏对语言学特征的融合机制，无法充分利用声调等语言信息来增强语义表示。

综上所述，壮文主题分类面临的主要挑战可归纳为：数据资源稀缺、语言特征复杂、现有模型对壮文特点适应性不足以及缺乏有效的多特征融合机制。这些挑战使得直接应用现有主题分类方法难以取得理想效果，亟需开发针对壮文特点的专用分类模型。

为应对上述挑战，本文针对壮文分类数据匮乏的问题，构建了一个壮文主题分类数据集。该数据集主要来源于《壮语文广播稿精选》和《广西民族报网》[2)]。在此基础上，提出了结合 ZHA_T 词嵌入与 GCN 的主题分类框架。具体而言，首先构建词-文档异构图，通过预训练的 ZHA_BERT 模型提取文本特征，将其与声调特征融合后输入 BiGRU 进行深层语义建模，最后利用 GCN 实现文本分类。本文的贡献可以概括为以下几点：

（1）构建了一个涵盖多领域的壮文主题分类数据集，并通过收集大量未标注壮文文本对 BERT 进行掩码语言模型预训练，使模型更好地适应壮文的语言特点；

（2）创新性地将预训练模型提取的文本特征与壮文的声调特征进行融合，丰富了语义表示，使模型能够更准确地理解和捕捉壮文的语义信息；

（3）改进了图神经网络的结构，通过 BiGRU 对融合特征进行深层语义建模，并结合 GCN 捕获文本的全局依赖关系，有效提升了分类性能。

本文第 2 章将详细回顾相关领域的研究工作；第 3 章将阐述模型框架，并深入介绍提出的创新方法的具体细节；第 4 章将针对实验结果展开详细讨论；最后总结全文并展望未来。

2 相关工作

本文旨在研究低资源壮文主题分类，相关领域涵盖低资源语言文本主题分类与图神经网络的应用。

2.1 低资源文本主题分类

低资源文本主题分类是指在标注数据匮乏的情境下，探究对文本进行有效分类的方法研究。

近年来，学术界对低资源语言的文本主题分类给予了广泛关注，研究者相继提出了诸多创新性的方法来提高分类表

1) https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes

2) http://www.gxmzb.net/zw.htm

现。例如，Li 等[17]提出了一种基于跨语言模型微调的方法，通过形态分析和词干提取构建低噪声的微调数据集，显著提升了乌兹别克语、哈萨克语和吉尔吉斯语等低资源语言的文本分类表现。Sazzed[18]针对孟加拉语这一资源匮乏的语言，创建并标注了一个大型孟加拉语评论语料库，并利用机器翻译技术，将英语资源用于孟加拉语的情感分析任务。Yao 等[19]引入外部知识库来弥补元训练和元测试任务之间的差距，提出了一种知识感知的元学习方法，用于解决低资源文本分类问题。Fesseha 等[20]探讨了卷积神经网络与词嵌入技术在提格里尼亚语这一低资源语言文本分类中的应用。

最近，An 等[21]针对藏文这一低资源语言，提出了一种基于提示学习的低资源藏文文本分类方法。该方法利用大规模预训练语言模型，通过设计合理的提示模板，即使在训练数据不足的情况下，也能提升藏文文本分类的效果，验证了提示学习在民族语言处理中的潜力。

2.2 基于图神经网络的主题分类

近年来，多项研究表明，图神经网络在文本分类任务中的表现良好。通过处理图结构数据，GNN 能够有效捕捉文本中的复杂关系，从而提升分类模型的性能。

例如，Wen 等[22]提出了一种结合图结构预训练与提示学习的方法，通过图交互式对比策略和下游分类任务中的提示，增强了低资源文本分类的表现。Yao 等[15]通过构建一个包含文档节点和单词节点的异构图，提出了基于图卷积网络的文本分类方法。实验表明，两层 GCN 结构相比于单层结构，能更好地捕获文本的上下文信息，增加更多层数并未进一步提高模型的准确性。为了更全面地获取上下文信息，Bert-GCN 模型[23-24]结合了大规模预训练与归纳学习的优势，通过图卷积网络来实现文本分类，取得了良好效果。

Huang 等[25]提出了一种创新的 GNN 方法，为每篇输入文本生成一个图，同时共享全局参数，减少了文本与整个语料库之间的依赖性。该方法不仅允许在线测试，还保留了全局信息，并通过较小的文本窗口捕获更多局部信息，有效减少了边的数量和内存使用。

现有研究表明，预训练语言模型和图神经网络在文本分类任务中具有显著效果。然而，在壮文主题分类这一特定任务中仍存在若干挑战：一方面，主流预训练语言模型未能充分适应壮文的语言结构和词汇特征，导致直接迁移应用效果欠佳；另一方面，传统图卷积网络（GCN）通常采用独热编码进行节点特征初始化，这种方法难以有效捕捉词语间的深层语义关联。特别值得注意的是，现有方法普遍忽略了与声调语言语义密切相关的声调信息特征。针对低资源条件下壮文主题分类任务的特殊性，本文创新性地提出了一种融合预训练语言模型、图卷积网络和声调信息的综合框架，致力于改进低资源壮文主题分类的特征表示能力与语义理解效果，为相关语言的文本处理任务探索新途径。

3 本文方法

3.1 壮文主题分类模型框架

本文的文本分类模型如图 2 所示。第一阶段（Stage I）为 BERT 模型的预训练，基于 30 多万条，共 28.5MB 大小的纯壮文文本语料，本文使用 BERT[1)] 进行掩码语言模型（Masked Language Model，MLM）预训练[2)]，学习壮文文本中的上下文信息，得到 ZHA_BERT 语言模型。

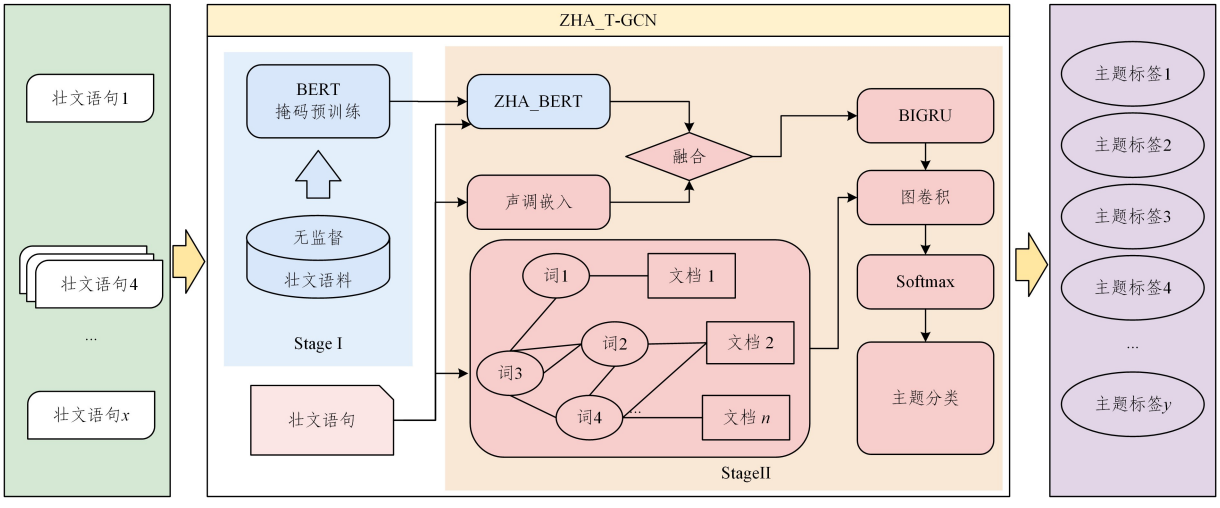


图 2 壮文文本分类模型框架

Fig. 2 Framework of Sawcuengh text classification model

第二阶段（Stage II）为分类模型训练。该模型利用词共现以及文档和词之间的关系构建了一个包含文档和词的异构图。在对原始文本进行预处理之后，将 ZHA_BERT 模型提取的句子级别的上下文特征与壮文独特的声调特征进行融合。然后，利用 BiGRU 对融合后的特征进行建模，进一步提取深层次语义信息，并用其初始化文档节点的特征表示。将节点特征输入 GCN 中，由 GCN 模块更新节点特征信息。最终的输出结果被视为文档节点的最终特征表示，并发送给 Softmax 层分类器进行分类。

3.2 壮文声调及处理

声调是指语言中一个音节读音的高低升降，能够影响词汇的语义。壮文，作为一种典型的有声调语言，共有 8 个声

1) https://huggingface.co/google-bert/bert-base-chinese

2) https://github.com/huggingface/transformers/blob/main/examples/pytorch/language-modeling/run_mlm.py

调，其中包括 6 个舒声调和 2 个塞声调。舒声调是指音节中以元音（如 a,e,o,i,u,w）或鼻辅音（如 m,n,ng）作为韵尾的声调；塞声调则是指音节中以塞音辅音（如 b,d,g,p,t,k）作为韵尾的声调。塞声调又可细分为两个调类：高音组塞声调和低音组塞声调，前者以 p,t,k 作韵尾表示，后者以 b,d,g 作韵尾表示。

如表 1 所列，壮文共有 22 个声母和 108 个韵母，壮文的 6 个舒声调，除第一调不用字母作调号外，其余舒声调分别用 z,j,x,q,h 5 个字母来表示并标写于韵母末尾。h 既作辅音，又作声调符号。塞声调不标声调符号。

表 1 壮文声母韵母声调表

Table 1 Sawcuengh consonants, vowels, and tones chart

类别	符号
声母	/b/,/mb/,/m/,/f/,/v/,/d/,/nd/,/n/,/s/,/l/,/g/,/gv/,/ng/,/h/,/r/,/c/,/y/,/ny/,/ngy/,/by/,/gy/,/my/
韵母	/a/,/e/,/i/,/o/,/u/,/w/,/ai/,/ae/,/ei/,/oi/,/ui/,/wi/,/au/,/aeu/,/eu/,/iu/,/ou/,/aw/,/am/,/aem/,/em/,/iem/,/im/,/om/,/oem/,/uem/,/um/,/an/,/aen/,/en/,/ien/,/in/,/on/,/oen/,/uen/,/un/,/wen/,/wn/,/ang/,/aeng/,/eng/,/ieng/,/ing/,/ong/,/oeng/,/ueng/,/ung/,/wng/,/ap/,/aep/,/ep/,/iep/,/ip/,/op/,/oep/,/uep/,/up/,/at/,/aet/,/et/,/iet/,/it/,/ot/,/oet/,/uet/,/ut/,/wet/,/wt/,/ak/,/aek/,/ek/,/iek/,/ik/,/ok/,/oek/,/uek/,/uk/,/wk/,/ab/,/aeb/,/eb/,/ieb/,/ib/,/ob/,/oeb/,/ueb/,/ub/,/ad/,/aed/,/ed/,/ied/,/id/,/od/,/oed/,/ued/,/ud/,/wed/,/wd/,/ag/,/aeg/,/eg/,/ieg/,/ig/,/og/,/oeg/,/ueg/,/ug/,/wg/
声调	/z/,/j/,/x/,/q/,/h/,/(以 p,t,k 为韵尾)/,/(以 b,d,g 为韵尾)/

声调在壮文中具有区分词义的作用，不同的声调会导致词汇的语义发生变化，因此，准确处理壮文中的声调信息显得尤为重要。

为了有效提取和标注壮文的声调信息，本文采用基于正则表达式的文本处理方法（见算法 1），对壮文中的声母、韵母和声调进行自动化分割和标注，如图 3 所示。

算法 1 壮文声调处理

```
Input: Sawcuengh text sequence //输入壮语序列
Output: FinalResult // 输出标注后的壮语文本
Tools: InitialDict, FinalDict, ToneDict
1. Syllables← Split(Sawcuengh TextSequence) /
2. AnnotatedSyllables
3. for each Syllable in Syllables do
4.     Initial, Final, ToneExtractComponents(Syllable, InitialDict, FinalDict, ToneDict)
5.     if Tone is empty then
6.         if Final ends with {b,d,g} then
7.             Tone ←"7"
8.         else if Final ends with {p,t,k} then
9.             Tone -"8"
10.        else
11.            Tone ←"1"
12.        end if
13.    end if
14.    if Final ends with {b,d,g,p,t,k} then
15.        if Final ends with {b,d,g} then
16.            Tone -"7"
17.        else
18.            Tone -"8"
19.    end if
20.    if Initial is not empty then
21.        AnnotatedSyllable ← Initial+"_"+Final +" "+Tone
22.    else
23.        AnnotatedSyllable- Final+"_"+Tone
24.    end if
25.    Append AnnotatedSyllable to AnnotatedSyllables
26. end for
27. FinalResult← Join(AnnotatedSyllables,"0")//用"0"连接音节
28. Output(FinalResult)
```

具体而言，本文首先根据壮文音节的 4 种常见组成形式进行处理：

(1)声母＋韵母＋声调：如"naz"（田）、"doiq"（退）等。

(2)声母＋韵母：如"bi"（年）、"da"（眼）等。

(3)韵母＋声调：如"iq"（小）、"at"（抵押）等。

(4)韵母：如"an"（安装）、"ien"（烟）等。

利用预先构建好的声母、韵母和声调字典对每个音节进行匹配，以准确识别其中的声母和韵母。然后通过提取音节的韵尾部分，进一步确定其对应的声调值。对于塞声调的处理，本文通过检查音节韵尾是否由塞音辅音（b,d,g,p,t,k）构成，来判断该音节所属的声调类别，并将其划分为第七调和第八调，从而实现调值标注。

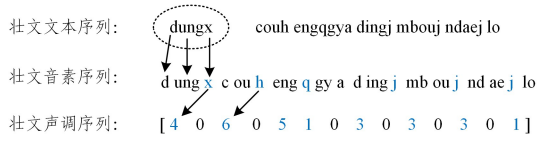


图 3 壮文声调处理

Fig. 3 The processing of Sawcuengh tone

3.3 壮文词嵌入处理

针对传统 GCN 模型依赖稀疏独热编码、难以捕捉语义关联的局限性，本节提出一种融合预训练语言模型与声调特征的多层次词嵌入方法（ZHA_T）对节点进行初始化。其思路如图 4 所示。

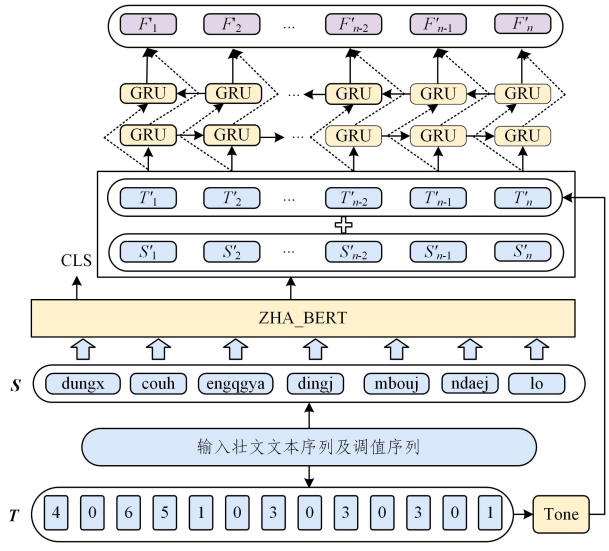


图 4 壮文词嵌入模型

Fig. 4 Sawcuengh word embedding model

该方法通过以下三阶段实现语义增强:首先,基于大规模壮文语料预训练的 ZHA_BERT 模型提取文本上下文特征;其次,结合壮文声调分析模块提取声调特征,并将其与 BERT 特征融合;最后,利用 BiGRU 模块对融合特征进行序列建模,将学习到的表示向量作为 GCN 节点的初始化特征。通过整合预训练语言模型的语义特征、声调特征以及序列模型的上下文建模优势,ZHA_T 词嵌入有效提升了低资源条件下壮文文本的特征表示质量。

基于壮文预训练模型 ZHA_BERT 提取文本的上下文语义特征,获取丰富的语义表示,图 4 中 $\boldsymbol{S}$ 是每个句子中的词向量表示,n 表示句子中的词数,CLS 不包含语义信息。$\boldsymbol{S}$ 通过 BERT 模型,利用其自注意机制学习文本中词与词之间的关系,学习不同语境下的词向量表示,具体公式如下:

$$S_i^{\text{BERT}} = BERT(S_i), i = 1, 2, \cdots, n \tag{1}$$

$$Output_{\text{BERT}} = \{S_1^{\text{BERT}}, S_2^{\text{BERT}}, \cdots, S_n^{\text{BERT}}\} \tag{2}$$

$$Output_{\text{BERT}} \in R^{n * d^{\text{BERT}}} \tag{3}$$

其中,d^{BERT} 为 BERT 模型的输出维数,设为 768。$\boldsymbol{T}$ 是句子转换成声调序列的词向量表示,m 表示声调序列中的声调数。$\boldsymbol{T}$ 通过声调分析模块 *Tone*,得到相应的声调嵌入 ToneEmbedding。

$$T_i^{\text{Tone}} = Tone(T_i), i = 1, 2, \cdots, m \tag{4}$$

BERT 模型可以有效地提取文本序列中的语义特征,为了进一步增强语义表征能力,将声调特征维度 m 进行 Padding 补零,使其与经过 BERT 的文本特征维度 n 一致,然后进行相加融合。

最后将融合后的特征 F 送入 BiGRU 网络来进一步学习上下文的语义特征,具体公式如式(4)—式(7)所示。

$$F_i^{\text{BERT-Tone}} = S_i^{\text{BERT}} + T_i^{\text{Tone}} \tag{5}$$

$$F_i^{\text{BERT-Tone-BiGRU}} = BiGRU(F_i^{\text{BERT-Tone}}) \tag{6}$$

$$Output_{\text{BERT-Tone-BiGRU}} = \{F_1^{\text{BERT-Tone-BiGRU}}, \cdots, F_n^{\text{BERT-Tone-BiGRU}}\}$$

$$Output_{\text{BERT-Tone-BiGRU}} \in R^{n * 2d^{\text{GRU}}} \tag{7}$$

其中,d^{GRU} 为 GRU 的输出维数,设为 384;$2d^{\text{GRU}}$ 为 BERT-Tone-BiGRU 模型的最终输出维度,设为 768。

3.4 图卷积网络

图卷积网络是一种处理图数据的神经网络架构,本文使用 TextGCN[15] 中的方法构造了一个词节点和文档节点的异构图,如图 2 所示,对标注的壮文数据集进行构造。其中文档节点和词节点之间的边权用词频-逆文档频率(term frequency-inverse document frequency,TF-IDF)表示,词节点和词节点之间的边权用点互信息(Point Mutual Information,PMI)表示,当计算 i 和 j 两个节点的边的权重 $A_{i,j}$ 时,如式(8)所示:

$$A_{i,j} = \begin{cases} PMI(i,j), & i,j = word, i \neq j \\ TF\text{-}IDF(i,j), & i = word, j = document \\ 1, & i = j \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

两个词 i 和 j 间的值 PMI 的计算式如式(9)—式(11)所示:

$$PMI(i,j) = \log \frac{p(i,j)}{p(i)p(j)} \tag{9}$$

$$p(i,j) = \frac{W(i,j)}{W} \tag{10}$$

$$p(i) = \frac{W(i)}{W} \tag{11}$$

其中,W 表示滑动窗口的总个数,$W(i)$ 表示包含单词的滑动窗口的个数,$W(i,j)$ 表示同时包含单词 i 和 j 的滑动窗口的个数。

TF-IDF 是一种统计方法,用于评估一个词对文档集中某个文档的重要性。TF-IDF 的值越大,说明该词对文档越重要。

$$TF\text{-}IDF = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{12}$$

其中,$tf_{i,j}$ 表示单词在文档 j 中出现的频率,N 表示整个文档集中的文档数,df_i 是包含单词的文档集中的文档数。

$$X = \begin{pmatrix} X_{ndoc} \\ 0 \end{pmatrix} \in R^{(ndoc+nword) * 2d^{GRU}}$$

$$X_{ndoc} \in R^{ndoc * 2d^{\text{GRU}}} \tag{13}$$

将 X 输入 GCN 模型中,通过节点间的边更新节点间的信息,其中,第 i 层 GCN 的输出特征矩阵 $\boldsymbol{L}^{(i)}$ 的计算式如下:

$$\boldsymbol{L}^{(i)} = \rho(\widetilde{\boldsymbol{A}} L^{(i-1)} W^{(i)}) \tag{14}$$

其中,ρ 为激活函数,$\widetilde{\boldsymbol{A}}$ 为归一化邻接矩阵,$W^{(i)}$ 是第 i 层的权值参数。$\boldsymbol{L}^{(0)} = X$ 为模型的输入特征矩阵。

$$Z_{\text{GCN}} = \text{Softmax}(g(X, A)) \tag{15}$$

其中,g 表示 GCN 模型,GCN 的输出被视为文档的最终表示,最后经 Softmax 层进行分类。

4 实验

4.1 实验设置

4.1.1 数据集构建

本研究所构建的数据集主要来源于《壮语文广播稿精选》和《广西民族报网》,涵盖了农业、文化、饮食等多个领域,提供了丰富的语言应用场景。

数据集包含文化、教育、法律、历史、医疗、地理、保健、饮食以及农业 9 个类别,如图 5 所示,共计 9 900 条数据。

内容	标签
hix miz di vunz youg ndawsingz cawx ranz,	0
daj ngoenz daihseig hwnj menh gueng	8
gwn haeux seiz,bouxsaenqgyau yinniz ceiq geih caemh aen vanj buenz ndeu nip gwn, hix mbouj gwn gijgwn bouxwng bungq gvag haenx,	7
aenvih yienghneix guh aiq hawj nohheuj deng sieng, yinx hwnj engqlai heujbingh.	4
de okrog bae youzlangh gaeng haujlai nanz lo.ndaw suenva daengx bi cungi miz gij sing'anggriu gyoengq lwgnyez.miz bi ndeu seizcou	0
nohmaklaeg 7 daengz 10 aen daem yungz	7
yungh gij nywj ginggvaq anhva gisuz cawqleix gvaq haenx daeuj gueng vaiz	8
ndaej ftengzre noncehlunz caeug binghnonsezbeih	4
ngoenzneix,gou ndaej caenndang depgaenh mwngz ho,aendap yenzanh	0
daj bi gouj ngeih dawz duz"ak gvaq duzromh" riuj bae le, lij hancangh gaeng haujlai bi mbouj ciengx gyauq roegvameiz lo.	3
doenghgij genj louzvuengz miz cawz youz cozyung	4
de umj aenbiengx ceiq doeklaeng roengz mbaek diegbingz, dangh roengz raemnx bae.	5
yinzminz fazyen wngdang ciugfap roengz linghhawjcienz.	2
ngoenzgvengq youg banringz gonglaeng 2-3 aen cungdaeuz? haeujsim biengi muengx,hawj dohraeuj ndaw muengx mbouj mauhgvag 30 doh,	8
mehginh da daengx cam:"dwg caen ha?" yenzluj couh naeuz:"gaej saeng de.de gangjriu ne, aen ci neix gwn giyouz."	7
gou hainduj dingh roengzdaeuj siengj mbangj saeh he, lumj gauj mbouj hwnj daihhag wnggai baenzlawz guh,	1
ndaej hanhhaed danjgucunz caeug ganhyouz sanhcij habbaenz.	6

图 5 数据集示例

Fig. 5 Dataset example

表 2 列出了数据集的各类别及样本数量情况。在数据

预处理阶段，标注人员首先对文本进行了分句处理，并挑选了具有明确标签的样本，以便模型可以更准确地捕捉文本中的语义特征。数据集按照 8∶1∶1 的比例随机划分为训练集、验证集和测试集。

表 2 壮语数据集

Table 2 Sawcuengh language dataset

类别	样本总数	训练集	验证集	测试集
文化	1820	1429	189	202
教育	522	425	59	38
法律	1523	1216	151	156
历史	335	263	40	32
医疗	1238	1017	105	116
地理	307	253	18	36
保健	356	286	32	38
饮食	968	770	104	93
农业	2831	2261	291	279
Total	9900	7920	990	990

4.1.2 评价指标与实验环境

在模型评估过程中，采用了多种评价指标来全面衡量模型性能，包括正确率(Accuracy)、准确率(Precision)、召回率(Recall)以及 F1 值。这些指标能够分别反映模型的整体性能、对各类别预测的精度、对各类别的覆盖程度，以及综合平衡精度的表现。为了提供更加全面的模型表现评价，采用所有类别的正确率、准确率、召回率和 F1 值的平均值作为整体评价指标。

本文采用 Python 3.9，Torch 2.1.2，使用的 GPU 型号为 NVIDIA GeForce RTX 3090 24 GB。主要模型参数如下：文本最大长度为 128。在模型训练中，Batch_Size 为 64，优化器使用 Adam。BERT 和 BiGRU 的学习率设置为 0.000 02，GCN 模块的层数为 2，学习率为 0.001，Dropout 为 0.5。

4.2 实验结果及分析

4.2.1 实验设置

为验证本文 ZHA_T-GCN 模型的有效性，选取 6 个具有代表性的基准模型开展对比实验，具体配置如下：

(1)FastText：轻量级文本分类模型。参数配置为学习率 0.001、文本最大长度为 32、*Dropout*=0.5、隐藏层为 256。

(2)TextRNN：基于循环神经网络，通过门控机制捕捉序列依赖，学习率为 0.001、文本最大长度为 64、*Dropout*=0.5、隐藏层为 128。

(3)DPCNN：深度金字塔卷积网络，分层提取多粒度特征，实验配置为学习率为 0.000 1、文本最大长度为 32、*Dropout*=0.5。

(4)TextCNN：经典卷积模型，通过多尺寸卷积核捕获局部语义，参数设置为学习率为 0.001、文本最大长度为 64、*Dropout*=0.5。

(5)BERT：预训练语言模型基线，采用中文 BERT-Base 模型，学习率为 0.000 05、文本最大长度为 32、*Dropout*=0.5、隐藏层为 768。

(6)BERT-GCN：结合预训练与图卷积的方法，使用官方模型框架，配置为学习率为 0.000 03、文本最大长度为 128、*Dropout*=0.5、隐藏层为 768、GCN 层数为 2。

4.2.2 基线模型对比

为全面评估本文提出的 ZHA_T-GCN 模型性能，选择了多个具有代表性的基线模型进行对比实验。实验结果如表 3 所列。

表 3 实验结果对比

Table 3 Comparison of experimental results

(%)

Model	*Accuracy*	*Precision*	*Recall*	*F1*
FastText[26]	57.81	47.47	44.93	45.28
TextRNN[27]	66.88	52.46	50.88	50.51
DPCNN[28]	70.00	72.75	54.86	57.54
TextCNN[29]	70.52	69.61	53.01	55.16
BERT[11]	75.42	69.17	68.63	68.67
BERT-GCN[23]	81.01	86.00	80.56	82.02
ZHA_T-GCN	**82.12**	**90.08**	**92.46**	**90.18**

在传统模型中，FastText 和 TextRNN 的表现相对较弱，准确率分别为 57.81%和 66.88%。这些模型结构相对简单，难以有效捕捉壮文文本中的复杂语义信息。虽然 DPCNN 和 TextCNN 通过深层卷积结构在特征提取方面有所改进，使准确率提升至 70%左右，但在处理壮文这类低资源语言时仍显不足。预训练模型 BERT 的引入带来了显著提升，准确率达到 75.42%，比传统模型提高了 5～18 个百分点。这表明，预训练模型通过大规模语料的预训练，能够更好地理解壮文文本的语义。进一步地，BERT-GCN 通过引入图结构建模，将准确率提升至 81.01%，F1 值达到 82.02%，验证了图神经网络在增强文本表示方面的优势。

本文 ZHA_T-GCN 模型在各项指标上均取得最优性能：准确率为 82.12%、精确率为 90.08%、召回率为 92.46%、F1 值为 90.18%。相比最强基线 BERT-GCN，各项指标分别提升了 1.11 个百分点、4.08 个百分点、11.9 个百分点和 8.16 个百分点。这一显著提升主要得益于 3 个方面的创新：壮文预训练策略、声调特征的融入以及改进的图神经网络结构。

4.2.3 消融实验

为深入分析模型各组件的贡献，本文设计了一系列消融实验，逐步验证各模块的有效性。表 4 列出了不同模型变体的性能对比结果。

表 4 模型各组件消融实验结果对比

Table 4 Ablation study results of model components

(%)

模型类别	模型变体	*Accuracy*	*Precision*	*Recall*	*F1*
基础模型	BERT	75.42	69.17	68.63	68.67
预训练改进	ZHA_BERT	77.50	74.21	67.23	69.07
序列建模	BERT-BiGRU	68.44	49.67	51.48	49.72
	BERT-Tone-BiGRU	72.60	66.04	59.99	59.81
	ZHA_BERT-BiGRU	78.02	77.54	67.38	70.01
	ZHA_T	79.17	77.20	68.45	70.49
图结构融合	BERT-GCN	81.01	86.00	80.56	82.02
	ZHA_BERT-GCN	81.31	83.67	82.94	82.75
	ZHA_BERT-BiGRU-GCN	81.62	94.29	89.29	89.63
	ZHA_T-GCN	82.12	90.08	92.46	90.18

(1)声调特征的影响

实验结果表明，引入声调特征能有效提升模型性能。BERT-Tone-BiGRU 相比 BERT-BiGRU，在准确率和 F1 值上分别提升了 4.16 个百分点和 10.09 个百分点(59.81% vs 49.72%)。类似地，ZHA_T 比 ZHA_BERT-BIGRU 的 F1 值提升了 0.48 个百分点(70.49% vs 70.01%)。这表明，声调特征能有效增强模型对壮文语义的理解。例如，在处理“ndaej”(得到)和“ndaex”(看)这样的近音词时，声调特征帮助模型准确区分它们的语义差异，提高了分类准确性。

(2)预训练策略的影响

预训练策略的效果体现在多个层面。首先，ZHA_BERT通过壮文预训练，相比原始BERT在F1值上提升了0.4个百分点(69.07% vs 68.67%)；其次，结合BiGRU的序列建模后，ZHA_BERT-BiGRU的F1值进一步提升至70.01%，说明针对壮文的预训练和序列建模能有效提升模型性能。

(3)GCN结构的贡献

引入GCN结构后，模型性能获得显著提升。基础BERT-GCN的F1值达到82.02%，比未使用GCN的BERT提升了13.35个百分点。在此基础上，ZHA_BERT-GCN和ZHA_BERT-BiGRU-GCN的F1值分别达到82.75%和89.63%，展现出预训练和序列建模与GCN的良好协同效果。最终的ZHA_T-GCN模型通过融合声调特征，将F1值提升至90.18%，较BERT-GCN提升了8.16个百分点，验证了声调特征对提升分类性能的重要作用。

这些结果充分说明，声调特征、预训练策略和GCN结构的组合能够有效提升壮文文本分类性能。特别是在召回率方面，ZHA_T-GCN达到了92.46%的优异表现，这表明模型能够更全面地识别各类别的文本特征。同时，90.18%的F1值也证实了模型在精确率和召回率之间取得了良好的平衡。

4.3 细粒度性能分析

4.3.1 各类别性能分析

从表5的分类详细结果可以看出，模型在不同类别上的表现存在显著差异。样本量充足的类别(如农业、文化、法律)普遍表现较好，F1值多在80%以上；中等规模类别(如医疗、饮食)表现次之，F1值在65%～80%；小样本类别(如历史、地理、保健)表现相对较弱，F1值普遍低于60%。这种性能差异主要源于训练数据的不平衡分布，暗示了在低资源场景下，数据量对模型性能的重要影响。

表5 BERT，BERT-BiGRU与ZHA_BERT-BiGRU的实验结果对比

Table 5 Comparison of experimental results between BERT，BERT-BiGRU，and ZHA_BERT-BiGRU

(%)

类别	*Precision*			*Recall*			*F1*		
	BERT	BERT-BiGRU	ZHA_BERT-BiGRU	BERT	BERT-BiGRU	ZHA_BERT-BiGRU	BERT	BERT-BiGRU	ZHA_BERT-BiGRU
文化	72.46	55.41	69.70	76.92	84.10	82.56	74.63	66.80	75.59
教育	86.49	78.57	90.62	86.49	89.19	78.38	86.49	83.54	84.06
法律	91.43	90.55	86.71	84.77	76.16	90.73	87.97	82.73	88.67
历史	61.54	20.00	88.89	51.61	6.45	25.81	56.14	9.76	40.00
医疗	66.30	50.77	60.58	54.46	58.93	74.11	59.80	54.55	66.67
地理	48.57	0	58.33	47.22	0	38.89	47.89	0	46.67
保健	44.44	0	70.00	54.05	0	56.76	48.78	0	62.69
饮食	67.31	71.59	85.00	76.92	69.23	74.73	71.79	70.39	79.53
农业	83.94	80.15	88.03	85.19	79.26	84.44	84.56	79.70	86.20
Macro平均	69.17	49.67	77.54	68.63	51.48	67.38	68.67	49.72	70.01

4.3.2 模型局限性分析

尽管本文模型整体表现优异，但仍存在一些局限性。在样本量低于500的类别中，模型性能显著下降，特别是地理类(307个样本)和保健类(356个样本)，即使采用了预训练方法，性能提升仍然有限。这些局限性为未来研究指明了方向，如探索更有效的小样本学习方法、引入领域知识等。

结束语 本研究针对低资源壮文文本分类的挑战，创新性地提出ZHA_T-GCN模型，结合预训练BERT、声调特征及GCN，有效提升了分类性能。实验结果显示，该模型在准确率、F1分数等指标上均有显著提升，验证了其在低资源语种中的潜力，并为壮文及其他低资源语言的文本分类研究提供了新思路。

然而，在低资源语言任务中，样本不平衡和数据稀缺仍然是影响模型性能的关键因素。未来的研究将继续优化现有模型，进一步探索声调特征和语言信息的多维度融合，以提升低资源语言的处理能力。通过更加精细的特征融合和模型改进，期望能够为其他民族语言的文本分类任务提供更加有效的解决方案。

参 考 文 献

[1] WANG A H. Don't follow me：Spam detection in twitter[C]// 2010 International Conference on Security and Cryptography (SECRYPT). IEEE，2010：1-10.

[2] PANG B，LEE L. Opinion mining and sentiment analysis[J]. Foundations and Trends © in Information Retrieval，2008，2(1/2)：1-135.

[3] KARIM A，AZAM S，SHANMUGAM B，et al. A comprehensive survey for intelligent spam email detection[J]. IEEE Access，2019，7：168261-168295.

[4] LI J Y. From Ancient Zhuang Characters to Zhuang Script：The Zhuang People Now Have Their Own Writing System [J]. Contemporary Guangxi，2019(Z1)：86.

[5] CHURCH K W. Word2Vec[J]. Natural Language Engineering，2017，23(1)：155-162.

[6] PENNINGTON J，SOCHER R，MANNING C D. Glove：Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014：1532-1543.

[7] LI Z，LIU F，YANG W，et al. A survey of convolutional neural networks：analysis，applications，and prospects [J]. IEEE Transactions on Neural Networks and Learning Systems，2021，33(12)：6999-7019.

[8] MIKOLOV T，KARAFIÁT M，BURGET L，et al. Recurrent neural network based language model[C]// Interspeech. 2010：1045-1048.

[9] 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.

[10] ZULQARNAIN M, GHAZALI R, GHOUSE M G, et al. Efficient processing of GRU based on word embedding for text classification[J]. International Journal on Informatics Visualization, 2019, 3(4): 377-383.

[11] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[12] WU L, CHEN Y, SHEN K, et al. Graph neural networks for natural language processing: A survey [J]. Foundations and Trends © in Machine Learning, 2023, 16(2): 119-328.

[13] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model [J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.

[14] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[C]// International Conference on Learning Representations. 2017.

[15] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 7370-7377.

[16] LIU X, YOU X, ZHANG X, et al. Tensor graph convolutional networks for text classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 8409-8416.

[17] LI X, LI Z, SHENG J, et al. Low-resource text classification via cross-lingual language model fine-tuning[C]// China National Conference on Chinese Computational Linguistics. Cham: Springer, 2020: 231-246.

[18] SAZZED S. Cross-lingual sentiment analysis in bengali utilizing a new benchmark corpus[C]// Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generate. 2020: 50-60.

[19] YAO H, WU Y, AL-SHEDIVAT M, et al. Knowledge-Aware Meta-learning for Low-Resource Text Classification[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 1814-1821.

[20] FESSEHA A, XIONG S, EMIRU E D, et al. Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya [J]. Information, 2021, 12(2): 52.

[21] AN B, ZHAO W N, LONG C J. Low-resource Tibetan Text-Classification Based on Prompt Learning[J]. Journal of Chinese Information Processing, 2024, 38(2): 70-78.

[22] WEN Z, FANG Y. Augmenting low-resource text classification with graph-grounded pre-training and prompting[C]// Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 506-516.

[23] LIN Y, MENG Y, SUN X, et al. BertGCN: Transductive Text Classification by Combining GNN and BERT[C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1456-1462.

[24] YUAN Y, LV S, BAO Z, et al. A Joint Model for Text Classification with BERT-BiLSTM and GCN[C]// Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition. 2022: 180-186.

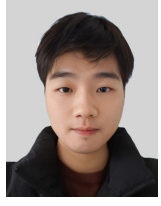
[25] HUANG L, MA D, LI S, et al. Text Level Graph Neural Network for Text Classification[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3444-3450.

[26] 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.

[27] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016: 2873-2879.

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

[29] GUO B, ZHANG C, LIU J, et al. Improving text classification with weighted word embeddings via a multi-channel TextCNN model[J]. Neurocomputing, 2019, 363: 366-374.



ZHAO Zhuoyang, born in 1999, master, is a student member of CCF (No. N5417G). His main research interest is natural language processing.

BAI Fengbo, born in 1978, Ph.D, lecturer, is a member of CCF (No. F6846M). His main research interests include artificial intelligence, data science and evidence science.