基于大语言模型的标准文献分类研究

刘春卉[1*] 高志春[2] 张 辉[3] 黄振远[3]

（1.中国标准化研究院；2.山西省市场监督管理局；3.北京航空航天大学）

摘 要：在当今大数据时代，随着标准等文献呈现爆炸性增长，文献的高效管理与服务面临着巨大挑战。由于产业的不断演进和多样化，传统的标准分类体系无法灵活适应不断变化的产业需求，导致标准分类与实际产业之间的鸿沟日益加深。尤其在信息时代，该问题显著突显，而传统标准分类的转型升级困难。因此，解决标准分类与产业匹配难题成为提升文献管理效能和服务质量的重要一环。在这一背景下，本文提出一种创新性方法，旨在弥合标准分类与产业之间的差距，提高产业分类的准确性，从而更好地满足不断发展的产业需求。同时，该方法注重解决在中文产业分类领域所面临的多语义、多类别和少标注数据等复杂问题。

关键词：大语言模型，语义表征，文献，标准

DOI编码：10.3969/j.issn.1674–5698.2024.12.007

Research on Standard Literature Classification Based on Large Language Model

LIU Chun–hui[1*] GAO Zhi–chun[2] ZHANG Hui[3] HUANG Zhen–yuan[3]

（1. China National Institute of Standardization; 2. Shanxi Provincial Administration for Market Regulation; 3. Beihang University）

Abstract: In today's era of big data, the explosive growth of standards and other literature poses significant challenges for the efficient management and services of documents. Due to the continuous evolution and diversification of industries, traditional standard classification systems struggle to adapt flexibly to the ever-changing demands of industries, resulting in a gap between standard classification and actual industrial need. In the information age, this problem is notably emphasized, and the transformation and upgrading of traditional standard classifications become challenging. Therefore, addressing the issue of standard classifications being difficult to align with industrial needs has become a crucial aspect of enhancing the efficiency and quality of document management and services. In this context, the innovative approach proposed in this paper aims to bridge the gap between standard classification and industries, enhancing the accuracy of industrial classification to better meet the evolving demands of industries. Simultaneously, this method focuses on addressing the complex challenges faced in the field of Chinese industrial classification, including issues such as multiple semantics, multiple categories, and limited annotated data.

Keywords: large language models, semantic representation, literature, standard

---

基金项目：本文受山西省市场监督管理局“支撑山西省高质量转型发展标准化专项”资助。

作者简介：刘春卉，通信作者，研究馆员，博士，研究方向为区域标准化、标准国际化、标准知识服务。

高志春，二级巡视员，研究方向为地方标准化、标准管理。

张辉，教授，研究方向为计算机科学与技术。

黄振远，博士研究生，研究方向为网络空间安全。

0 引言

在当今数字化转型时代，信息量呈指数级增长，记载了各个领域的数字化过程及成果，形成了庞大的数据海洋。标准化领域同样经历着这一潮流，如：山西省在长期的科研攻关及生产应用过程中，产出了大量的标准文献资源。这些标准按照适用范围划分，包括国家标准，行业标准，地方标准和企业标准，按标准化对象划分，又有产品标准、过程标准、服务标准等。其不仅分类繁多，且各产业本身也有其独立的分类体系，因此标准分类很难与产业分类进行直接关联。在信息化时代之前，由于标准相对较少，主要依赖相关领域的专家进行人工分析，确定标准所属产业的类别信息。然而，如今大量标准文献的涌现使得人工产业分类方法不再适用。一方面传统方法不仅需要大量人力投入，而且分类效率不高。因此，迫切需要利用深度学习等技术构建一套能够自动化分析标准文献、提取产业类别信息的模型[1]。

本文针对标准分类体系无法适应不断变化的产业需求这一挑战，通过深度学习技术提出一种先进的自动化标准产业分类模型。分类作为数据挖掘的一个重要分支，近年来在各种实际应用中受到广泛研究和应用。标准文献是一种文本，基于文本的分类模型可以分为3个阶段：特征提取、分类器选择和模型评估。其中特征提取是文本分类的关键步骤之一，它有助于将文本数据转化为机器学习算法可以理解的形式。传统的文本分类方法主要依赖于手工设计的特征和统计学习方法。典型的方法包括n-gram模型[2]、支持向量机[3]等传统机器学习算法。这些方法在一定程度上取得了成功，但受限于特征表达和模型的表达能力，难以捕捉文本中的复杂语义和上下文信息。随着深度学习的兴起，基于神经网络的文本分类方法取得了显著的突破。特别是利用预训练的大型语言模型（如：BERT[4]、GPT[5]）进行文本表示学习，使得模型能够更好地理解语境和语义关系，大语言模型在各种文本分类任务中拥有卓越的性能，超越了传统方法的限制。此外，卷积神经网络（CNN）和循环神经网络（RNN）等架构也被广泛应用于文本分类[6]。CNN通过卷积操作捕捉局部特征，而RNN则能够建模序列信息，两者结合使用更能有效地处理文本数据。

综合而言，文本分类领域取得了长足的进展，从传统的特征工程方法到基于深度学习的方法，各种技术层出不穷。本文基于大语言模型，提出一种标准产业分类模型，以更有效地应对标准领域分类的复杂性，并适应不断多样化的产业分类需求，最终为标准大数据的分类管理及分类检索提供高效的解决方案。

分类模型的核心是表征学习技术，该技术是机器学习领域的核心，旨在学习数据的有效表示，从而更好地理解和利用数据[8]。随着人工智能的发展，在文本分类任务中，研究者们提出了两类主要的方法，分别是基于机器学习和基于深度学习的文本分类模型。

1 基于机器学习的分类模型

在文本分类领域，基于机器学习的分类模型一直扮演着重要的角色。这一类模型在处理自然语言文本时，通过传统的特征工程和经典的机器学习算法，取得了一系列显著成果。其中，特征工程是关键的步骤之一，其目的在于将文本转化为计算机可处理的形式，提取文本的关键信息。最典型的特征表示方法包括词袋模型（Bag-of-Words，BoW）和词频逆文档频率（TF-IDF）。在这些表示基础上，研究者们应用了支持向量机（Support Vector Machine，SVM）、朴素贝叶斯等传统的分类算法[18]。这些基于机器学习的文本分类模型在处理中小规模文本数据集时表现出色，取得了令人满意的分类性能。然而，基于机器学习的分类模型也面临一些挑战。首先，传统的特征工程往往需要大量的领域专业知识，而且在处理大规模、高维度的文本数据时，手动提取特征变得非常困难且耗时。其次，这类模型通常无法捕捉到文本中的深层语义信息，因为它们缺乏对文本全局结构和上下文关系的理解。因此随着深度学习的兴起，基于深度学

习方法逐渐成为新的研究热点，也为文本分类任务带来了全新的技术范式。

2 基于深度学习的分类模型

深度学习的兴起为表征学习带来了重大突破，深度神经网络在图像处理、自然语言处理、语音识别等领域实现了广泛应用。表征学习的基本思想是通过神经网络自动学习数据的最佳表示，从而取代了传统的手工特征工程方法，使机器能够更好地理解和处理各种类型的数据。随着深度学习方法的崛起，自动特征学习开始成为焦点。早期的Word2Vec词嵌入模型，将单词映射到连续向量空间，以捕获单词的语义信息[9]。它有两种主要变种模型，CBOW和Skip-gram，分别用于从上下文预测目标单词或从目标单词预测上下文。Word2Vec的训练允许单词在向量空间中相似的单词更加接近，使其广泛用于文本分类、语义搜索、情感分析等NLP任务，引领了自然语言处理领域的重大变革。

自2017年Transformer模型[8]问世以来，预训练模型开始引领表征技术的新时代。这一模型架构引入了多头自注意力机制，使其在自然语言处理领域取得了显著的突破。随后，BERT[4]的发布将预训练技术推向高潮，通过大规模语料库的无监督学习，为多个NLP任务提供了最先进的性能。从那时起，预训练模型如：GPT[5]、RoBERTa[10]、XLNet[11]等的不断涌现，将表征学习技术推向了更广泛的领域，包括计算机视觉和多模态任务。这些模型的成功阐释了预训练技术对于自动学习数据的强大能力，为解决各种复杂问题提供了新的可能性。

最近，代表性的大语言模型，如：GPT-4[12]，LLaMA[13]，已经成为人工智能研究领域的一大重要突破。这些模型展现出了令人瞩目的人机对话和任务求解能力，引发了广泛的关注和讨论。它们不仅在自然语言处理任务中表现出色，还在各种应用中取得了卓越的成绩。因此本文选择大语言模型作为表征模型进行产业分类。结合大语言模型进行多类型文献表征，能够实现更准确和语义丰富的文献分类，提高了文献资源的组织和管理效率。这一方法不仅有助于更好地理解文献内容，还可以为研究人员提供更多深入的领域洞察力，推动科学研究的进展。

3 基于大语言模型的标准文献分类

本文通过对不同类型的文献资源进行统一表征，已解决文献资源的异构性。通过大语言模型强大的语义表征能力对不同文献资源进行建模，使其具有标准化的数学表达，帮助不同类型文献资源的统一产业分类。本文提出的产业分类模型架构图如图1所示。该分类模型架构中选择LLaMA大语言模型作为文本表征模型，该模型基于Transformer架构中的Decoder，但相比Transformer，LLaMA有一些改进机制，如：预归一化、SwiGLU激活函数等。LLaMA模型中使用了多头注意力机制，该机制能用于捕获输入序列中的不同关注点。在该机制中，输入序列经过多个不同的自注意头，每个头都学习到不同的权重分布，可表示为：$MultiHead(Q,K,V)=Concat(head_1,head_2,\cdots,head_h)W^O$

其中，Q是查询矩阵，K是键矩阵，V是值矩阵，$head_1$表示第i个注意力头，W^O是输出权重矩阵。LLaMA模型旨在解决文献资源库中多类型文献的表征问题，包括标准、专利、论文等多种类型的文献。其次本文选择多层感知机作为分类器。在该分类模型中，多层感知机分类器发挥着至关重要的作用，利用其多层结构和学习到的权重参数，能够有效地对文本数据进行分类和预测。最终，该分类器可以输出文献属于每个类别的概率值。这一方法的关键优势在于其能够提供详细的类别概率信息，为文献资源的分类决策提供了更全面的参考。在对分类模型训练时，本文选择交叉熵损失函数和随机梯度下降算法更新模型参数。

交叉熵损失（Cross-entropy loss）函数是深度学习中常用的一种损失函数，通常用于分类问题。该函数度量了模型预测结果与实际结果之间的差异，是优化模型参数时的一个关键指标，可表示为：$\ell(x,y)=-\sum_{i=1}^{C}x_i\log y_i$

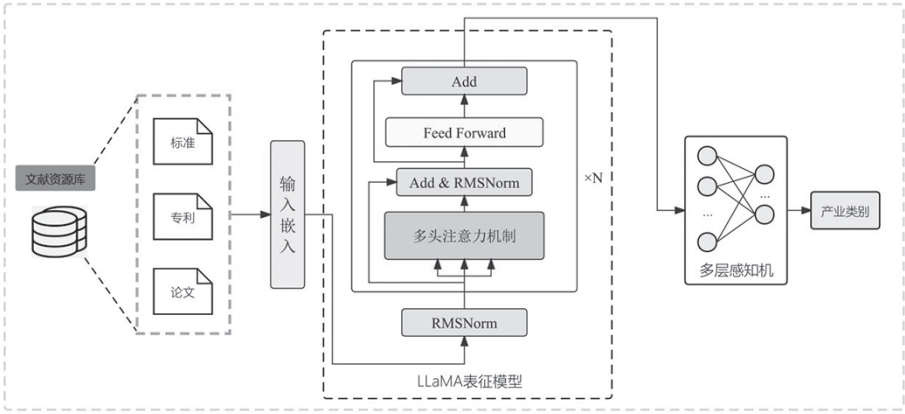

图1 标准文献产业分类模型架构图

其中，x_i是模型输入，y_i是目标标签，C表示分类数量。

本文采用微调（Finetune）技术对模型进行训练。微调这一技术的兴起始于基于Transformer架构的大型语言模型（LLM），如：GPT-4[12]和BERT[4]。微调技术是深度学习领域的关键技术，一直以来都受到广泛的研究和应用。通常，它指的是研究人员使用在大规模数据上进行了预训练的神经网络模型，在特定任务或领域上对其进行进一步的调整和优化的过程。这个过程使得模型能够在新任务上表现出色，因为在预训练阶段它已经学到了通用的特征和知识。微调技术在自然语言处理、计算机视觉、语音识别等领域都取得了显著的成功，其经过不断改进以适应不同类型的任务和数据，成为深度学习中的一个关键工具。在分类任务中，微调技术的重要性尤为显著。通过将预训练的模型以特定的分类任务进行微调，能够实现出色的产业分类性能，而无需从头开始构建和训练一个全新的模型。通过不断改进微调策略，例如：微调层次结构、调整学习率和优化损失函数等，进一步提高了分类性能。

4　实验结果与分析

4.1 标准文献数据集构建

本文采用山西省标准文献数据作为数据集。为提高文献资源的产业分类效果，需要对山西省科研攻关及生产管理过程中产生的标准文献进行产业标注，以实现更准确和可靠的文献分类。该任务的目标是将标准文献资源标注为14个产业，共15个

表1 数据集详情

所属产业类别	论文数据量	标准数据量	专利数据量	总数
半导体	323	707	501	1531
碳基新材料	587	765	432	1784
大数据融合创新	533	755	436	1724
信息技术应用创新	776	698	506	1980
特种金属材料	533	694	716	1943
煤机智能制造	786	656	731	2173
轨道交通装备制造	553	791	297	1641
通用航空	369	641	303	1313
新能源	302	377	377	1056
新能源汽车	376	330	330	1036
非常规天然气	314	522	320	1156
节能环保	369	386	262	1017
现代医药和大健康	458	180	243	881
有机旱作农业	556	8	327	891
其他	702	743	654	2099

类别，具见表1。在数据标注的过程中，针对每个领域，精心选择相关的关键词，这些关键词将成为标注的依据，确保它们能够充分反映每个领域的主题特点。随后，将标准文献根据其内容和关键词进行标注，以便将其归类到正确的产业类别中。这一标注工作可以借助领域专家的知识和经验来提高准确性。在进行模型训练时，本文将数据集划分为训练集、验证集和测试集，以便在模型开发和评估过程中有足够的数据支持，其中划分方式为随机从所有样本中采样70%的样本作为训练集，10%的样本作为验证集，20%的样本作为测试集。

4.2 实验结果

为了验证本文提出的模型相对于当前存在的基线模型的性能优势，研究中选择了4种不同的基线模型。

（1）WideMLP[14]：该模型是一个基于词袋的多层感知器，包含一个具有1024个线性单元的单一隐藏层。这个模型作为一个有用的基准，用于度量实际科学进展的水平。

（2）LSTM（Long short-term memory）[15]：LSTM是一种循环神经网络（RNN）的变体，专门设计用于解决传统RNN中梯度消失和梯度爆炸的问题。其通过精妙的门控结构，能够有效地捕捉和记忆长期依赖关系，使其成为处理序列数据的强大工具。

（3）DADGNN[16]：深度注意扩散图神经网络（DADGNN）是一种基于图的方法，旨在解决图神经网络中的过度平滑问题。它通过引入注意扩散机制，允许堆叠更多的层和采用解耦技术。这种解耦技术对于短文本尤其有利，因为它能够在深度图网络中捕获明显的隐藏特征。

（4）ConTextING-BERT[17]：将图神经网络（GNN）与BERT结合，提供基于文档的上下文嵌入，以用于归纳文本分类。此混合模型充分利用了GNN对图结构的建模能力和BERT对上下文的深层理解，使得文本分类任务在更丰富的语境下得以执行。

以上所选的基线模型包括一种传统机器学习模型（WideMLP），以及其他3个基线模型（LSTM、DADGNN、ConTextING-BERT），他们都是来自最新研究论文中的先进模型。这些模型的选择旨在覆盖不同的方法和技术，以全面评估提出模型的性能。通过与这些基线模型进行比较，可以更清晰地了解提出模型在文本分类任务中的优越性和创新之处。同时，本文选择在3个公开数据集进行模型比较，分别为R8、MR和SST-2数据集。R8是新闻数据集可用于8种类别的文本分类任务。MR是一个广泛用于文本分类的数据集，其中包含了平均长度为20.39个标记的影评文件。SST-2是情感库的一个子集，是一个细粒度的情感分析数据集，其中中立的评论已被删除，数据集仅包含积极或负面标签，因此可用于进行二分类文本分类任务。通过在这些不同领域和任务的数据集上进行模型比较，可以更全面地评估提出模型的泛化性能和适用性。

通过对比表2的实验结果，可以明显观察到本文所提出的模型在分类任务中表现更为卓越，这进一步证明了本文模型的有效性和性能优势。

表2 公共数据集中准确率对比

模型	数据集		
	R8	MR	SST-2
WideMLP	96.98	76.48	82.26
LSTM	96.34	74.99	79.95
DADGNN	97.28	74.54	82.81
ConTextING-BERT	97.91	86.01	–
Ours	98.06	86.77	84.25

为了进一步验证提出的模型在实际应用中的有效性，本文进行了多组实验，使用构建的数据集来评估模型在产业分类任务中的性能，同时使用精准率、召回率和F1值等指标来衡量模型的优越性。首先，观察不同迭代次数下模型的训练结果，见表3。随着迭代次数的增加，模型在测试集上的性能逐渐提升，直至模型趋于收敛。从表3中可以看出，模型在测试集上的F1值最高可达到98%以上。这表明我们的模型在产业分类任务中表现出色，具有高度的性能和实用性。

图2展示了本文提出的分类模型在测试集上得到的混淆矩阵，反映了模型的分类性能。从图中可以观察到，在所有类别中，该模型表现出超过95%的准确率。这表明提出的模型在各个类别上都取得了显著的分类准确性，展现了其在多类别文献分

表3 在标准文献数据集中训练轮次对模型结果影响

Data	Metric	轮次-2	轮次-4	轮次-6	轮次-8	轮次-10
标准文献数据集	Pre	93.74	97.04	97.46	98.12	98.28
	Rec	94.78	96.96	97.32	97.77	98.19
	F1	93.93	97.04	97.34	97.76	98.19

表4 在标准文献数据集中学习率超参数对实验结果影响(%)

Data	Metric	学习率（LR）				
		5e-7	1e-6	5e-6	1-e5	5e-5
文献资源数据集	Pre	42.41	68.67	95.75	97.96	98.20
	Rec	38.49	71.85	96.30	97.85	98.15
	F1	31.61	65.83	95.87	97.92	98.18

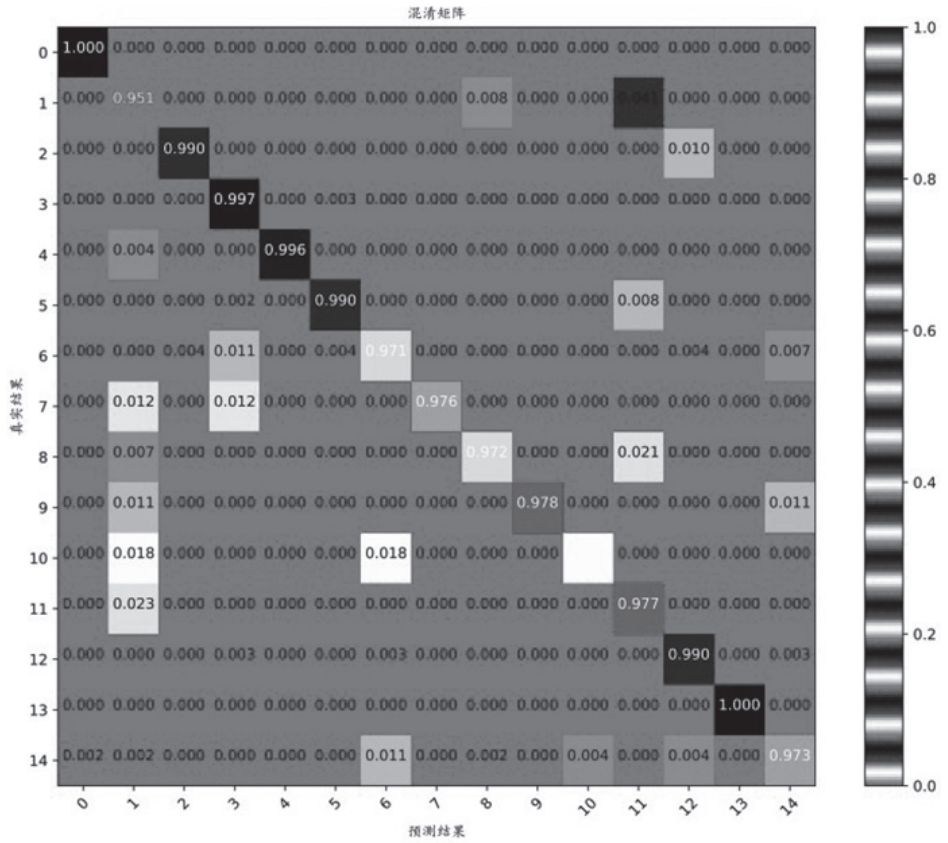

图2 在标准文献测试集中分类模型的测试结果

类任务中的优越性和鲁棒性。

此外，实验还进行了对比分析，研究了超参数中学习率（LR）对实验结果的影响。学习率在深度学习模型中扮演着至关重要的角色，它决定了模型在训练过程中参数更新的速度和方向。因此，学习率的选择对模型的性能和收敛速度具有显著影响。在表4中，本文呈现了经过5次迭代后，模型在不同学习率下对产业分类效果的影响。从表中可以清晰地看出不同学习率对模型的收敛速度和最终结果产生了显著的影响。这突出了学习率选择在深度学习中的重要性。因此，在实际应用中，合适的学习率选择需要谨慎权衡，以取得最佳的模型性能和训练效率。

5 结论

本文提出了一种基于大型语言表征模型的多类型文献资源产业分类方法，旨在应对标准文献按产业进行分类时的多语义、多类型、少标注的挑战。通过实验证明，所提出的模型有效提高了文献资源分类的准确性和效率。在方法的实施中，本文充分利用大型语言模型，对标准文献资源进行语义表征，从而生成了文献的丰富语义信息。该方法有助于解决文献中的多语义问题，使得模型能够更好地理解文献中的隐含语义和关联性。其次，本文采用多层感知机模型进行产业分类。最终，结合大型语言模型的语义表征和多层感知机的分类能力，本文提出的方法成功应用在山西“111”创新工程项目中，对于推动科技文献资源更好地应用于服务具有重要意义。

参考文献

[1] Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and Machines, 2020,30(2): 1-14.

[2] 杜宇晨.基于Word2Vec和N-Gram的短文本情感分类方法研究[D].杭州：浙江工业大学, 2018.

[3] 奉国和.SVM分类核函数及参数选择比较[J]. 计算机工程与应用, 2011,47(03):123-124+128.

[4] 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.

[5] Radford, A., Narasimhan, K., Salimans, T., et al. Improving language understanding by generative pre-training.

[6] 汪家伟,余晓. 基于深度学习的文本分类研究综述[J]. 电子科技, 2024(1):81-86.

[7] Yang F J. An implementation of naive bayes classifier[C]//2018 International conference on computational science and computational intelligence (CSCI). IEEE, 2018: 301-306.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[9] Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J]. arXiv preprint arXiv:1402.3722, 2014.

[10] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[11] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.

[12] OpenAI. GPT-4 technical report[R]. 2023.

[13] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.

[14] Galke L., Scherp A. Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP[C]. 2022 arXiv preprint arXiv:2109.03777.

[15] Long short-term memory[J]. Neural computation, 2010, 9(8): 1735-1780.

[16] Liu Y H, Guan R C, Giunchiglia F., Deep attention diffusion graph neural networks for text classification[C]//Proceedings of the 2021 conference on empirical methods in natural language processing.

[17] Huang Y H, Chen Y H, Chen Y S. ConTextING: Granting Document-Wise Contextual Embeddings to Graph Neural Networks for Inductive Text Classification[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 1163-1168.

[18] 李旭然,丁晓红. 机器学习的五大类别及其主要算法综述[J]. 软件导刊, 2019,18(07):4-9.