计算机科学
COMPUTER SCIENCE

基于DAN与FastText的藏文短文本分类研究

李果, 陈晨, 杨进, 群诺

引用本文

李果, 陈晨, 杨进, 群诺. 基于DAN与FastText的藏文短文本分类研究[J]. 计算机科学, 2024, 51(6A): 230700064-5.

LI Guo, CHEN Chen, YANG Jing, QUN Nuo. Study on Tibetan Short Text Classification Based on DAN and FastText [J]. Computer Science, 2024, 51(6A): 230700064-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700064

基于 DAN 与 FastText 的藏文短文本分类研究

李 果[1,2] 陈 晨[1,2] 杨 进[1,3] 群 诺[1]
1 西藏大学信息科学技术学院 拉萨 850000
2 藏文信息技术教育部工程研究中心 拉萨 850000
3 四川大学网络空间安全学院 成都 610000
(2216643277@qq.com)

摘 要 随着藏文信息不断融入社会生活,越来越多的藏文短文本数据存在网络平台上。针对传统分类方法在藏文短文本上分类性能低的问题,文中提出了一种基于 DAN-FastText 的藏文短文本分类模型。该模型使用 FastText 网络在较大规模的藏文语料上进行无监督训练获得预训练的藏文音节向量集,使用预训练的音节向量集将藏文短文本信息转化为音节向量,把音节向量送入 DAN(Deep Averaging Networks)网络并在输出阶段融合经过 FastText 网络训练的句向量特征,最后通过全连接层和 softmax 层完成分类。在公开的 TNCC(Tibetan News Classification Corpus)新闻标题数据集上所提模型的 Macro-F1 是 64.53%,比目前最好评测结果 TiBERT 模型的 Macro-F1 得分高出 2.81%,比 GCN 模型的 Macro-F1 得分高出 6.14%,融合模型具有较好的藏文短文本分类效果。

关键词:藏文短文本分类;特征融合;深度平均网络;快速文本
中图分类号 TP391.1;TP183

Study on Tibetan Short Text Classification Based on DAN and FastText

LI Guo[1,2],CHEN Chen[1,2],YANG Jing[1,3] and QUN Nuo[1]
1 School of Information Science and Technology,Tibet University,Lhasa 850000,China
2 Engineering Research Center of Tibetan Information Technology Ministry of Education,Tibet University,Lhasa 850000,China
3 School of Cyber Science and Engineering,Sichuan University,Chengdu 610000,China

Abstract As Tibetan information continues to be integrated into social life,more and more Tibetan short text data is available on online platforms. Aiming at the low classification performance of traditional classification methods on Tibetan short texts,a Tibetan short text classification model based on DAN-FastText is proposed. The model uses the FastText network to perform unsupervised training on a large-scale Tibetan corpus to obtain the pre-trained Tibetan syllabic vector set,uses the pre-trained syllable vector set to convert the Tibetan short text information into syllable vector,sends the syllable vector into the deep averaging networks(DAN) network and fuses the sentence vector features trained by the FastText network in the output stage,and finally completes the classification through the fully connected layer and the softmax layer. On the publicly available tibetan news classification corpus(TNCC) news headline dataset,Macro-F1 is 64.53%,which is 2.81% higher than that of the TiBERT model and 6.14% higher than that GCN model,and the fusion model has a better Tibetan short text classification effect.

Keywords Tibetan short text classification,Feature fusion,Deep averaging networks,Fast text

1 概述

藏文是一种具有悠久历史和丰富文化的少数民族语言,随着藏文信息化技术不断普及,越来越多的藏文短文本数据存在网络平台上,如何有效利用这些短文本数据挖掘出有价值的信息具有重要意义。藏文短文本分类将藏文短文本按照不同的主题进行归类的任务,是藏文自然语言处理基础和重要的研究内容,具有词汇稀疏、语义模糊、类别不平衡等通用的短文本分类问题,需要采用有效的方法来提高分类性能。目前,短文本分类任务的研究主要集中在文本表示、主题模型拓展和外部数据增强上。

在文本表示方面,Salton 等[1]提出了向量空间模型(Vector Space Model,VSM),VSM 的假设是将文本看作一个特征序列的随机排序而不考虑词序,也被称作词袋模型(Bag-of-words Model,BOW),它通过把每一个独立单词表示成以 0 和 1 组成的二元向量,使得文本可以被计算机所处理。向量空间模型具有表示方法简单和易于计算的特点,在普通文本的建模及特定领域的应用中取得了跨越式的效果。然而,仅仅采用二元向量来表示词表中的所有单词,并且忽视词在文本中的顺序关系,导致词与词之间的语义信息无法挖掘。Mikolov 等[2]提出了 word2vec 词向量模型,从大量文本语料中无监督学习词语的语义知识,将词语表示为低维的稠密向

基金项目:国家自然科学基金(61872254,62162057)
This work was supported by the National Natural Science Foundation of China(61872254,62162057).
通信作者:杨进(yangjin_abc123@163.com)

量，从而反映词语之间的相似性和关联性，可以有效地捕捉词语的语义信息，使得相似或相关的词语在向量空间中具有较高的相似度。虽然不断有研究对文本向量化进行优化，但是短文本特征稀疏的特点让模型对文本的向量表示始终有限。

在主题模型扩展方面，一些学者尝试利用概率生成模型来提取短文本的潜在特征，如 LSA(Latent Semantic Analysis)模型[3]、PLSA(Probability Latent Semantic Analysis)模型[4]和 LDA(Latent Dirichlet Allocation)模型[5]。与以往的方法相比，概率生成模型通过推理策略获取短文本的主题特征，并将其与文档的原始特征进行融合，从而实现较好的分类效果。Cao 等[6]创建 LSA 微博文本模型降低话题检测的错失率，提高微博话题检测的性能。Wang 等[7]基于 PLSA 学习概率分布语义信息，提出了新型多标签分类算法，在多标签的公开数据集上的对比实验表明，PLSA 能够提高多标签算法的性能。Sun 等[8]将 LDA 模型引入到日志异常检测能够获得较高的查准率、查全率和调和分数，并且对于新日志模板的接入，模型也能有较高的稳定性。BTM(Biterm Topic Model)主题模型由 Yan 等[9]于 2013 年提出，是一种短文本主题提取模型，通过将语料库中的词进行聚合，得到稳定词共现频率进而清楚揭示词之间的相关性，相比传统主题模型对每个文档中的词共现建模，以潜在的方式反映语料库的语意结构，解决短文本数据稀疏问题。

在外部数据增强方面，Jiang 等[10]利用知识图中的外部知识来丰富文本信息，而且可以通过图神经网络来利用知识图中的结构信息以促进对文本的理解，实验结果表明利用外部知识及其结构信息进行短文本分类是有效的，具有良好的可解释性。He 等[11]整合医学知识图谱到复杂的医学文本分类任务中，在诊断相关组分类中优于所有基线和当前最先进的模型。Li 等[12]在模型中将知识图谱实体链接进行嵌入，引入外部知识以获取语义特征，同时使用双重注意力机制提高模型对短文本中有效信息提取的效率。

相比中英文短文本分类，藏文短文本分类起步较晚，但是通过学者的研究，已经取得很大的进展，其中主要关注在藏文情感分类任务上。Jiang 等[13]根据藏文微博的表述特点，提出了基于多特征的情感倾向性分析算法，实验表明具有藏汉双语表述的微博文本比纯藏文表述的微博文本有更好的正确率。Yan 等[14]先手工构建词典，基于词性词典的藏文情感分析具有很好的倾向性。Zhu 等[15]采用 Albert 预训练模型构建词向量，融合数据集中标注的情感词，使用图神经网络模型进行训练，最终达到 98.6%的准确率。Meng 等[16]融合藏文音节和词条特征，融合多核卷积神经网络、双向长短时记忆网络、多头注意力机制获取藏文文本的多维度特征，在测试集上的实验表明，在分类准确率指标上融合音节和词条特征的模型优于基于音节的模型和基于词条的模型。除了对藏文情感分类做了大量研究之外，在多分类的藏文短文本方面，Qun 等[17]使用端到端的神经网络模型在 TNCC 数据集上对长短文本进行研究，在短文本上长短期记忆网络优于卷积神经网络和神经词袋模型。Xu 等[18]通过音节和文档共现关系构建文本图，使用图卷积神经网络对文本节点进行分类，在 TNCC 标题数据集上达到了 61.94%的准确率。Liu 等[19]基于预训练模型 BERT，收集藏文网站上的大规模文本数据进行训练，提出藏文预训练模型 TiBERT，TiBERT 模型在下游的藏文文本分类任务和问题生成任务上进行评测，在 TNCC 标题数据集上，TiBERT 模型的 Macro-F1 值是 61.72%，是当前最好的评测结果。

传统方法对藏文短文本分类的准确率较低，使用大语言模型进行微调虽然能够取得较好的分类结果，但是需要耗费大量的计算成本和时间成本。在公开数据集 TNCC 的标题数据集上，本文提出的 DAN-FastText 模型能够在本地性能一般的 PC 机上使用较短的训练时间和推理时间获得比目前最好评测结果的 TiBERT 模型的 Macro-F1 得分高出 2.81%。

2 DAN-FastText 模型

DAN-FastText 模型使用 FastText 网络在基于音节的较大规模藏文新闻内容语料上进行无监督训练，预训练的音节向量集将藏文短文本转化为音节向量特征，通过把音节向量特征送入 DAN 网络并在输出阶段融合经过 FastText 网络的在数据集上有监督训练的句向量特征，最终结果由 DAN 网络的特征和 FastText 网络的句向量特征进行线性插值。DAN-FastText 模型的结构如图 1 所示。

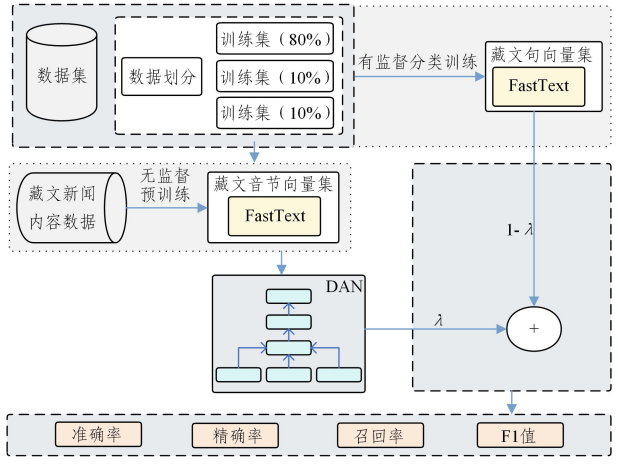

图 1 DAN-FastText 模型的结构

Fig. 1 Structure of DAN-FastText model

2.1 DAN 模型

DAN 模型是一个简单无序的深度神经网络，能够将无序函数的速度和有序函数的准确率结合起来，在一些分类和问答任务上，只需要很短的训练时间，能够达到甚至优于一些有序函数的模型。DAN 模型在神经词袋模型(NBOW)基础上堆砌多个非线性层，每一层都是学习前一层更加抽象的表示，这样在 DAN 模型中每一层都会逐渐放大词嵌入平均值中微小但是又有实际意义的差异，反映在原始文本中能够使表达含义相近的文本有更加相似的向量表示，能够使 DAN 模型比标准的 NBOW 模型更好地捕获文本输入中细微的变化，同时每一层非线性变化只是一次矩阵计算，计算复杂性随着层数增加，而与输入文本的长度无关，DAN 模型和 NBOW 模型在训练时间上没有明显的差异。深度无序的 DAN 网络只需要在普通的笔记本电脑上进行几分钟的训练，就能在句子或者文档分类任务上获得很好的结果，先对输入的文本的词嵌入进行向量平均化，然后将平均值传入一个或多个非线性层，最后将最后一层的向量表示进行分类，为了提高网络的鲁棒性，对输入的文本进行 word dropout 正则化，即把输入文本分词之后随机丢弃一些单词。具体计算式如式(1)－式(4)所示：

$$\boldsymbol{r}_w \sim \text{Bernoulli}(p) \tag{1}$$

$$\hat{X}=\{w \mid w\in X \text{ and } \boldsymbol{r}_w>0\} \tag{2}$$

$$\boldsymbol{z}_{\mathrm{DAN}}^0=g(w\in X)=\frac{\sum_{w\in\hat{X}} v_w}{|\hat{X}|} \tag{3}$$

$$\boldsymbol{r}_{\mathrm{DAN}}^i=g(\boldsymbol{r}_{\mathrm{DAN}}^{i-1})=\rho(\boldsymbol{W}_i\cdot\boldsymbol{z}_{\mathrm{DAN}}^{i-1}+\boldsymbol{b}_i) \tag{4}$$

其中，式(1)和式(2)对输入文本进行分词之后为单词序列 X，对于每个单词 w 以满足概率 p 的伯努利分布保留得到新的单词序列$\hat{X}$，式(3)是将$\hat{X}$中每个单词进行向量化并求平均值得到向量$\boldsymbol{z}_{\mathrm{DAN}}^0$，式(4)是对平均值向量进行多次非线性变化，其中 ρ 是激活函数，在本文中使用 RELU 函数，经过多次 g 函数变化之后得到最终文本表示向量$\boldsymbol{z}_{\mathrm{DAN}}$。DAN 模型的结构如图 2 所示。

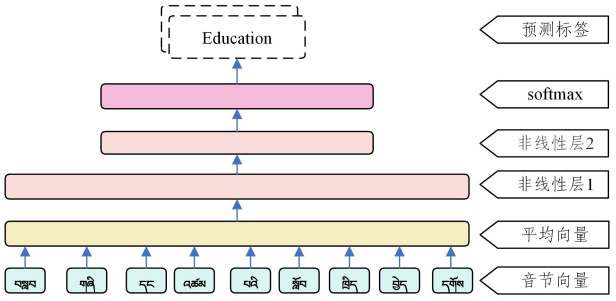

图 2　DAN 模型的结构

Fig. 2　Structure of DAN model

图 2 中的音节向量是先对原始音节列表 word dropout 正则化，然后通过预训练的音节向量集合进行嵌入表示，计算 9 个藏文音节向量的平均值得到平均值向量，即$\boldsymbol{z}_{\mathrm{DAN}}^0=\sum_{i=1}^{9}\frac{c_i}{9}$，将平均向量$\boldsymbol{z}_{\mathrm{DAN}}^0$作为非线性层 1 的输入得到向量$\boldsymbol{z}_{\mathrm{DAN}}^1$，即$\boldsymbol{z}_{\mathrm{DAN}}^1=\mathrm{RELU}(\boldsymbol{W}_1\cdot\boldsymbol{z}_{\mathrm{DAN}}^0+b_1)$，又将向量$\boldsymbol{z}_{\mathrm{DAN}}^1$输入到第二个非线性层得到向量$\boldsymbol{z}_{\mathrm{DAN}}^2$，之后经过 softmax 层得到预测标签 Education，其中 softmax 函数是 $\mathrm{softmax}(\boldsymbol{q})=\frac{\exp\boldsymbol{q}}{\sum_{j=1}^{k}\exp\boldsymbol{q}_j}$。

2.2　FastText 模型

FastText 是开源、免费、轻量级的库，能够学习词的向量表示和句子分类。在词向量表示方面，FastText 是将字符 n-gram 纳入到 skipgram 模型中，将每个单词表示一袋字符的 n-gram，向量表示与每个字符的 n-gram 相关联，单词被表示为关联的字符向量相加。这种词向量表示方法有很快的训练速度，同时结合子词信息的词向量表示方法能够让 FastText 模型在任何语言的文本语料库上进行训练，一些生僻词尽管在训练语料中并未出现，也能获得这些生僻词很好的向量表示。整个词向量的训练过程是无监督的，不需要对语料进行预处理，FastText 模型优于不考虑字词信息或者依赖形态分析的大多数方法。在句子分类方面，FastText 的分类准确率和一些深度学习分类器相当，然而训练和评估速度要快好几个数量级，在标准的多核 CPU 上，1 分钟之内能将 50 万句子分类到 31.2 万类别中。FastText 的模型结构特别简单，模型结构如图 3 所示。

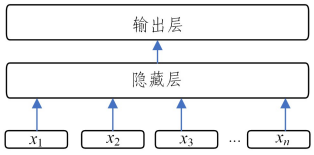

图 3　FastText 模型的结构

Fig. 3　Structure of FastText model

FastText 模型有 3 层：输入层、隐藏层、输出层。输入层是多个向量表示的单词，隐藏层是对多个词向量的叠加平均，输出层是将隐藏层向量映射到标签的特征向量。

2.3　DAN-fast Text 模型

DAN-FastText 模型是在 DAN 网络的第三个非线性层融合经过 FastText 网络有监督训练的句向量特征，如式(5)所示，其中 λ 的取值为 0.5。

$$\boldsymbol{Z}=\lambda\boldsymbol{z}_{\mathrm{DAN}}^3+(1-\lambda)\boldsymbol{Z}_{\mathrm{FastText}} \tag{5}$$

其中，$\boldsymbol{z}_{\mathrm{DAN}}^3$是 DAN 网络第三个非线性层输出的向量，$\boldsymbol{Z}_{\mathrm{FastText}}$是经过 FastText 网络有监督训练的 DAN 网络中对应的同一条文本的句向量，$\boldsymbol{Z}$ 是由 DAN 网络的输出特征和 FastText 网络的句向量特征进行线性插值的特征向量表示，其中维度是 200，利用多层感知机(MLP)把向量 $\boldsymbol{Z}$ 维度降为标签类别数，最后使用 softmax 函数进行分类，使用交叉熵损失函数优化训练参数。

3　实验及结果分析

3.1　数据集

已有工作对藏文文本语料进行了一些研究，目前公开的语料有 TNCC 数据集。TNCC 数据集是 2017 年在 Qun 等[17]的论文中公开的，从中国西藏网上获取，数据集分两部分，一部分是基于新闻标题的短文本数据集，另一部分是基于新闻内容的长文本数据集，本文仅使用短文本数据集，数据样例如表 1 所列。

表 1　TNCC 标题的数据样本

Table 1　TNCC headlines data sample

Text content	Class Label
འཕགས་སོང་ལྷ་ས་རྩོང་ཁྲིན་ནི་ བང་རྒྱལ་ནུབ་རྒྱལ་ཁྱོན་ཀྱི་ ས་མཚམས་ཤིག་ལས་ཤས་ཆེ་བའི་ རྩོང་ཁྲིན་ ཚོགས་མ་ཏུ་ འགྲུབ་རྒྱུ་ ཡིན་ པ། (2025 年拉萨市将成为我国首个高原生态城市)	Environment
རོང་ཀྱི་ རྒྱུག་ གཞིས་སྤྲོག་ གསོའི་ ཉིད་ སྐོས་ ལས། ཉེད་ དང་ མང་ ཚོགས་ཀྱི་ འཚེལ་ བ་ ཐབ་ཏུ་ བཀང་ པ། (西藏双语教育活动密切了干群关系)	Education
ཚོས་ རྗེ་ དོན་ གྲུབ་ རིན་ ཆེན་ ཀྱི་ གསུང་འབུམ་ བང་ རྒྱུན་ འཚོལ་ ཞིབ་ དང་ དོད་ རིག་ པའི་ གཞུང་ བསྒྲིགས་ བཙུགས་ ཚོགས་ ཆེན་ གཞིས་ པ། ཇི་ ལིང་ རྒྱས་ ཚོགས་ པ། (曲吉东珠仁青全集出版发行和第二次藏学学术研讨会在西宁召开)	Literature
དེ་ ལོའི་ ཟླ་ དང་ དུས་ བཅུ་ བར་ རྩོ་ ཤ་ ཁྲེལ་ ཀྱི་ ཡུལ་ སྤོར་ པོང་ འཚའ་ ཟོར་ ནུབ་ ཐུབ་ ༡༥༥ཡོངས་ འགྲུབ་ ཐུང་ པ། (今年一月至九月山南旅游收入达到 155 亿元)	Tourism
དེ་ ལོ་ རྒྱས་ བཅུང་ ནོར་ སྐྱེ་ ལོའི་ རྒྱུན་ བཙོལ་ ཁེན་ འགག་ འཚ་ འཁགས་ ཅོག་ པོངས་ རྒྱས་ ཁྲལ་ ཀྱི་ ཚེ་ ཁོལ་ གཏོང་ རྒྱ་ རེད་ འདུག (从今年开始将全面推行大病医疗保制度)	Medicine

数据集以藏文音节为单位划分，数据总共 9276 条，按照 8∶1∶1 随机划分训练集、验证集和测试集，每个标题的平均长度是 16 个藏文音节，TNCC 藏文新闻标题短文本数据集划分如表 2 所列。

表 2 数据集划分

Table 2 Dataset partition

数据集	训练集	验证集	训练集	平均长度
TNCC	7420	928	928	16

藏文新闻短文本数据集包括 12 个类别，其中 Education 类数量最多，有 2132 条数据，约占总体数据量的 30%，Instruments 类数量最少，有 255 条数据，约占总体数据量的 3%，可见数据集的类别数量分布有很大的不平衡性，数据集标签分类如图 4 所示。

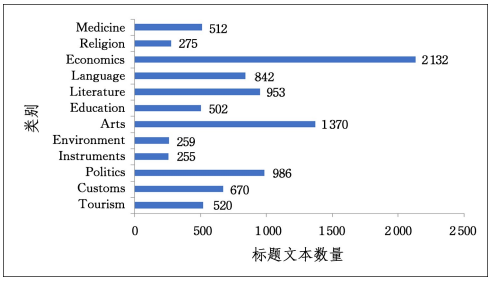

图 4 数据集标签分布

Fig. 4 Dataset label partition

3.2 参数设置

在本地普通 PC 机上使用 15.7G 的藏文语料对 FastText 模型进行 20 个 epoch 无监督训练获得维度为 300 的预训练藏文音节向量集，并在 TNCC 标题数据集上对 FastText 模型进行 30 个 epoch 有监督分类训练获得维度为 200 的藏文句向量。DAN 与 FastText 特征融合实验中，第一个非线性层维度设置为 300，第二、三个非线性层维度设置为 200，使用 Adam 优化器，学习率设置为 2×10^{-3}，epoch 设置为 30，word dropout 正则化设置为 0.1。在融合 DAN 和 FastText 阶段，最终结果是 λ 乘以深度平均网络特征和 $(1-\lambda)$ 乘以 FastText 网络的句向量特征相加，其中 λ 的值为 0.5。

3.3 评估指标

藏文新闻短文本分类的评估指标使用准确度 Accuracy，A)、精确度(Precision，P)、召回率(Recall，R)和综合指标 Macro-F1 来衡量分类模型的性能，其计算式如式(6)－式(9)所示。其中，TP(True Positive，真阳例)表示预测是正类，实际也是正类的文本个数；TN(Ture Negative，真阴例)表示预测是负类，实际也是负类的文本个数；FP(False Positive，假阳例)表示预测是正类，实际是负类的文本个数；FN(False Negative，假阴例)表示预测是负类，实际是正类的文本个数；通过以下公式计算得到每一类的 F1 值，再将各类的 F1 值求均值，即为评估指标 Macro-F1。

$$A=\frac{TP+TN}{TP+FP+TN+FN}*100\% \tag{6}$$

$$P=\frac{TP}{TP+FP}*100\% \tag{7}$$

$$R=\frac{TP}{TP+FN}*100\% \tag{8}$$

$$F1=\frac{2PrecRec}{Prec+Rec}*100\% \tag{9}$$

3.4 实验结果及分析

为了说明 DAN-FastText 模型能够相比其他主流模型有更好的藏文短文本分类性能，在公开数据集 TNCC(Tibetan News Classification Corpus)标题短文本数据集上进行对比实验，与 CNN 模型[17]、LSTM 模型[17]、Transformer 模型[19]、TextCNN 模型[19]、DPCNN 模型[19]、TextRCNN 模型[19]、TiBERT 模型[19]、TiBERT＋CNN 模型[19]、GCN 模型[18]进行对比实验，在 TNCC 标题数据集上实验结果如表 3 所列。

表 3 在 TNCC 标题数据集上的实验结果

Table 3 Experimental results on TNCC headlines dataset

(%)

Model	*A*	*P*	*R*	*F1*
CNN	54.42	49.22	48.34	48.64
LSTM	62.65	58.33	56.43	56.65
Transformer	46.88	54.21	46.88	42.91
TextCNN	60.94	59.58	60.94	58.90
DPCNN	64.06	59.05	64.06	59.68
TextRCNN	65.62	63.12	65.62	60.80
TiBERT	65.62	62.88	65.62	61.72
TiBERT＋CNN	65.62	59.47	65.62	60.93
GCN	61.94	61.24	56.91	58.39
DAN-FastText	**65.84**	**64.05**	**65.67**	**64.53**

实验结果表明，在公开数据集 TNCC 新闻标题短文本数据上，DAN-FastText 模型的当前实验结果 Macro-F1 值比藏文预训练模型 TiBERT 高出 2.81%，比图神经网络模型 GCN 高出 6.14%，并且只是在普通 PC 机器上用几分钟就完成训练和推理，然而 TiBERT 预训练模型的微调和推理需要具有 GPU 的服务器并花费相当长的时间才能完成，图卷积神经网络 GCN 是一种归纳式学习模型，是将文本数据构建文本图全量读入内存，需要较多的内存资源和训练及推理时间，DAN-FastText 模型用少量的计算成本显著提升了藏文短文本分类的性能，获得当前 TNCC 新闻标题短文本数据集下最好的分类结果。

为了进一步验证本文模型的有效性，将该模型进行分解进行消融实验，消融实验结果如表 4 所列。

1)DAN：使用预训练音节向量集将文本转成音节向量作为 DAN 网络的输入，两次非线性变换后经过 Softmax 层分类。

2)FastText：通过 FastText 网络得到藏文短文本向量表示，经过 Softmax 层分类。

表 4 在 TNCC 标题数据集上的消融实验结果

Table 4 Ablation experimental results on TNCC headlines dataset

(%)

Model	A	P	R	F1
DAN	63.58	61.50	62.80	61.92
FastText	63.79	63.24	62.17	62.24
DAN-FastText	**65.84**	**64.05**	**65.67**	**64.53**

本文提出了基于 DAN-FastText 的藏文短文本分类算法，DAN 网络能够获取短文本中细微而有实际意义差异的局部特征，FastText 网络在数据集上有监督训练的句向量能够获取短文本的全局特征，两者的融合增强了模型对藏文短文本数据的表示能力，从而有效提高了分类任务的效果。

3.5 有效的 λ 值

λ 影响着 DAN 特征向量和 FastText 句特征向量之间结

果的折衷。图 5 给出了 DAN-FastText 模型在 TNCC 标题短文本数据集下不同 λ 对应 macro-F1 的值。从实验结果可以得出，λ 值越大，macro-F1 的值越高，当 λ＝0.5 时，macro-F1 的值最高，继续增大 λ 的值，macro-F1 的值反而逐渐减小，其中 λ 值在 0.3 和 0.6 之间，macro-F1 的平均值在 63%以上。因此 λ 值取 0.5，能够使 DAN 模型和 FastText 模型预测结果得到很好融合，从而使模型对藏文文本的分类效果最佳。

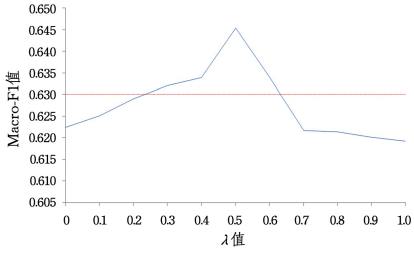

图 5　参数 λ 对分类 macro-F1 值的影响

Fig. 5　Effect of parameter λ on macro-F1

结束语　本文对 DAN 网络和 FastText 网络进行特征融合，显著提升了藏文短文本分类的性能，在公开数据集 TNCC 新闻标题短文本数据集上验证了融合模型的有效性。在当前的工作中，本文只是对两个模型进行简单的线性插值，一定程度上提升了藏文新闻短文本的表示能力。另外，本文存在一定的不足，只是在 TNCC 数据集上进行实验，TNCC 数据集文本数量相对较少，FastText 模型预训练藏文音节向量集依赖外部语料的数量和质量，这些都会对实验结果产生影响。在未来的研究中，将采集更多的藏文新闻标题数据并融合主题模型挖掘短文本潜在的语义信息，并引入外部知识来扩充短文本的外部特征，解决藏文短文内容简短、数据稀疏、语义具有奇异性的问题，提升融合模型对藏文短文本语义的表示能力和分类准确率。

参 考 文 献

[1] SALTON G，WONG A，YANG C A. A vector space model for automatic indexing[C]// Communications of the ACM. 1975：613-620.

[2] MIKOLOV T，SUTSKEVER I，CHENK，et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. 2013：3111-3119.

[3] DUMAIS，SUSAN T. Latent semantic analysis. AnnualReview of Information Science and Technology[J]. Annual Review of Information Science and Technology，2004，38：189-230.

[4] CHRISTOS H P，PRABHAKAR R，HISAO T，et al. Latent semantic indexing：a probabilistic analysis[J]. Journal of Computer and System Sciences，2000，61(2)：217-235.

[5] BLEI D M，NG A Y，JORDANM I. Latent dirichlet allocation [J]. Machine Learning Research Archive，2003，3(Jan)：993-1022.

[6] CAO C P，CUI H C. Microblog topic detection based on LSA and structural property[J]. Application Research of Computers，2015，32(9)：2720-2723.

[7] WANG Y B，ZHENG W J，CHENG Y S，et al. Multi-label classification algorithm based on PLSA learning probability distribution semantic information[J]. Journal of NANJING University(Natural Science)，2021，57(1)：75-89.

[8] SUN X K，DAI H，ZHOU J H，et al. LTTFAD：log template topic feature-based anomaly detection[J]. Computer Science，2023，50(6)：313-321.

[9] YAN X H，GUO J F，LAN Y Y，et al. A biterm topic model for short texts[C]// WWW 2013-Proceedings of the 22nd International Conference on World Wide Web. 2013：1445-1456.

[10] JIANG X H，SHEN Y H，WANG Y Z，et al. BaKGraSTeC：a background knowledge graph based method for short text classifications[C]// 2020 IEEE International Conference on Knowledge Graph(ICKG). IEEE，2020：360-366.

[11] HE Y，WANG C，ZHANG S，et al. KG-MTT-BERT：knowledge graph enhancedbert for multi-type medical text classification [J]. arXiv：2210.03970，2022.

[12] LI B H，XIANG Y X，FENG D I，et al. Short text classification model combining knowledge aware and dual attention[J]. Journal of Software，2022，33(10)：3565-3581.

[13] JIANG T，YUAN B，YU H Z. Multi-feature based sentiment analysis of Tibetan microblogs[J]. Journal of Chinese Information Processing，2017，31(3)：163-169.

[14] YAN X D，HUANG T. Tibetan sentence sentiment classification based on emotion dictionary[J]. Journal of Chinese Information Processing，2018，32(2)：75-80.

[15] ZHU Y L，DEJI K Z，QUN N，et al. Sentiment analysis of Tibetan short texts based on graphical neural networks and pre-training models[J]. Journal of Chinese Information Processing，2023，37(2)：71-79.

[16] MENG X H，YU H Z. Tibetan text sentiment classification combining syllables and words[J]. Journal of Chinese Information Processing，2023，37(2)：80-86.

[17] QUN N，LI X，QIU X，et al. End-to-End neural text classification for Tibetan[C]// The Sixteenth China National Conference on Computational Linguistics. 2017：1-8.

[18] XU G X，ZHANG Z X，YU S N，et al. Tibetannews text classification based on graph convolutional networks[J]. Data Analysis and Knowledge Discovery，2022，7(6)：73-85.

[19] LIU S S，DENG J J，SUN Y，et al. TiBERT：tibetan pre-trained language model[C]// 2022 IEEE International Conference on Systems. IEEE，2022：2956-2961.

LI Gu，born in 1994，postgraduate. His main research interest includs natural language processing.

YANG Jing，born in 1980，professor. His main research interests include cyberspace security and artificial intelligence.