计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200023

基于多语言嵌入图卷积网络的仇恨言论检测方法

赵弘毅　李志远　卜凡亮
中国人民公安大学信息网络安全学院　北京 100045
(2022211430@ppsuc.edu.cn)

摘　要　随着社交媒体的广泛应用,网络仇恨言论的传播问题日益严重,尤其在网络匿名性的掩护下,仇恨言论得以快速扩散,为仇恨言论检测带来严峻挑战。为了有效应对这一问题,提出了一种基于多语言嵌入图卷积网络(Multi-language Embedding Graph Convolutional Network,MEGCN)的多语言仇恨言论检测方法。该方法充分融合了序列建模与图建模的优势,利用多语言预训练模型进行特征提取,从而能够处理不同语言间的复杂关系。同时,提出了一种基于插值预测的联合训练方式,以提升模型的准确性和鲁棒性。通过在4个公开数据集上的实验,结果表明,MEGCN相比所有对比模型,均在多语言仇恨言论检测任务中取得了更优的性能。该方法不仅能够保持较高的序列建模精度,还能够有效地捕捉文本间的结构性关系,进而提升模型在多语言环境中的表现,尤其在不同语言之间的语义对应关系方面展现出显著优势。

关键词:仇恨言论检测;图卷积网络;多语言预训练模型;自然语言处理
中图分类号　TP391

Multi-language Embedding Graph Convolutional Network for Hate Speech Detection

ZHAO Hongyi,LI Zhiyuan and BU Fanliang
School of Information Network Security,People's Public Security University of China,Beijing 100045,China

Abstract　With the widespread use of social media,the issue of the spread of online hate speech has become increasingly severe,especially under the cover of anonymity on the Internet,allowing hate speech to spread rapidly,posing a serious challenge to the detection of hate speech. In order to effectively address this issue,this paper proposes a cross-lingual hate speech detection method based on Multi-language Embedding Graph Convolutional Network(MEGCN). This method fully integrates the advantages of sequence modeling and graph modeling,and uses multi-language pre-trained models for feature extraction,thus being able to handle complex relationships between different languages. At the same time,this paper proposes a joint training method based on interpolation prediction to improve the accuracy and robustness of the model. Experiments on four public datasets show that MEGCN achieves better performance than all existing comparative models in the task of cross-lingual hate speech detection. This method not only maintains a high sequence modeling accuracy,but also effectively captures the structural relationships between texts,thereby improving the performance of the model in multi-language environments,especially in terms of semantic correspondence between different languages.

Keywords　Hate speech detection,Graph convolutional network,Multi-language pre-trained model,Natural language processing

1　引言

在全球化和社交媒体平台普及的背景下,网络仇恨言论的传播已对社会和谐与个人安全构成显著威胁。随着跨文化交流的增加,不同语言、种族和文化背景的人们频繁互动,虽然促进了全球沟通,但也加剧了仇恨言论的蔓延,给仇恨言论的检测带来了前所未有的技术挑战。现有的单语言检测方法难以应对这一多样化的语言环境,尤其在跨文化、多语言的场景中,情感表达、俚语、方言等差异加剧了模型识别的难度[1-2]。为此,构建一种高效、准确的多语言仇恨言论检测模型,对于遏制仇恨言论的跨国传播、保护不同文化群体的权益具有重要意义。

在仇恨言论检测建模过程中,主要采用两种方式:序列建模与图建模。如图1所示,序列建模将文本拆分为单词序列,通过预训练语言模型,如BERT和XLM,结合多头注意力机制对文本进行深度嵌入,已经在该领域展现出优异的性能和广泛的通用性[3]。与此不同,图建模将文本和单词视为节点,将单词间以及单词与句子之间的关系构建为边,通过点互信息(PMI)计算单词之间的边权重,并利用词频-逆文档频率(TF-IDF)计算句子与单词之间的边权重[4]。由于自然语言本质上具有非欧结构特性,图建模能够有效捕捉文本之间的结构关系,并揭示不同语言间的对应关系,为多语言仇恨言论检测提供了新的技术方向[5]。

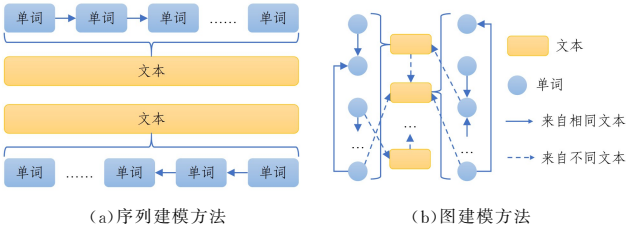
尽管图建模通过节点与边的结构较好地表达了不同语言之间的映射关系,但在捕捉多种语言的通用特征方面仍存在一定局限性。相比之下,序列建模通过构建一个包含多种语

基金项目:中国人民公安大学双一流创新研究项目(2023SYL08)
This work was supported by the Double First-Class Innovation Research Project of People's Public Security University of China(2023SYL08).
通信作者:卜凡亮(bufanliang@sina.com)

言的大规模词表，并将单词映射到统一的特征空间中，能够有效捕捉这些通用特征，但在揭示不同语言之间的映射关系方面有所不足。

多语言任务和跨语言任务是自然语言处理领域的两个重要研究方向。多语言任务涉及在多个语言上训练一个统一模型，使得该模型能够处理不同语言的文本。这种类型的任务假设不同语言的文本具有较强的代表性，且模型可以从多种语言中学习到通用特征[6]。但是，跨语言任务则更加关注如何在一种语言上训练模型，并使其能够在其他语言上取得良好的性能。跨语言任务通常涉及语言之间的映射和转移，面临的挑战在于如何有效利用源语言与目标语言之间的有限或无监督对应关系。

针对上述问题，本文提出了一种多语言嵌入图卷积网络（Multilingual Embedded Graph Convolutional Network，MEGCN），用于多语言仇恨言论检测。该方法结合了序列建模和图建模的优势，旨在提升模型在多语言环境下对仇恨言论的识别能力。具体而言，MEGCN 利用先进的多语言预训练模型（XLM）对文本进行特征提取，并且能够处理不同语言的多样化表达。随后，将这些特征嵌入图卷积神经网络，MEGCN 能够充分发挥序列建模在语言语义理解上的优势，确保文本中复杂的语义关系得到准确捕捉。此外，MEGCN 利用图卷积网络强大的非欧结构建模能力，能够有效提取文本中单词之间的相互关系、单词与句子之间的依赖性以及句子与句子之间的语义联系。在多语言情境下，图卷积网络还可以识别并对齐不同语言之间的语义模式，从而进一步增强模型的多语言检测能力。这种结构不仅保留了序列建模在捕捉语言语义特征上的精确性，还能够充分挖掘文本之间的潜在相似性和依赖性，提升仇恨言论检测在多语言情境下的表现。实验结果表明，MEGCN 在多个公开数据集上均表现出色，尤其在多语言文本分类任务中，相比于传统的单语言方法，显著提高了多语言仇恨言论检测的准确性和鲁棒性。



（a）序列建模方法　　（b）图建模方法

图 1　两种仇恨言论检测建模方式的区别

Fig. 1　Difference between the two types of hate speech detection modeling

本文的主要贡献如下：

1）提出了多语言嵌入图卷积网络（MEGCN），创新性地结合了序列建模和图建模方法，解决了现有方法在处理多语言文本时面临的语义对齐和结构建模的挑战。

2）引入了插值预测的联合训练策略训练多语言嵌入图卷积网络，有效解决了多语言数据集在训练过程中语言间的差异问题。

3）在多个公开数据集上进行的广泛测试验证了 MEGCN 在仇恨言论检测任务中的优越性能。实验结果表明，MEGCN 能够有效捕捉不同语言间的文本关系，减少语言障碍对检测效果的影响，并在多语言仇恨言论检测上优于传统方法，解决了多语言文本分类中的精度与可扩展性问题。

本文第 2 章总结了相关工作，第 3 章详细介绍了本文方法，包括图的构建过程和模型框架；第 4 章给出了实验过程与详细的结果分析；最后总结全文。

2　相关工作

2.1　多语言仇恨言论检测二级标题

在近年来的研究中，多语言仇恨言论检测逐渐成为自然语言处理领域的重要课题。Sebastian 等[7]在 2018 年提出了一种基于 Word2Vec 和扩展 2-gram 组合的方法来检测仇恨言论。文献[7]强调了迁移学习的重要性，尤其是从英语到其他低资源语言的迁移，这为提升多语言仇恨言论检测的效果提供了新的视角。Debora[8]则指出了零镜头跨语言仇恨言论检测的局限性，特别是非仇恨的禁忌感叹词常常被误解为仇恨言论的信号，作者认为模型设计时需考虑语言特有的表达方式与文化背景。Irina 等[9]通过结合双语文本嵌入与传统全监督神经网络（如 CNN 和 RNN），实现从英语到德语的跨语言仇恨言论检测。这种方法不仅提高了检测准确性，还展示了在多语言环境中有效利用双语数据的潜力。

随着基于 Transformer 模型[10]的多语言预训练模型的出现，如 MBERT[11]，XLM[12]和 XLM-R[13]等，越来越多的多语言仇恨检测任务开始依赖这些强大的预训练模型。Teodor 等[14]对多种多语言预训练模型在多语言仇恨言论检测任务上的能力进行了比较，结果表明，Transformer 架构在捕捉复杂语义关系方面具有显著优势。Yang 等[15]提出了 CINO 预训练模型，该模型在多种少数民族语言语料上基于 XLM-R 进行二次预训练，显著提升了对藏语、蒙语、维吾尔语、哈萨克语（阿拉伯体）、朝鲜语、壮语等少数民族语言与方言的理解能力。Sai 等[16]对多语言仇恨言论检测进行广泛评估，发现逻辑回归结合 LASER 嵌入在低资源场景中的表现最佳，但在高资源场景中，基于 BERT 的模型则展现出更强的能力。

2.2　用于自然语言处理的图神经网络

近年来，图神经网络（GNN）在自然语言处理（NLP）中的应用取得了显著进展。Yao 等[4]提出的 TextGCN 首次将图卷积神经网络应用于文本建模，取得了与传统序列建模方法相当的效果。随后，Lin 等[17]提出了 BertGCN，将预训练模型与图神经网络结合，通过 BERT 的特征嵌入提升文本表示能力，从而提高模型的性能。Yang 等[18]提出了异质图注意力网络（HGAT），创新性地将短文本构建为异质图结构，解决了数据稀疏和歧义问题，展示出图神经网络在处理复杂文本关系时的潜力。Wu 等[19]对 GNN 在 NLP 中的应用进行全面综述，讨论当前面临的挑战及未来研究方向，提出了图结构建模在语义理解和推理中的优势，同时指出计算效率和可解释性方面存在的挑战。近年来，诸如 GraphBERT[20]和 GAT2[21]等工作进一步提升了图神经网络在文本分类、情感分析等任务中的表现，同时也提出了改进的图卷积结构和图注意力机制，以更好地捕捉文本中的深层语义关系。最近的研究表明[22]，层次化图卷积网络和多模态图神经网络在处理复杂文本任务中的潜力不断被挖掘，为 NLP 任务提供了新的思路与解决方案。

3　本文方法

本章将详细阐述所提出的多语言嵌入图卷积网络

MEGCN，包括文本图的构建过程以及详细的模型框架设计思路。具体而言，首先利用语料库构建一个多语言文本图，并通过预训练语言模型提取文本特征，将其嵌入文本图中。其次，通过图卷积操作实现文本和单词节点之间信息的有效聚合。最终，利用插值预测的联合训练策略优化整个模型的性能。MEGCN 的框架如图 2 所示。

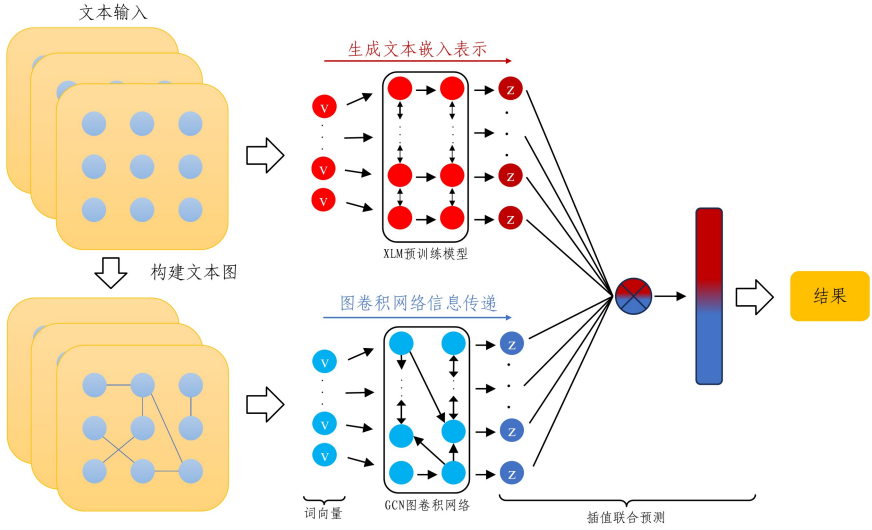


图 2　MEGCN 模型框架

Fig. 2　Framework of MEGCN model

3.1　文本图构建

本文构建的异构图包含两种节点类型：文本节点和单词节点，以及 3 种边类型：单词-单词边、单词-文本边和文本-文本边。每种边都带有相应的边权重，用于量化节点间的关系。通过这种方式，模型能够综合考虑节点的语义信息及其在图中的结构关系，从而有效提高文本分类的表现。

3.1.1　单词-单词边

在构建单词-单词边时，利用点互信息（Pointwise Mutual Information，PMI）作为衡量单词之间联系的度量[4]。具体来说，首先使用滑动窗口方法对文本进行扫描，并记录单词的共现情况。设定滑动窗口总数为 W，其中 $W(i)$ 表示窗口中单词 i 出现的次数。单词 i 的独立出现概率可以表示为 $P(i)=\frac{W(i)}{W}$。此外，$W(i,j)$ 表示窗口中同时出现单词 i 和 j 的次数，共现概率为 $P(i,j)=\frac{W(i,j)}{W}$。

基于上述概率表示，任意两单词之间的 PMI 值可以通过式(1)计算：

$$PMI(i,j)=\begin{cases}\log_2\left(\dfrac{P(i,j)}{P(i)P(j)}\right), & \text{if } P(i,j)\geqslant k\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

其中，k 是一个超参数，用于设定一个阈值，只有当共现概率 $P(i,j)$ 超过该阈值时，才会计算其 PMI 值。经过实验验证，当 $k=0.4$ 模型表现最佳。这样的策略有助于避免计算共现频次较低的单词对，从而减少计算开销并降低噪声的影响。

3.1.2　单词-文本边

在构建单词-文本边时，使用词频-逆文档频率（TF-IDF）来衡量单词与文档之间的关联强度[17]。TF-IDF 是衡量单词在文档集合中重要性的常用统计方法，其计算过程如式(2)：

$$TF\text{-}IDF(t,d)=TF(t,d)\times IDF(t) \tag{2}$$

其中，$TF(t,d)$ 是单词 t 在文档 d 中的词频，定义为单词 t 在文档 d 中出现的次数除以文档中所有单词的总数；$IDF(t)$ 是单词 t 的逆文档频率，计算过程可表示为：

$$IDF(t)=\log\left(\frac{N_t+1}{N_t}\right) \tag{3}$$

其中，N 为文档总数，N_t 为包含单词 t 的文档数。TF-IDF 通过加权单词的频率与它在语料库中的稀有度，能够有效反映单词在文档中的重要性。

3.1.3　文本-文本边

为深入挖掘文本之间的内在联系，本文提出了一种基于文本-文本边的边构建方法。首先利用多语言嵌入模型（如 XLM）对每个文本进行向量化，得到每个文本的密集嵌入表示。具体来说，设文本集合为 T，每个文本 t 通过嵌入模型 E 映射为一个 d 维向量 $\boldsymbol{v}_t$，即 $\boldsymbol{v}_t=\boldsymbol{E}(t)$，其中，$t\in T$，$\boldsymbol{v}_t\in R^d$，$\boldsymbol{E}(t)$ 表示文本 t 的嵌入向量表示。随后，通过计算文本嵌入向量之间的余弦相似度来衡量文本之间的语义关联，计算方法为式(4)：

$$\cos(\boldsymbol{v}_t,\boldsymbol{v}_s)=\frac{\boldsymbol{v}_t\cdot\boldsymbol{v}_s}{\|\boldsymbol{v}_t\|\,\|\boldsymbol{v}_s\|} \tag{4}$$

其中，$\boldsymbol{v}_t$ 和 $\boldsymbol{v}_s$ 分别表示文本 t 和文本 s 的嵌入向量，$\|\cdot\|$ 表示向量的欧几德得范数，$\cdot$ 表示向量的点积操作。

为了构建文本-文本之间的边，本文为每个文本 t 找到与其相似度最高的 L 个文本，从而形成边的集合 B。边的权重 $w(t,s)$，即为文本 t 和文本 s 之间的余弦相似度，计算过程为：

$$w(t,s)=\begin{cases}\cos(\boldsymbol{v}_t,\boldsymbol{v}_s), & \text{if } s\in Top\text{-}L(t)\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

其中，$Top\text{-}L(t)$ 表示与文本 t 相似度最高的 L 个文本集合。该方法通过捕捉文本之间的语义相似度，能够增强文本间的关系表达能力。总之，本文提出的异构图的构建方法通过构建不同类型的边，使得 MEGCN 模型能够从多个层面进行特征聚合和信息传播。单词-单词边通过点互信息（PMI）来衡量词语之间的关系，单词-文本边通过 TF-IDF 来表示单词和文档之间的相关性，文本-文本边则通过文本嵌入向量之间的余弦相似度来衡量文本间的语义相似性。这样的构建方法不

仅提高了多语言文本之间的语义对齐和表示能力，还能精确地计算文本之间的语义相似度，优化了边的构建过程，在多语言仇恨言论检测任务中，提升了模型的性能和泛化能力。

3.2 MEGCN

MEGCN 的核心思想是将多语言预训练模型 XLM 生成的文本嵌入与图卷积网络 GCN 结合，以充分利用 XLM 在文本语义理解方面的优势，并发挥 GCN 在捕捉图结构关系方面的能力。在传统的多语言预训练模型和图卷积网络结合中，直接拼接两者会导致图卷积网络的节点向多个节点传播梯度，造成计算难度增加。因此，MEGCN 通过插值策略联合训练两种模型，减少了计算复杂度，提升了模型的训练效率与预测能力，从而在文本分类任务中显著提高了性能，尤其是在不同语言的文本语义和结构关系建模上。

此外，传统的多语言预训练模型与图卷积网络结合时，通常直接拼接这两个模型的输出。然而，这种拼接方式可能导致图卷积网络的节点需要向多个节点传播梯度，造成 XLM 模型同时更新多个不同梯度，计算负担极大。因此，MEGCN 模型引入了一种插值策略，避免了直接拼接造成的计算开销，并且能够在训练过程中动态调整 XLM 和 GCN 模型的贡献，使得模型既能利用语义信息，也能有效聚合图结构信息。其中，插值策略的引入不仅减少了计算量，还使得模型能够根据具体任务的需要调整两个模型的贡献比例。通过设置超参数 λ，模型能够灵活地调节 XLM 和 GCN 在最终预测中的贡献比重，确保在仇恨言论检测任务中获得最佳性能。

3.2.1 文本图节点初始化

在 MEGCN 模型的初始阶段，利用 XLM 预训练模型对输入的文本 T 进行处理，生成文本的嵌入表示。XLM 是基于 Transformer 架构的大规模预训练模型，旨在处理多语言文本。通过 XLM，文本 T 被转换为高维度的密集特征向量 $\boldsymbol{X}_0$，并将所有文本节点初始化为这些由 XLM 生成的特征向量。这一过程的计算方法为：

$$\boldsymbol{X}_0 = f(T) \tag{6}$$

其中，$f(T)$ 表示通过 XLM 生成的文本嵌入，$\boldsymbol{X}_0$ 是文本节点的初始特征向量。这些初始特征向量携带文本的语义信息，能够为后续的图卷积操作提供丰富的语义基础。同时，词语节点的初始特征向量被设置为零向量，表示在初始化时，词语节点不包含任何语义信息。随着模型训练的进行，词语节点会通过图卷积更新其特征，逐步融入图中的语义信息。

3.2.2 图卷积网络信息传递

在图卷积网络(GCN)部分，模型通过多层的图卷积操作对节点的特征向量进行更新和传播。每一层图卷积网络通过邻接矩阵 $\boldsymbol{A}$ 和度矩阵 $\boldsymbol{D}$ 来描述节点之间的连接关系，并利用这些矩阵对节点特征进行更新。第 n 层到第 $n+1$ 层的节点特征更新规则为：

$$X^{(n+1)} = \sigma(D^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{\frac{1}{2}}X^{(n)}\boldsymbol{\omega}^{(n)}) \tag{7}$$

其中，$\omega^{(n)}$ 为第 n 层的权重矩阵，σ 表示激活函数(sigmoid)，$\boldsymbol{A}$ 表示图的邻接矩阵，$\boldsymbol{D}$ 表示度矩阵，X 表示第 n 层的节点特征。通过这种更新规则，每个节点都能够从邻接节点接收信息，并根据邻接节点的特征进行更新。具体地，$\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{\frac{1}{2}}$ 表示图的归一化邻接矩阵，其能够有效调整图中不同节点的连接强度，使得信息传播更加平衡。图卷积操作确保了节点的特征不仅能够体现自身的信息，还能够反映其邻接节点的特征，从而使节点表示更加丰富。

通过多层的图卷积网络，节点会逐步聚合来自其邻居节点的信息，捕捉到图结构中的潜在模式和节点间的依赖关系。这种信息传播机制使得 MEGCN 模型能够在图结构上进行有效的特征学习，为后续的分类任务提供更强的特征表示。

3.2.3 模型预测

在模型预测策略上，本文通过梯度传播路径的工程化设计，实现 XLM 多语言预训练模型与图卷积网络(GCN)的协同优化。分别提取文本节点的特征 X 和通过 XLM 生成的特征 X_0。这两个特征分别经过 sigmoid 激活函数和线性变换层处理，得到两个预测值：文档节点的预测值 Z_1 和 XLM 生成的预测值 Z_2。计算过程如式(8)所示：

$$Z_1 = \sigma(\boldsymbol{\omega}_1 X), Z_2 = \sigma(\boldsymbol{\omega}_2 X_0) \tag{8}$$

其中，$\boldsymbol{\omega}_1$ 和 $\boldsymbol{\omega}_1$ 分别表示两个线性层的权重矩阵，它们分别控制 GCN 和 XLM 生成的特征在最终预测中的影响。通过这种方式，模型能够同时基于 GCN 和 XLM 提供的不同特征进行预测，最大化地利用两种模型的优势。

为实现高效联合训练，MEGCN 使用分路径梯度控制机制：对于 GCN 路径的预测结果 Z_1，其梯度仅反向传播至 GCN 参数；而 XLM 编码器的参数更新仅通过直连路径的预测结果 Z_2 进行反向传播。梯度的这种创新分离策略既保证了两个模块的协同训练，又避免了传统级联结构中因多层参数耦合带来的计算复杂度问题。若采用常规的端到端反向传播，每次图结构更新都将触发 XLM 的全参数更新，其时间复杂度将随节点数呈指数级增长。预测结果通过可学习的插值系数进行动态融合，最终的预测结果 Z 通过式(9)计算得到：

$$Z = \lambda Z_1 + (1-\lambda)Z_2 \tag{9}$$

其中，λ 是一个超参数，用于调节 XLM 和 GCN 在最终预测中的贡献比重。通过这种加权机制，模型可以根据具体任务的需求，动态调整两个模型的贡献，以实现最佳的分类效果。明显地，当 $\lambda=1$ 时，模型完全依赖 XLM 的预测结果；当 $\lambda=0$ 时，模型完全依赖 GCN 的预测结果；而在 $0<\lambda<1$ 的情况下，模型的最终预测结果是 XLM 和 GCN 预测结果的加权组合。这种方式不仅增强了模型对图结构信息和语义信息的综合表达能力，还能根据任务需求优化特征融合，从而有效提高多语言文本分类任务的性能，解决了现有方法在特征融合和多任务适应性方面的不足。

3.2.4 损失函数

为了评估模型的性能并优化参数，本文利用二分类交叉熵损失函数(Binary Cross-Entropy Loss)作为优化目标，来衡量预测结果与真实标签之间的差异。该损失函数的计算式为：

$$L_{\mathrm{BCE}} = -y\log(Z) - (1-y)\log-(1-Z) \tag{10}$$

其中，y 表示真实标签，Z 表示模型的预测值。交叉熵损失函数可以有效度量模型预测的准确性，并通过反向传播优化模型的参数，使得模型在训练过程中逐步提高分类性能。

3.3 可行性与优势分析

本文结合 XLM 模型与 GCN 模型进行多语言仇恨言论检测，具有以下优势。

1)图卷积网络与预训练模型结合的优势：通过结合 XLM

的高质量语义表示与 GCN 的图结构信息，MEGCN 模型能够同时利用文本的全局语义和局部结构信息，通过图卷积网络持续优化基于优质语义特征的结构建模，从而显著提升多语言文本分类的性能。该方法的优势在于通过插值融合的策略，有效地减少了计算量，并保证两个模型的互补优势。

2）训练优化：在训练过程中，为避免图卷积网络和预训练模型的计算瓶颈，采用插值策略并引入交叉熵损失函数来进行模型优化。通过梯度路径的物理隔离避免了参数更新冲突，同时平衡 GCN 与 XLM 的贡献，能够灵活调整模型的表现，适应不同的任务需求。

3）内存和计算效率：通过对节点特征进行有效的更新和优化，并使用超参数 λ 来调整模型的权重，MEGCN 在保证高性能的同时，也能保持 XLM 参数的稳定更新，避免被高频的图结构更新干扰。

4 实验与结果分析

本章通过全面的实验来验证 MEGCN 在多语言仇恨言论检测任务上有效性。4.1 节介绍了实验设置，4.2 节给出了对比实验的结果分析，4.3 节探究了不同微调策略的影响，4.4 节分析了模型部分重要参数的敏感性。

4.1 实验设置

4.1.1 数据集

为体现模型的特性并检验其在多语言环境下的仇恨言论识别能力，本文采用利用多种语言的数据集进行实验，涵盖了汉语、英语、法语和西班牙语。各数据集的详细信息如下。

1）COLD（Chinese Offensive Language Detection）[23]：COLD 是一个汉语仇恨言论数据集，包含 37 480 条带有二进制攻击性标签的评论，涵盖种族、性别和地区等多种主题。

2）SE（SemEval 2019-Task 5）[24]：该数据集源自 SemEval 2019 任务 5，该任务的目标是检测从 Twitter 中提取的西班牙语和英语信息中，针对移民和妇女的仇恨言论。该任务由两个相关的子任务组成：一个二元子任务用于检测仇恨言论的存在，另一个细粒度子任务用于识别仇恨内容中的其他特征，例如攻击性态度和被骚扰的目标，以区分是对个人的攻击还是针对群体的煽动。

3）FHSS（French Hate Speech Superset）：FHSS 是一个由多个法语仇恨言论数据集组成的集合，数据来源于多个社交平台，涵盖骚扰、性别歧视、种族歧视等类别的仇恨言论。该数据集从 Huggingface 平台下载，包含以下子数据集：MLMA 数据集[25]、CAA 数据集[26]、FTR 数据集[27]、“An Annotated Corpus for Sexism Detection in French Tweets”数据集[28]、UC-Berkeley-Measuring-Hate-Speech 数据集[29]（从英语翻译而来）。由于正负样本数量差异较大，实验中对数据进行了筛选。

这些数据集为本文的多语言仇恨言论识别任务提供了丰富的实验基础，使得模型能够在不同语言背景下进行有效评估和比较。

4.1.2 对比基线

本文使用 MBERT，XLM 与 XLM-R 作为对比基线模型，以下是对这 3 种模型的详细说明。

1）MBERT（Multilingual BERT）[11]：MBERT 是一种基于 BERT 架构的多语言预训练模型，采用多语言的参数共享机制，能够处理多种语言的文本数据。在训练过程中，MBERT 通过掩盖语言模型（MLM）和下一句预测（NSP）任务，学习不同语言的通用表示。这使得该模型在多语言任务中表现出了较为优异的性能，能够有效地处理多种语言的文本理解任务。

2）XLM（Cross-lingual Language Model Pretraining）[12]：XLM 是一种跨语言预训练模型，基于 Transformer 架构，并结合了双语词嵌入和跨语言注意力机制来捕捉不同语言之间的关联。XLM 在训练过程中采用了双语语言模型（Bilingual LM）和翻译语言模型（Translation LM），使得模型能够有效地学习不同语言之间的语义信息，从而提升其跨语言理解的能力。

3）XLM-R（XLM-Roberta）[13]：XLM-R 是基于 RoBERTa 架构的跨语言预训练模型，它在 XLM 的基础上进行了改进，采用了更大规模的模型容量和更丰富的训练数据。通过训练大量的跨语言语料，XLM-R 能够学习到更加丰富的跨语言表示，因此在多种跨语言任务中取得了优异的性能，尤其在处理不同语言的文本理解和生成任务时展现了较强的能力。

4）SS-GAN-mBERT[30]：SS-GAN-mBERT 结合生成对抗网络（GANs）和预训练语言模型（PLMs），从预训练的 PLM 模型出发，引入了 GAN 层以执行半监督学习。通过对抗生成器和判别器来改进分类效果，证明了其有效性。在多语言、零样本跨语言和单语言训练场景中均表现出显著的高性能。

此外，为避免参数量对实验结果的影响，本文还构建了 XLM-R＋CNN 和 XLM-R＋BiLSTM 作为对比模型。这两个模型在 XLM-R 的基础上分别增加了卷积神经网络（CNN）和双向长短时记忆网络（BiLSTM）结构，以探究不同神经网络结构对模型性能的影响。通过对比实验可以更准确地评估图卷积神经网络在本文模型中的贡献。

4.1.3 评价指标

在本文中，模型的性能评估选用了准确率（Accuracy，Acc）和 F1 值（F1-score）作为主要衡量标准。

1）准确率（Acc）：准确率是指模型预测正确的样本数占总样本数的比例。其计算过程如下：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

其中，TP（True Positive）表示真正例数，TN（True Negative）表示真负例数，FP（False Positive）表示假正例数，FN（False Negative）表示假负例数。

（2）F1 值（F1-score）：F1 值是精确率（Precision）与召回率（Recall）的调和平均数，是一种综合考虑分类模型性能的指标。其计算过程如式（12）－式（14）所示：

$$Precision=\frac{TP}{TP+FP} \tag{12}$$

$$Recall=\frac{TP}{TP+FN} \tag{13}$$

$$F1\text{-}score=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{14}$$

其中，精确率（Precision）表示所有被分类为正类的样本中，实际为正类的比例；召回率（Recall）表示所有真实为正类的样本中，模型能够正确分类为正类的比例。为确保评估结果的稳

定性和可靠性，本文进行了 5 次独立实验（使用 10 倍交叉验证），并对每次实验的结果取平均值，所有评估指标（Acc 和 F1 值）的结果保留 4 位小数。

4.1.4 实验环境

本实验在 Windows 10 操作系统上进行，硬件环境包括 Intel Core i7 处理器，RTX 3090 GPU（24 GB）。在 Python 3.8，PyTorch 1.18 的环境下搭建模型，并利用 Transformers，NumPy，scikit-learn 等相关库。训练过程中，堆叠 3 层 GCN，模型训练的 epoch 设置为 10，batch size 为 64，优化算法选择 AdamW，学习率为 5×10^{-4}，权重衰减系数为 0.01。在预测阶段，超参数 λ 设置为 0.7，用于平衡图建模与序列建模插值。本实验选用了 XLM-R 作为嵌入模型，确保了高效的训练与推理性能，并且，所选取的对比模型均按照原文网络结构与参数设置进行复现。

4.2 对比实验

本文在 4 个公开数据集 COLD（汉语）、SE（英语）、FHSS（法语）、SE（西班牙语）上进行充分的对比实验，实验结果如表 1 所列。

表 1 各方法的仇恨言论检测结果比较

Table 1 Comparison of hate speech detection results of each method

方法	COLD（汉语）		SE（英语）		FHSS（法语）		SE（西班牙语）	
	Acc	*F1-score*	*Acc*	*F1-score*	*Acc*	*F1-score*	*Acc*	*F1-score*
mBERT	0.7915	0.7808	0.4966	0.6175	0.9250	0.9192	0.7710	0.7246
XLM	0.7649	0.7641	0.5150	0.6192	0.9311	0.9248	0.7773	0.7324
XLM-R	0.7894	0.7796	0.5397	0.6282	0.9373	0.9317	0.7807	0.7450
XLM-R+CNN	0.8113	0.8083	0.5226	0.6296	0.9357	0.9294	0.7813	0.7469
XLM-R+Bi-LSTM	0.8197	0.7995	0.5341	0.6325	0.9380	0.9333	0.7863	0.7489
SS-GAN-mBERT	0.8132	0.7938	0.5136	0.6175	0.9314	0.9227	0.7880	0.7369
MEGCN(ours)	0.8235	0.7983	0.5793	0.6242	0.9466	0.9373	0.7902	0.7553

可以看到，MEGCN 方法在多语言文本分类任务中展现出显著优势，其性能在 4 种语言数据集上均优于对比模型。通过引入动态插值策略，MEGCN 在语义建模和结构关系捕捉之间实现了高效平衡，并在多语言环境中展现了强大的泛化能力。在中文数据集 COLD 上，MEGCN 的 Acc 达到 0.8235，F1-score 为 0.7983，显著优于基线模型 XLM-R（0.7894 和 0.7796）以及结合神经网络的改进模型 XLM-R+BiLSTM（0.8197 和 0.7995），说明了其在高资源语言任务中的细粒度语义建模能力。在低资源英语数据集 SE 上，MEGCN 克服了资源稀缺的限制，准确率达到 0.5793，F1-score 为 0.6242，相比 XLM（0.5150 和 0.6192）和 XLM-R（0.5397 和 0.6282）均有明显提升，表明 MEGCN 在低资源语言中更好地整合了不同语言表示和结构特征。在法语（FHSS）和西班牙语（SE）数据集中，MEGCN 的表现同样卓越，其在 FHSS 数据集上的准确率达到 0.9466，F1-score 为 0.9373，略高于 XLM-R+BiLSTM（0.9435 和 0.9333）；在西班牙语数据集 SE 中，其准确率为 0.7902，F1-score 为 0.7553，优于所有对比模型。

此外，MEGCN 在各语言数据集上优越的一致性，源于其创新性的设计——利用 XLM 的强大语义表示能力作为基础，通过图卷积网络（GCN）捕捉文本间隐含的结构关系，并通过插值策略动态调整两部分的贡献比例。这种设计既提升了模型对不同语言语义特征的敏感度，又降低了传统拼接策略带来的计算复杂度，确保了模型的高效性与适应性。此外，与单纯增加神经网络结构（如 CNN 和 BiLSTM）的方式相比，MEGCN 能够更好地平衡语义信息与结构关系，使其在低资源语言环境中仍然表现强劲。

4.3 不同微调策略的影响

为验证不同微调策略对模型性能的影响，本文设计了一系列消融实验。这些实验包括以下 4 种情况：1）随机初始化嵌入向量；2）不微调嵌入模型而直接联合训练；3）微调嵌入模型后冻结参数进行图卷积训练；4、微调嵌入模型后使用插值联合训练模型。本文利用 FHSS 数据集进行上述实验，实验结果如表 2 所列。

总体来看，随机初始化嵌入向量的模型表现最差，准确率（Acc）仅为 0.8776，F1 得分为 0.7826，表明随机初始化缺乏足够的语义信息支持，导致模型难以有效学习任务相关的特征。相比之下，不微调嵌入模型而直接联合训练的策略性能有所提升，Acc 达到 0.8983，F1 提高至 0.8354，这说明预训练的嵌入模型为特征学习提供了一定的帮助，但由于缺乏任务相关的微调，未能充分释放潜力。其次，采用微调嵌入模型后冻结参数进行图卷积训练的策略显著提升了模型性能，Acc 从前两种策略的不足 0.9 大幅提高至 0.9295，F1 则达到 0.9078，表明微调后的嵌入模型在任务数据上的表现更优，有效增强了图卷积网络的输入质量。然而，最终采用插值联合训练嵌入模型与图卷积模型的策略实现了最佳效果，Acc 进一步提升至 0.9466，F1 高达 0.9373，展现出最优的任务适配能力。

表 2 不同微调策略对 Acc 和 F1-score 的影响

Table 2 Effects of different fine-tuning strategies on Acc and F1-score

指标	微调策略			
	随机初始化	不微调嵌入模型	冻结嵌入模型	插值联合训练
Acc	0.8776	0.8983	0.9295	0.9466
F1-score	0.7826	0.8354	0.9078	0.9373

综上，插值联合训练策略在准确率和 F1 得分上均显著超越其他方法，这表明其不仅能够动态结合嵌入模型与图卷积模型的特征，还能够充分利用两者的互补优势。这种方法在不同特征表示的协同优化上表现出色，为处理复杂图数据提供了强有力的技术支持。

4.4 敏感性分析

1）如图 3 所示，插值策略中的参数 λ 对模型性能影响显著。随着 λ 从 0.1 增加到 0.7，模型准确率逐步提升，在 $\lambda=0.7$ 时达到最高值 0.9466，表明此时 XLM 和 GCN 的预测结果被合理融合，发挥了各自的优势。当 λ 超过 0.7 时，准确率开始下降，这可能是因为模型过于依赖 XLM 的预测结果，弱

化了 GCN 在图结构关系建模中的作用。因此，λ 的最佳取值为 0.7，可以在 XLM 和 GCN 的语义和结构信息之间实现平衡。

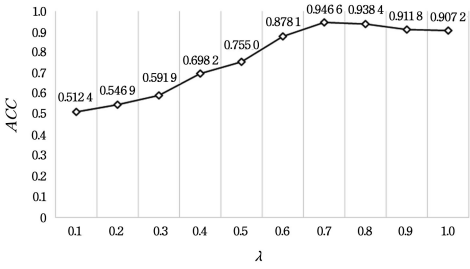


图 3　参数 λ 对准确率的影响

Fig. 3　Effect of λ on accuracy

2)从表 3 中可以看到，在 GCN 层数的实验中，层数从 1 增加到 3 时，准确率逐渐提升，在 3 层时达到最佳值 0.9466，这表明适度增加层数可以更充分地聚合节点之间的信息。但当层数超过 3(如 4 或 5 层)时，准确率开始下降，原因是过度平滑问题导致节点特征同质化，从而影响了模型性能。因此，GCN 的最佳层数为 3，能够在局部与全局信息聚合之间取得良好的平衡。

表 3　GCN 层数对准确率的影响

Table 3　Effect of GCN layers on accuracy

GCN 层数	*ACC*
1	0.7342
2	0.8626
3	0.9466
4	0.9275
5	0.9087

3)batch size 的调整实验表明(见表 4)，当 batch size 为 32 和 64 时，模型分别达到了较高的准确率，其中 batch size 为 64 时准确率最高，为 0.9466。这说明较小的 batch size 能保证梯度更新的频繁性，有助于提升模型的收敛效果。而当 batch size 增大至 128 或 512 时，准确率有所下降，这表明较大的 batch size 降低了模型的学习效率。综上，batch size 为 64 是性能与计算效率的最佳平衡点。

表 4　GCN 层数对准确率的影响

Table 4　Effect of GCN layers on accuracy

batch size	*ACC*
32	0.9233
64	0.9466
128	0.9134
512	0.8907

结束语　本文提出了一种基于多语言嵌入图卷积神经网络(MEGCN)的仇恨言论检测方法，能够有效应对社交媒体中多语言仇恨言论文本带来的挑战。该方法通过创新性地结合序列建模和图建模的优势，并引入插值预测的联合训练策略，实现了文本与单词之间信息的有效聚合。实验结果表明，MEGCN 不仅能保持鲁棒的序列建模能力，还能捕捉文本间的结构关系以及不同语言间的对应关系，有效提升多语言仇恨言论检测的性能。但是，图构建方法和图聚合策略仍有优化空间，未来将探究基于随机游走的方法与图结构感知 Transformer 的优化算法，以进一步提升模型的检测精度和泛化能力。

参 考 文 献

[1] AGOSTINA C, LEONARDO N, NEIL S, et al. Explainability and Hate Speech: Structured Explanations Make Social Media Moderators Faster[C]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2024: 398-408.

[2] WANG X L, WANG Y H, ZHANG S X, et al. Gender Discrimination Speech Detection Model Fusing Post Attributes[J]. Computer Science, 2024, 51(6): 338-345.

[3] CHEN H Y, ZHANG L. Very Short Texts Hierarchical Classification Combining Semantic Interpretation and DeBERTa[J]. Computer Science, 2024, 51(5): 250-257.

[4] YAO L, MAO C S, LUO Y. Graph Convolutional Networks for Text Classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 7370-7377.

[5] HUANG R, XU J. Text Classification Based on Invariant Graph Convolutional Neural Networks[J]. Computer Science, 2024, 51(S1): 120-124.

[6] STEPHEN M, EMANUELA B, ANTOINE D, et al. Multilingual Epidemiological Text Classification: A Comparative Study [C]// International Conference on Computational Linguistics (COLING). 2020: 6172-6183.

[7] SEBASTIAN K, DENNIS M R, STEFFEN H, et al. Discussing the Value of Automatic Hate Speech Detection in Online Debates[C]// Multikonferenz Wirtschaftsinformatik. 2018.

[8] DEBORA N. Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2021: 907-914.

[9] IRINA B, VIKTOR H, ALEXANDER F. Cross-Lingual Transfer Learning for Hate Speech Detection[C]// Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Kyiv. Association for Computational Linguistics, 2021: 15-25.

[10] ASHISH V, NOAM S, NIKI P, et al. Attention is All You Need [J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[11] WU S H, DREDZE M. Are All Languages Created Equal in Multilingual BERT? [C]// Proceedings of the 5th Workshop on Representation Learning for NLP. Association for Computational Linguistics, 2020: 120-130.

[12] LAMPLE G, ALEXIS C. Cross-lingual Language Model Pretraining[J]. arXiv: 1901.07291, 2019.

[13] GCONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised Cross-lingual Representation Learning at Scale[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 8440-8451.

[14] TEODOR T, ZUBIAGA A. Cross-lingual Hate Speech Detection using Transformer Models[J]. arXiv: 2111.00981, 2021.

[15] YANG Z Q, XU Z H, CUI Y M, et al. CINO: A Chinese Minority PRE-trained Language Model[C]// Proceedings of the 29th

International Conference on Computational Linguistics. International Committee on Computational Linguistics, 2022: 3937-3949.

[16] SAI S A, BINNY M, PUNYAJOY S, et al. A Deep Dive into Multilingual Hate Speech Classification[C]// European Conference on Machine Learning and Knowledge Discovery in Databased. 2021: 423-439.

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

[18] YANG T C, HU L M, SHI C, et al. HGAT: Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification[J]. ACM Transactions Information Systems, 2021, 39(3): 1-29.

[19] 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.

[20] ZHANG J, ZHANG H, SUN L, et al. Graph-Bert: Only Attention is Needed for Learning Graph Representations[J]. arXiv: 2001.05140, 2020.

[21] SHAKED B, URI A, ERAN Y. How Attentive are Graph Attention Networks? [J]. arXiv: 2105.14491, 2021.

[22] YAO L, MAO C S, LUO Y. Graph Convolutional Networks for Text Classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 7370-7377.

[23] DENG J W, ZHOU J Y, SUN H, et al. COLD: A Benchmark for Chinese Offensive Language Detection[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022: 11580-11599.

[24] VALERIO B, CRISTINA B, ELISABETTA F, et al. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter[C]// Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2019: 54-63.

[25] PATRICIA C, VÉRONIQUE M, FARAH B, et al. An Annotated Corpus for Sexism Detection in French Tweets[C]// Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, 2020: 1397-1403.

[26] OUSIDHOUM N, LIN Z Z, ZHANG H M, et al. Multilingual and Multi-Aspect Hate Speech Analysis[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2019: 4675-4684.

[27] OLLAGNIER A, CABRIO E, VILLATA S, et al. CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game[C]// Language Resources and Evaluation Conference. 2020.

[28] VANETIK N, MIMOUN E. Detection of Racist Language in French Tweets[J]. Information, 2022, 13(7): 318.

[29] KENNEDY, CHRIS J, GEOFF B, et al. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application[J]. arXiv: 2009.10277, 2020.

[30] MNASSRI K, FARAHBAKHSH R, CRESPI N. Multilingual Hate Speech Detection Using Semi-supervised Generative Adversarial Network[C]// International Conference on Complex Networks and Their Applications. 2024: 192-204.

ZHAO Hongyi, born in 2000, postgraduate. His main research interests include natural language psrocessing and hate speech detection.

BU Fanliang, born in 1965, Ph. D, professor, Ph. D supervisor. His main research interests include artificial intelligence and security engineering.