计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：结合社交网络图的多模态虚假信息检测模型
作者：叶舟波，罗舜，于娟
DOI：10.19734/j.issn.1001-3695.2023.11.0565
收稿日期：2023-11-23
网络首发日期：2024-02-02
引用格式：叶舟波，罗舜，于娟．结合社交网络图的多模态虚假信息检测模型[J/OL]．计算机应用研究. https://doi.org/10.19734/j.issn.1001-3695.2023.11.0565

网络首发时间：2024-02-02 17:27:36
网络首发地址：https://link.cnki.net/urlid/51.1196.TP.20240201.1441.006

结合社交网络图的多模态虚假信息检测模型 *

叶舟波，罗 舜，于 娟†

(福州大学 经济与管理学院，福州 350108)

摘 要：针对现有虚假信息检测方法主要基于单模态数据分析，检测时忽视了信息之间相关性的问题，提出了结合社交网络图的多模态虚假信息检测模型。该模型使用预训练 Transformer 模型和图像描述模型分别从多角度提取各模态数据的语义，并通过融合信息传播过程中的社交网络图，在文本和图像模态中加入传播信息的特征，最后使用跨模态注意力机制分配各模态信息权重以进行虚假信息检测。在推特和微博两个真实数据集上进行对比实验，所提模型的虚假信息检测准确率稳定为约 88%，高于 EANN、PTCA 等现有基线模型。实验结果表明所提模型能够有效融合多模态信息从而提高虚假信息检测的准确率。

关键词：网络舆情；虚假信息检测；多模态融合；跨模态注意力；社交网络图

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2023.11.0565

Multimodal misinformation detection model with social network graph

Ye Zhoubo, Luo Shun, Yu Juan†

(*College of Economics & Management*, *Fuzhou University*, *Fuzhou Fujian* 350108, *China*)

Abstract: To address the issues of existing misinformation detection approaches, which primarily focus on single-modal data analysis and ignore the correlation between information during detection, this paper proposed a multimodal misinformation detection model combined with the social network graph, the MMD-SNG model. This model used the pre-trained Transformer model and the image caption model to extract the semantics of each modality from multiple perspectives. It incorporated the features of propagated information into the text and image data by fusing the social network graph of the information dissemination process. Finally, this model used a multimodal co-attention mechanism to allocate the weights of each modality for misinformation detection. This paper conducted comparative experiments on two real datasets including Twitter and Weibo, the proposed MMD-SNG model achieved a consistent detection accuracy of approximately 88%, which was higher than existing misinformation detection approaches such as EANN and PTCA. The experimental results demonstrate that the proposed model can fuse multimodal information effectively to improve the accuracy of misinformation detection.

Key words: online public opinion; misinformation detection; multimodal fusion; multimodal co-attention; social network graph

0 引言

随着我国数字经济的发展与通信基础设施的日益完善，我国网民规模已达到 10.67 亿，互联网在城乡居民生活中得到了越来越广泛的普及与应用[1]。以微博、推特、微信朋友圈等为主的在线社交媒体平台逐步取代传统纸质与电视媒体，成为人们获取信息、分享观点、交换意见的主要场所。在线社交媒体平台正在促进文本、图片等多模态形式的信息在互联网上快速传播，而这些海量的多模态信息对我国网民和社会舆情有着广泛且深远的影响。虽然在线社交媒体平台拥有许多优秀内容，但由于其用户规模庞大且内容来源难以检测，其同样易于被利用来传播各种形式的虚假信息，如图 1 所示。不实信息借助在线社交媒体平台进行快速传播，往往会误导用户甚至引发严重的网络舆情管理问题[2]。因此，作为及时阻止虚假信息传播的基础环节，准确检测社交平台中的虚假信息就显得至关重要。

模态是指信息的来源或者信息表示形式。文本、图像、视频、声音和种类繁多的传感器信号都可以称为一种模态[3]。其中，文本和图像是当前在线社交媒体的主要内容形式。不同模态表示的信息往往存在交叉或互补，融合具有交叉或互补内容的多模态数据能够为管理决策提供更多且更准确的信息，提高决策准确率。因此如何处理利用所获取的多模态信息是提高决策总体准确率的关键[4]。

传统的虚假信息检测方法，主要从单一图像模态或单一文本模态入手进行检测，或是对两种模态进行简单的融合，一方面从文本方面对在线社交媒体平台内容制定文本规则或关键词屏蔽，另一方面从图片方面进行图片特征检测[5]。这类方法的主要问题有二：a)虚假信息检测过程中大多是对各条信息独立检测，忽视了信息之间的相关性；b)简单的模态融合易丢失图像模态与文本模态之间的交互信息，影响了检测

收稿日期：2023-11-23；**修回日期：**2024-01-22 **基金项目：**国家自然科学基金资助项目(71771054，72171090)；福建省自然科学基金资助项目(2023J01393)

作者简介：叶舟波(1999—)，男，福建南平人，硕士研究生，主要研究方向为数据挖掘与商务智能；罗舜(1997—)，男，福建龙岩人，博士研究生，主要研究方向为数据科学与智能系统；于娟(1981—)，女(通信作者)，山东青岛人，教授，博导，博士，主要研究方向为数据科学与智能信息系统(yujuan@fzu.edu.cn).

的准确率。

图 1　社交媒体平台多模态信息样例

Fig. 1　Sample multimodal information on social media platforms

多模态信息的融合处理主要包括模态表示、模态转换、模态对齐与融合等步骤。为了利用在线社交媒体平台中各信息之间丰富的关联来检测虚假信息，本文将社交网络图融入多模态转换、对齐、融合的过程，提出结合社交网络图的多模态虚假信息检测模型(Multimodal Misinformation Detection Model with the Social Network Graph, MMD-SNG 模型)。该模型将在线社交媒体平台上的用户与内容连接起来，在信息检测过程中共享各信息中的知识，协助彼此之间的检测；同时通过改进模型损失函数，在模态融合之前进行信息与社交网络图的对齐；最后使用跨模态注意力机制动态分配图像模态、文本模态、图像描述模态、社交网络特征的权重，将结果输入至多层全连接层,进行虚假/非虚假信息分类，生成检测结果。

本文主要贡献：

a) 本文将社交网络图结合进虚假信息检测模型中，通过社交网络图连接原本独立进行检测的信息，使各信息之间能够进行知识共享，从而提高虚假信息的检测准确率。

b) 本文将现有图像描述模型和跨模态注意力机制引入虚假信息检测，并改进模型损失函数，对多模态信息进行转换、对齐与融合，提高对信息中各模态相关性的利用率，减少信息冗余，提高检测效率。

1　相关研究

虚假信息(Misinformation)是指凭空捏造或被有意扭曲的消息，不能真实反映客观事物的本来面貌。近些年，国内外学者针对虚假信息的各类属性与传播特征，提出了很多虚假信息检测方法，包括制定文本规则、文本可信度评估等文本模态信息检测方法以及融合多模态信息的深度学习方法等。

1.1　单模态虚假信息检测

虚假信息检测方法的早期研究主要基于人工制定的特征规则，从文本内容、用户信息以及传播链的角度检测信息。例如，Castillo 等人[6]通过制定规则对 Twitter 传播的信息文本和传播用户进行主题发现与可信度评估来发现社交媒体中的虚假信息。这类方法起到了一定检测效果，但手工制定规则的成本较高且无法及时适应实时更新的信息检测。

近年来随着计算机算力的不断提升与深度学习技术的日益完善，机器学习和深度学习模型被应用于获取虚假信息检测的有效特征并进行分类[7]。MA 等人[8]基于递归神经网络(Recurrent Neural Network, RNN)学习微博文本的连续表示，通过学习微博帖子的上下文信息以及文本随时间变化的特征来识别虚假信息。Nasir 等人[9]提出了一种混合深度学习模型，结合卷积神经网络(Convolutional Neural Network, CNN)和递归神经网络来进行虚假信息分类。Sotto 等人[10]通过将所提取错误信息的关键特征与深度学习方法进行融合以检测虚假的健康信息。这类方法通过自动提取信息中各模态的特征以检测虚假信息，但因其仅提取单一模态信息而无法提取多模态信息完整语义，所以准确率不够令人满意。

1.2　多模态虚假信息检测

随着在线社交媒体平台信息从基于文本的帖子发展为带有图片或视频的多媒体推文，学者们开始研究包含文本模态和图像模态的多模态虚假信息检测方法。Wang 等人[11]提出基于生成对抗网络(Generative Adversarial Networks, GAN)的 EANN 模型，将多模态数据进行简单融合后输入 GAN 判别器中判定事件类别，利用学习到的每个事件的共享特征与独有特征强化虚假信息分类器。Khattar 等人[12]提出基于双模态变分自编码器和二元神经网络分类器的 MVAE 模型，通过优化观测数据的边际似然值的边界，来学习概率潜在变量模型。张国标等人[13]通过对图片与文本进行语义一致性计算，构建虚假新闻检测模型。Zhou 等人[14]基于多模态特征联合和跨模态相似性来学习社交媒体信息表示的 SAFE 模型，并通过计算相似度与损失函数相结合改进分类模型。

在融合图像模态与文本模态进行虚假信息检测的同时，为了提高检测准确率，学者们提出基于社交网络图的虚假信息检测方法。Yang 等人[15]提出基于图卷积网络与对抗训练的 CGAT 模型，通过训练鉴别器以区分在社交媒体中伪装身份的虚假信息发布者。Wei 等人[16]提出基于贝叶斯方法的考虑信息传播过程中潜在关系的 EBGCN 模型，该方法能够自适应地确定传播图中边的权重值以处理传播结构中的不确定性。Zheng 等人[17]在社交网络中考虑隐藏链接预测并且通过损失函数对齐来提升模态融合效果。韩雪明等人[18]在传播树结构上利用 Transformer 架构学习帖子的语义关系，提出了一种基于传播树的节点及路径双注意力虚假信息检测模型 DAN-tree，进一步对传播结构上的深层结构和语义信息进行融合。这类虚假信息检测方法多是对文本模态、图片模态与社交网络进行两两组合，尚未充分利用数据中的所有信息。

综上，当前虚假信息检测少有对多个模态信息进行融合、转换、对齐的虚假信息检测研究，尚未充分利用数据语义信息提高检测准确率。因此，在过去研究基础上，本文结合社交网络图特征，融合文本模态与图像模态以提高虚假信息检测的准确率。

2　多模态虚假信息检测模型 MMD-SNG

结合社交网络图的多模态虚假信息检测模型(MMD-SNG 模型)主要由社交网络图以及多模态表示、转换、对齐、融合模块组成，其结构如图 2 所示。

MMD-SNG 模型将输入待检测的多模态信息转换为集合 $Info=\{info_1,info_2,\cdots,info_n\}$。在其中，每条多模态信息 $info_i \in Info, info_i=\{t_i,p_i,u_i,c_i\}$，$t_i$ 表示该信息所包含的文本，p_i 表示该信息所包含的图片，u_i 表示在社交平台发布该信息的用户，c_i 表示该信息所获得的评论集合。一条多模态信息可以包含多条来自不同用户的评论，即 $c_i=\{c_i^1,c_i^2,\cdots,c_i^j\}$。

根据输入数据集建立初始社交网络图 $G=\{N,\boldsymbol{A}\}$，其中 N 表示社交网络图中的节点，本文将社交平台中的信息、用户、评论作为社交网络图的节点，N 的个数表示为 k；社交网络图中的邻接矩阵表示为 $\boldsymbol{A}\in\{0,1\}^{k\times k}$，当 $a_{ij}=1$ 时表示节点 $node_i$ 与节点 $node_j$ 之间存在一条无向边。

2.1　社会网络图特征

将社交平台中的信息、用户、评论作为社交网络图的节点。对于信息与评论节点的初始化，选用下文提取其中文本信息的句向量作为节点初始化特征，而对于用户节点，则选

用其所发布的所有信息与评论的句向量平均值作为节点初始化特征。为了缓解社会网络图中链接稀疏的问题，本文通过计算节点间的余弦相似度对社交网络图进行潜在链路预测，如式(1)所示。

$$\beta_{i,j}=\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|},\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathbb{R}^{1\times512} \tag{1}$$

其中，$\boldsymbol{x}_i,\boldsymbol{x}_j$ 为对应节点 $node_i$，$node_j$ 的初始化特征向量，$\beta_{i,j}$ 为两节点的余弦相似度。

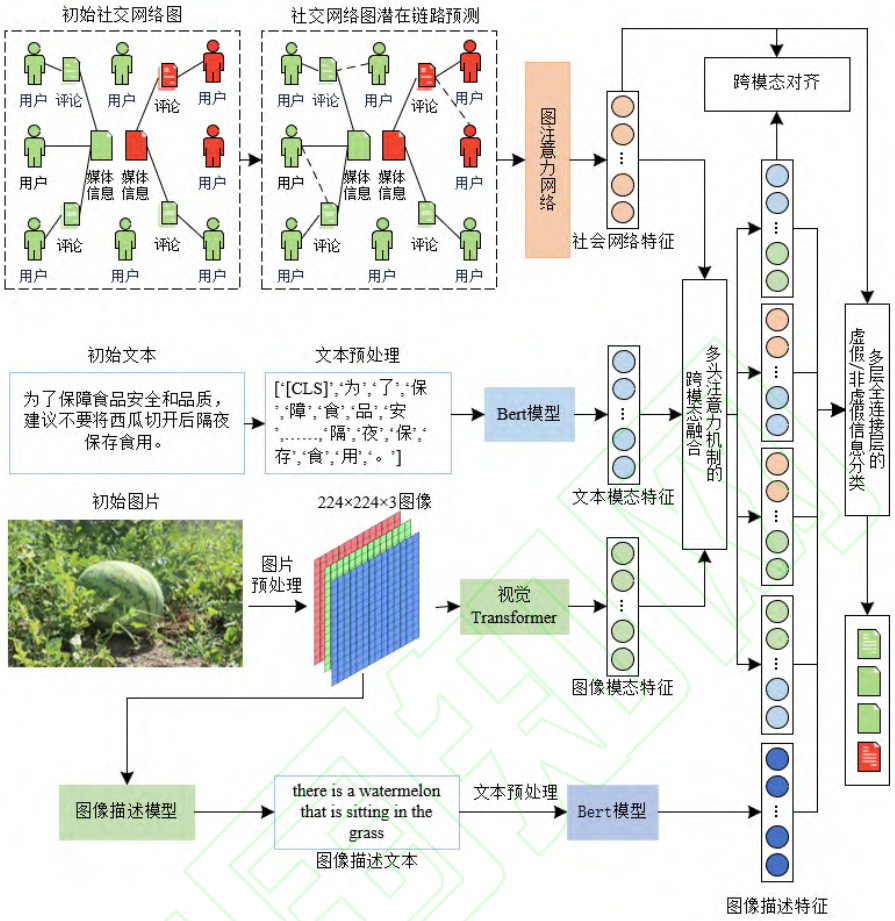

图 2　结合社交网络图的多模态虚假信息模型结构

Fig. 2　Structure of multimodal misinformation detection model with social network graph

本文认为余弦相似度大于一定阈值的节点间可能存在潜在链路，因此在预测后邻接矩阵 $\boldsymbol{A}'$ 中将余弦相似度大于所设阈值 θ 的两节点连接设置为 1，如式(2)所示。

$$a'_{i,j}=\begin{cases}1, if\ a_{i,j}=1\ or\ \beta_{i,j}>\theta\\0, otherwise\end{cases} \tag{2}$$

使用图注意力网络[19]提取社会网络图特征，能够对不同的邻接节点赋予不同的权重，更有利于关键信息的汇聚。对于节点 $node_i$，逐个计算其与邻接节点 $node_j$ 的注意力权重，如式(3)所示。

$$e_{i,j}=LeakyReLU\left(Concat\left(\boldsymbol{x}_i\boldsymbol{W},\boldsymbol{x}_j\boldsymbol{W}\right)\times\boldsymbol{a}\right) \tag{3}$$

其中，$e_{i,j}$ 表示节点 $node_i$ 与其邻接节点 $node_j$ 的注意力权重，$node_j\in\mathcal{N}_i$，$\mathcal{N}_i$ 表示与节点 $node_i$ 相邻接的节点的集合，$LeakyReLU$ 为激活函数，$\boldsymbol{W}\in\mathbb{R}^{512\times300}$ 与 $\boldsymbol{a}\in\mathbb{R}^{600\times1}$ 为神经网络中可学习的参数。

得到节点 $node_i$ 与其邻接节点的注意力权重 $e_{i,j}$ 后，使用 $softmax$ 进行注意力权重归一化。在社交媒体中，一条评论可以对一条信息提出同意观点或反对观点，由注意力机制会相应得到正注意力权重与负注意力权重。在 $softmax$ 函数中，会将负注意力权重值赋予相对较小的注意力分数，导致反对观点所获得的注意力较少[20]。因此，在计算注意力分数时同时计算取负注意力权重时的注意力分数。得到节点 $node_i$ 与其邻接节点的正注意力分数 $\alpha_{i,j}$ 与负注意力分数 $\alpha'_{i,j}$，如式(4)(5)所示。

$$\alpha_{i,j}=softmax(e_{i,j})=\frac{exp(e_{i,j})}{\sum_{k\in\mathcal{N}_i}exp(e_{i,k})} \tag{4}$$

$$\alpha'_{i,j}=softmax(-e_{i,j})=\frac{exp(-e_{i,j})}{\sum_{k\in\mathcal{N}_i}exp(-e_{i,k})} \tag{5}$$

将正负注意力分数与邻接节点特征进行加权求和后进行拼接，拼接后的向量经过一层全连接层即可得到经过图注意力网络汇聚后的社会网络图特征 $\boldsymbol{x}_i^{'}$，如式(6)所示。

$$\boldsymbol{x}_i^{'}=\sigma\left(Concat\left(\sum_{k\in\mathcal{N}_i}\alpha_{i,j}\times\boldsymbol{x}_k,\sum_{k\in\mathcal{N}_i}\alpha'_{i,j}\times\boldsymbol{x}_k\right)\boldsymbol{W}\right) \tag{6}$$

其中，$\sigma(\cdot)$ 表示激活函数，$concat$ 表示向量拼接，$\boldsymbol{W}\in\mathbb{R}^{1024\times300}$ 为神经网络中可学习参数，$\boldsymbol{x}'$ 经过全连接层后得到社会网络图特征 $\boldsymbol{f}_s$。

2.2　文本模态特征

利用预训练模型提取文本模态特征。此处使用 Bert 模型[21]，该模型利用自注意力机制对文本进行全局上下文理解，能够更好地捕捉文本的语义信息。对于多模态社交媒体信息数据集 $Info$ 中的文本 $T=\{t_1,t_2,\cdots,t_n\}$，首先进行去除停用词与分词等预处理并在文本前添加标识分类。对于每一条文本 t_i，利用基于无监督预训练的 Bert 模型对其进行向量表示，即 $B=\{\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_n\}$，如式(7)所示。

$$\boldsymbol{b}_i=Bert(t_i),\boldsymbol{b}_i\in\mathbb{R}^{L\times d_w} \tag{7}$$

其中，L 表示句子长度；d_w 表示词向量维度，本文设置为 512；每条 $\boldsymbol{b}_i$ 第一行的词向量 $\boldsymbol{b}_{i[CLS]}$ 对应该文本的分类标识，将分类

标识所对应的词向量作为该文本的句向量。句向量经全连接层降维后作为跨模态注意力中文本模态的输入 $\boldsymbol{f}_t$ 。

2.3 图像模态特征

对于图像模态的表征，CNN 使用卷积核来获取图像信息，但不擅长融合其他模态的信息，而 Transformer 的输入不需要保持二维图像，通常可以直接对像素进行操作得到初始嵌入向量，更适合多模态信息融合，因此本文使用 Transformer 提取多模态社交媒体信息中的图像模态特征。此处使用 CLIP 模型[22]中预训练的 ViT[23]模型。

对于多模态社交媒体信息数据集中的图像 $P=\{p_1,p_2,\cdots,p_n\}$，首先进行图像变换与图像切块等预处理，并添加标识分类以及每个图像块的位置编码。对于每一张图片 p_i，利用基于对比学习预训练的 ViT 模型抽取其特征，即 $V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$，如式(8)所示。

$$\boldsymbol{v}_i=ViT(p_i),\boldsymbol{v}_i\in\mathbb{R}^{197\times768} \tag{8}$$

其中每条 $\boldsymbol{v}_i$ 第一行的向量 $\boldsymbol{v}_{i[CLS]}$ 对应该图片的分类标识，本文将分类标识所对应的向量作为该图片的全局特征向量，经全连接层降维后作为跨模态注意力中图像模态的输入 $\boldsymbol{f}_p$ 。

同时，为了多角度获取图像语义，进一步通过 BLIP 模型[24]为预处理后的图像生成对应的图像描述文本，再同样使用本文 3.2 节所述的 Bert 模型得到图像描述模态特征向量 $IC=\{\boldsymbol{ic}_1,\boldsymbol{ic}_2,\cdots,\boldsymbol{ic}_n\},\boldsymbol{ic}_i\in\mathbb{R}^{1\times512}$，$\boldsymbol{ic}_i$ 经全连接层降维后得到图像描述特征向量 $\boldsymbol{f}_{ic}$ 。

2.4 多头注意力机制的跨模态融合

在已有多模态虚假信息检测研究中，多采用直接相加融合或点乘融合，在模态融合时未进行权重动态分配，无法对各模态之间相关联语义进行针对性融合。为此，本文引入多头跨模态注意力机制，以使模型更好地捕捉不同模态之间的关联性，在更好地学习不同模态之间共性的同时减少信息冗余。多头注意力机制的跨模态融合计算方法如式(9)~(11)所示。

$$Attemtion(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{9}$$

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_h)\boldsymbol{W}^O \tag{10}$$

$$\boldsymbol{head}_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{11}$$

其中，h 表示注意力头数，d_{model} 表示模型输入向量的维度，d_k 表示各注意力头输入向量的维度且 $d_k=d_{model}/h$，神经网络可学习参数 $\boldsymbol{W}_i^Q\in\mathbb{R}^{d_{model}\times d_k},\boldsymbol{W}_i^K\in\mathbb{R}^{d_{model}\times d_k},\boldsymbol{W}_i^Q\in\mathbb{R}^{d_{model}\times d_k},\boldsymbol{W}_i^K\in\mathbb{R}^{d_{model}\times d_k}$ 。

如图 2 所示，分别将所获取的文本模态特征 $\boldsymbol{f}_t$ 、图像模态特征 $\boldsymbol{f}_p$ 、社交网络图特征 $\boldsymbol{f}_s$ 作为注意力机制中的 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$，从而获得文本-图像融合特征 $\boldsymbol{f}_{tp}$ 、图像-文本融合特征 $\boldsymbol{f}_{pt}$ 、社交网络图-文本融合特征 $\boldsymbol{f}_{st}$ 、社交网络图-图像融合特征 $\boldsymbol{f}_{sp}$ 。

通常从社交媒体上获取的图像-文本对是弱相关的，即文本包含与图像无关的文字或图像包含文本中没有描述的实体[25]。因此跨模态注意力机制对于图像模态和文本模态映射仍停留在其各自的空间，难以学习到跨模态信息之间的交互。为了改善在噪声数据中的学习问题，引入均方差损失 $\mathcal{L}_{align}$，以对齐社交网络图、图像模态、文本模态三者之间的信息，如式(12)所示。

$$\mathcal{L}_{align}=\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{f}_s\boldsymbol{W}_s-\boldsymbol{f}_{tp}\boldsymbol{W}_{tp}\right)^2 \tag{12}$$

其中，$\boldsymbol{f}_s$ 为输入模型的社交网络图特征，$\boldsymbol{f}_{tp}$ 为上小节计算出的文本-图像融合特征，n 表示样本数量。神经网络可学习参数 $\boldsymbol{W}_s\in\mathbb{R}^{300\times300}$，$\boldsymbol{W}_{tp}\in\mathbb{R}^{300\times300}$ 。

2.5 虚假信息检测分类

通过前述的多模态特征提取和跨模态对齐，获得输入多层全连接层的所有输入，且所有特征均通过全连接层降至同一维度。将这些特征向量拼接后输入多层全连接层，得到最终虚假信息检测分类结果，如式(13)(14)所示。

$$\boldsymbol{x}_{input}=Concat(\boldsymbol{f}_s,\boldsymbol{f}_{tp},\boldsymbol{f}_{pt},\boldsymbol{f}_{sp},\boldsymbol{f}_{st},\boldsymbol{f}_{ic}) \tag{13}$$

$$\hat{y}=softmax(\boldsymbol{x}_{input}\boldsymbol{W}+\boldsymbol{b}) \tag{14}$$

其中，神经网络可学习参数 $\boldsymbol{W}\in\mathbb{R}^{1800\times2}$，$\boldsymbol{b}\in\mathbb{R}^{1\times2}$

虚假信息检测模型的分类损失函数为交叉熵函数 $\mathcal{L}_{detection}$，将其与上述对齐损失函数相加即为模型最终损失函数 $\mathcal{L}_{model}$，如式(15)(16)所示。

$$\mathcal{L}_{detection}=\frac{1}{n}\sum_{i=1}^{n}-\left(y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\right)^2 \tag{15}$$

$$\mathcal{L}_{model}=\lambda_1\mathcal{L}_{detection}+\lambda_2\mathcal{L}_{align} \tag{16}$$

其中，$y_i\in\{0,1\}$ 为数据集标签，0 表示为虚假信息，1 表示为非虚假信息，$\hat{y}$ 为模型预测标签，n 表示数据集数据量，λ_1 与 λ_2 为可调节超参数。

3 实验及结果分析

3.1 实验数据

采用微博[26]与推特[27]两个真实社交媒体虚假信息检测数据集进行实验，分析本文方法的性能。微博与推特均为目前主流的社交媒体平台，其数据具有代表性。两个数据集均包含文本、图像、评论以及标注。其中，微博数据集中的文本内容为中文，推特数据集中的文本内容为英文。

为了对比分析多模态虚假信息检测方法，在数据集的预处理阶段，本文专注于检测具有多模态特征的虚假信息，因此删除了数据集中仅含文本或图片的数据以及重复数据，同时为了保证数据质量删除了经 Word Piece 分词后 token 序列长度小于 5 的数据或图片单边分辨率低于 112 的数据。设置句子 token 序列长度为 80，对于超长的句子，截去超出部分；对于长度不足 80 的句子则用标识符补齐。同时，将图片大小调整为 224×224×3，以方便模型输入。经过预处理后的数据集如表 1 所示。

表 1　微博与推特数据集统计数据

Tab. 1　Weibo and Twitter dataset data

数据集来源	虚假信息数	非虚假信息数	图片数	用户数	评论数	社交网络图节点数
微博	590	877	1467	985	4534	6963
推特	590	1428	2018	894	7388	4329

3.2 实验设置及评价指标

本文实验均基于单张 NVIDIA RTX3090 24GB GPU 训练，使用的 CUDA 版本为 11.3。本文通过基于 Python 的深度学习框架 PyTorch 实现 MMD-SNG 模型，所使用的预训练模型

权重来自 HuggingFace，将预处理完的数据输入进模型进行训练。

将数据集按 7:1:2 的比例随机划分训练集、验证集与测试集，所有实验均在相同的训练集、测试集上完成。在训练过程中，本文使用 AdamW 作为目标函数的优化器，学习率设置为 0.001，权重衰减系数设置为 0.01，全连接层激活函数为 GELU，训练批次大小设置为 96，训练轮数设置为 80。本文在训练中采用早停策略，如损失函数在 10 轮训练中未下降则停止训练。

由于本文选用数据集中的两类数据数量不同，因此本文选用准确率(Accuracy)，加权平均精确率(Weighed-Precision)，加权平均召回率(Weighed-Recall)，加权平均 F1 分数(Weighed-F1-score)作为评价指标，上述评价指标在后续表中分别简称为 A、P、R、F1。

3.3 基线模型

本文选择了几种在公开数据集中表现出优异性能的虚假信息检测模型，将其与本文所提出的模型作对比，具体如下。

EANN(Event Adversarial Neural Network)[11]：该模型采用 Text-CNN 模型提取文本模态特征，采用 VGG-19 模型提取图像模态特征，两者通过连接融合后进行虚假信息检测。该模型提出采用生成对抗网络学习各事件的独立性，有利于模型学习各事件的独立特征。

GLAN(Global-Local Attention Network)[28]：该模型基于社交网络图，将相关转发的语义信息与注意力机制进行融合，为每条信息生成包含社交语义的局部特征表示，同时构建一个全局异质图捕获丰富的结构信息以用于虚假信息检测。

SAFE(Similarity-Aware Fake news detection)[14]：该模型通过引入了一个额外的全连接层的扩展 Text-CNN 来提取文本模态特征与图像特征，并在损失函数中引入文本模态与图像模态的微调余弦相似性以进行虚假信息检测。

EBGCN(Edge-enhanced Bayesian Graph Convolutional Network)[16]：该模型采用贝叶斯方法，自适应地重新考虑了社交网络图中的潜在关系，结合了文本嵌入层与图卷积神经网络共同提取社交网络图特征，然后通过节点更新模块与边推理模块捕获图结构特征，将两者拼接以进行虚假信息检测。

PTCA(Pre-trained Transformer and Cross Attention)[29]：该模型采用预训练的 Transformer 分别提取文本和图像特征，同时使用 Text-CNN 多角度提取语义特征后通过交叉注意力机制获取两个模态的融合特征后进行虚假信息检测。

其中 EANN、SAFE、PTCA 同时提取了文本模态与图像模态特征进行虚假信息检测，GLAN、EBGCN 在提取社交媒体信息文本模态特征的基础上融合了社交网络图进行虚假信息检测。

3.4 对比实验

本文提出的结合社交网络图的多模态虚假信息检测模型 MMD-SNG 与其他模型在推特数据集上的实验对比结果如表 2 所示，在微博数据集上的实验对比结果如表 3 所示。

由表 2 和 3 所示，本文提出的 MMD-SNG 模型在两个真实虚假信息检测数据集中评价指标上优于当前先进的基线模型，相关结果分析如下。

a) 在与 EANN 模型的对比中，MMD-SNG 在两个数据集的检测准确率上平均高出了 7.1%。EANN 在融合阶段仅采用了高维的拼接融合，表明 MMD-SNG 中采用跨模态注意力融合机制能够有效减少模态融合过程中各个模态所包含的噪声信息，动态计算的注意力分数能够将权重赋予各模态中的有效信息，从而提高模态融合效率。

b) GLAN 将原信息、转发信息与用户信息之间的结构关系建模为社交网络异质图，但在建模过程中只考虑社交媒体中的文本模态信息，忽略了图像模态信息，在与 GLAN 模型的对比中，表明图像模态的加入有助于丰富社交媒体信息的特征表示，多模态融合后的特征具有更加丰富的语义信息，能够增强虚假信息检测效果。

c) SAFE 在文本与图像模态融合阶段，通过计算相似度的方式将两种模态间的特征进行融合，其实验结果表明在模态融合阶段考虑各模态间相似度以对齐模态有助于提高模态融合性能。

d) 与 GLAN 相似，EBGCN 采用贝叶斯方法，在社交网络图中考虑了节点间的潜在关系可靠性，但同样仅抽取文本模态特征，表明了在虚假信息检测中考虑社交网络图的潜在链路预测能够提升检测性能。

e) PTCA 使用了两个经预训练的 Transformer 模型(BERT 和 ViT)分别提取文本和图像特征，一定程度上克服了样本数量不足的局限性，该模型还采用 Text-CNN 模型多角度提取文本语义信息，最后通过交叉注意力机制进行特征融合。同时，由于基于 Transformer 模型中的自注意力机制带来了二次方计算复杂度，与本文相同，模型在表现出优异性能的同时，其计算时间与算力消耗方面与上述模型相比较大。

表 2　模型在推特数据集中的对比实验结果

Tab. 2　Experimental results of the model in the Twitter dataset

模型	A(%)	P(%)	R(%)	F1(%)
EANN	78.09	72.46	72.26	72.13
GLAN	84.96	83.31	80.39	81.19
SAFE	82.33	81.10	80.31	80.38
EBGCN	83.64	82.04	79.99	80.46
PTCA	86.75	87.04	86.42	86.87
MMD-SNG	**88.57**	**88.39**	**88.57**	**88.34**

注：粗体表示在该指标下取得的最优结果。

表 3　模型在微博数据集中的对比实验结果

Tab. 3　Experimental results of the model in the Weibo dataset

模型	A(%)	P(%)	R(%)	F1(%)
EANN	83.22	82.56	82.14	82.27
GLAN	84.46	84.71	82.57	83.19
SAFE	85.80	85.80	85.86	85.82
EBGCN	85.15	87.58	83.30	83.16
PTCA	86.44	86.58	86.44	86.50
MMD-SNG	**88.14**	**88.16**	**88.14**	**88.15**

注：粗体表示在该指标下取得的最优结果。

3.5 交叉验证实验

为避免偶然性，更全面评估模型表现，本节将两个数据集中虚假信息与非虚假信息两类数据平均分成五份进行分层五折交叉验证。交叉验证实验其余设置与对比实验相同，五次实验结果取平均值，结果如表 4 所示。

表 4　MMD-SNG 交叉验证实验结果

Tab. 4　Cross validation experimental results of the MMD-SNG

数据集来源	A(%)	P(%)	R(%)	F1(%)
推特	88.31	88.43	88.31	88.27
微博	88.24	88.28	88.25	88.26

由表 4 可以看出，MMD-SNG 模型在两个真实数据集上交叉验证实验表现与对比实验中结果相近，证明了该模型的可靠性。

3.6　消融实验

为验证本文所提出模型中各部分模块功能，通过对本文所提出模型去除对应模块进行消融实验。在两个数据集上的消融实验结果如表 5 和 6 所示，MMD-SNG 为本文提出模型，MMD 表示在本文模型的基础上去除社交网络特征与模态对齐，MMD-V 表示在本文模型的基础上去除视觉模态特征、图像描述特征与模态对齐，MMD-T 表示在本文模型的基础上去除文本模态特征与模态对齐，MMD-IC 表示在本文模型的基础上去除图像描述特征。

表 5　模型在推特数据集中的对比实验结果

Tab. 5　Results of ablation experiments in the Twitter dataset

模型	A(%)	P(%)	R(%)	F1(%)
MMD-SNG	**88.57**	**88.39**	**88.57**	**88.34**
MMD	85.71	85.46	85.71	85.53
MMD-V	84.94	84.60	84.94	84.36
MMD-T	75.84	74.40	75.84	73.07
MMD-IC	86.75	87.48	86.75	86.98

注：粗体表示在该指标下取得的最优结果。

表 6　模型在微博数据集中的对比实验结果

Tab. 6　Results of ablation experiments in the Weibo dataset

模型	A(%)	P(%)	R(%)	F1(%)
MMD-SNG	**88.14**	**88.16**	**88.14**	**88.15**
MMD	86.44	87.17	86.44	86.61
MMD-V	69.15	67.70	69.15	65.96
MMD-T	83.73	84.24	83.73	83.88
MMD-IC	84.75	84.95	84.75	84.82

注：粗体表示在该指标下取得的最优结果。

由表 5、6 数据，对消融实验结果分析如下。

a) 与去除了社交网络图的 MMD 相比，MMD-SNG 中社交网络图特征的加入将各条独立的信息通过图结构连接起来，使得在模型检测时能够同时考虑多条信息语义，能够提高多模态虚假信息检测模型性能。

b) MMD-IC 的性能下降表明，图像描述模态的加入对原有信息作了数据增强，能够利用起信息中的全部语义，以提升模型各项性能。

c) MMD-V 与 MMD-T 的性能结果表明，多模态虚假信息检测相比于单一模态检测，能够从多角度提取信息特征，使检测性能得到显著提高；且在不同数据集中，文本模态与图片模态所表现出的信息量不同，同时利用多种模态数据能够有效提升虚假信息检测模型的鲁棒性。

3.7　参数敏感性实验

在本文 2.1 节潜在链路预测式 2 与 3.5 节的模型损失函数式 16 中，潜在链路预测阈值 θ 与分类损失函数与对齐损失函数前的系数 λ_1 和 λ_2 是本文所提出的模型中的重要超参数，因此，本节对不同参数的取值导致模型性能的影响进行评估。

参数敏感性实验结果分别如图 3 和 4 所示。从图 3 中可以看出，当其他参数固定时，改变 θ 的值能对模型在不同数据集上的性能均能产生较大影响，当节点相似度阈值设为 0.5 时进行潜在链路预测能够使模型性能取得最优。

从图 4 中可以看出，当固定 θ 为 0.5 时，在一定范围内，模型性能总体趋势为随着 λ_1 和 λ_2 的增加而增加，说明模态对齐能够使得模态间数据具有一致性，对齐损失函数与分类损失函数的结合能够有效提升模型性能。经多次实验结果表明，当 λ_1 的取值为 1.8，λ_2 的取值为 2.4 时，模型的性能取得最优。

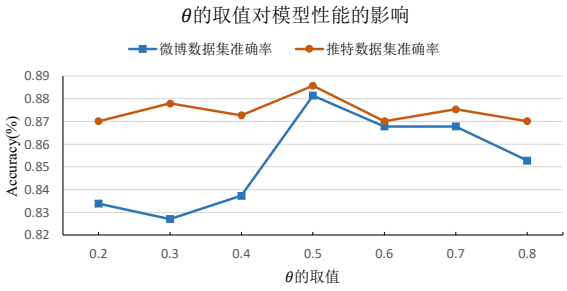

图 3　θ 的取值对模型性能的影响

Fig. 3　The Effect of Values of θ on Model Performance

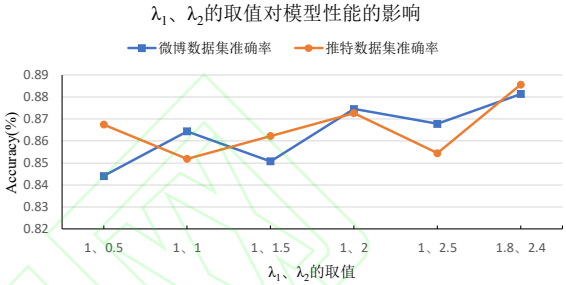

图 4　λ_1 和 λ_2 的取值对模型性能的影响

Fig. 4　The Effect of Values of λ_1 and λ_2 on Model Performance

3.8　样例分析

为了更加直观证明本文所提出的模型有效性，将本文提出模型的结果与现有基线模型进行比较，结果如图 5 所示。图 5 为在两个数据集中被本文所提模型 MMD-SNG 正确检测的社交媒体信息，其中绿框中的为真实信息，红框中的为虚假信息。从图 5 中可以看出多模态虚假信息主要涉及图像窜改、夸大文本内容、图像与文本描述不符等情况。图 5(a)与图 5(b)为常识类与谚语类文本信息，比较模型(如 EANN、GLAN)对其进行了错误分类，本文提出模型由于采用经过预训练的语言模型提取文本特征，能够正确检测该类信息。图 5(c)为图片信息中含有地图信息，比较模型(如 EANN、PTCA)对其进行了错误分类，本文提出模型由于具有图像描述模块，对于识别图片中的文字以及地图类图片时，能够准确识别到图片中的文字信息并提高检测正确率。

MMD-SNG检测成功，但比较方法检测失败样例

图 5　MMD-SNG 检测样例

Fig. 5　Detection Examples of MMD-SNG

在图 5(d)中所示图片与现实具有差距，可能是一张伪造的图片、图 5(e)中文字与图片相比具有夸大成分、而图 5(f)

中则是图片内容与文本不符。以上三个例子均具有上述虚假信息主要特征，本文所提模型均正确将其检测为虚假信息，而比较模型(如 PTCA、EBGCN 等)将其分类为非虚假信息。上述结果表明本文所提出的模型能够有效地提取多模态社交媒体信息中的文本信息和图像信息，与社交网络图的结合以及利用图像描述进行数据增强能够更全面地检测社交媒体信息的真实性，证明其在多模态虚假信息检测任务中的有效性。

4　结束语

针对当前在线社交媒体平台虚假信息检测方法中存在的忽略信息间相关性和多模态融合不充分问题，本文提出结合社交网络图的多模态虚假信息检测模型(Multimodal Misinformation Detection Model with the Social Network Graph, MMD-SNG 模型)。该模型采用社交网络图连接通常独立进行检测的信息，将图信息以节点特征的形式融合进虚假信息检测模型；同时，提出基于图像描述模型的多模态数据增强方法，将经过模态转换后的图像描述文本融合进最终的决策模型中；最后，经过模态对齐与跨模态注意力机制，将社交网络特征、文本模态特征、图像模态特征、图像描述特征进行融合以作虚假信息检测。在推特和微博上的实验结果表明，相比已有模型取得了更好的准确率。

本研究的不足之处在于，构建社交网络图时仅考虑了文本模态信息，无法全面捕获多媒体信息中的关联性；并且，本文所使用的图像描述模型仅可产生英文文本描述，在中文数据集上会与现有的中文文本模态产生特征间隔，从而造成信息冗余。未来研究可在 MMD-SNG 模型的基础上，完善社交网络图的构建以及探索使用中文图像描述模型，进一步优化多模态虚假信息检测模型的计算效果。

参考文献：

[1] 中国互联网网络信息中心. 第51次中国互联网络发展状况统计报告[R/OL]. (2023-03-02) [2023-05-31]. https://www. cnnic. net. cn/n4/2023/0303/c88-10757. html. (China Internet Network Information Center. The 51st statistical report on the development of China's Internet [R/OL]. (2023-03-02) [2023-05-31]. https://www. cnnic. net. cn/n4/2023/0303/c88-10757. html.)

[2] 王晰巍, 邱程程, 李玥琪. 突发公共事件下社交网络谣言辟谣效果评价及实证研究 [J]. 情报理论与实践, 2022, 45 (12): 14-21. (Wang Xiwei, Qiu Chengcheng, Li Yueqi. Evaluation and empirical research on the effect of social network rumors refutation under public emergencies [J]. Information Studies: Theory & Application, 2022, 45 (12): 14-21.)

[3] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35 (8): 1798-1828.

[4] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2018, 41 (2): 423-443.

[5] Guo Haoming, Huang Tianyi, Huang Huixuan, *et al.* A systematic review of multimodal approaches to online misinformation detection [C]// Proc of the 5th IEEE International Conference on Multimedia Information Processing and Retrieval. Piscataway, NJ: IEEE Press, 2022: 312-317.

[6] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter [C]// Proc of the 20th International Conference on World Wide Web. New York: ACM Press, 2011: 675-684.

[7] Mishra S, Shukla P, Agarwal R. Analyzing machine learning enabled fake news detection techniques for diversified datasets [J]. Wireless Communications and Mobile Computing, 2022, 2022 (1): Article ID 1575365.

[8] Ma Jing, Gao Wei, Mitra P, *et al.* Detecting rumors from microblogs with recurrent neural networks [C]// Proc of the 25th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 3818-3824.

[9] Nasir J A, Khan O S, Varlamis I. Fake news detection: a hybrid CNN-RNN based deep learning approach [J]. International Journal of Information Management Data Insights, 2021, 1 (1): 100007.

[10] Di Sotto S, Viviani M. Health misinformation detection in the social web: an overview and a data science approach [J]. International Journal of Environmental Research and Public Health, 2022, 19 (4): 2173.

[11] Wang Yaqing, Ma Fenglong, Jin Zhiwei, *et al.* Eann: event adversarial neural networks for multi-modal fake news detection [C]// Proc of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2018: 849-857.

[12] Khattar D, Goud J S, Gupta M, *et al.* Mvae: multimodal variational autoencoder for fake news detection [C]// Proc of World Wide Web Conference. New York: ACM Press, 2019: 2915-2921.

[13] 张国标, 李洁. 融合多模态内容语义一致性的社交媒体虚假新闻检测 [J]. 数据分析与知识发现, 2021, 5 (5): 21-29. (Zhang Guobiao, Li Jie. Detecting social media fake news with semantic consistency between multi-model contents [J]. Data Analysis and Knowledge Discovery, 2021, 5 (5): 21-29.)

[14] Zhou Xinyi, Wu Jindi, Zafarani R. Safe: similarity-aware multi-modal fake news detection [C]// Proc of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham, Switzerland: Springer, 2020: 354–367.

[15] Yang Xiaoyu, Lyu Yuefei, Tian Tian, *et al.* Rumor detection on social media with graph structured adversarial learning [C]// Proc of the 29th International Conference on International Joint Conferences on Artificial Intelligence. Freiburg: IJCAI Press, 2021: 1417-1423.

[16] Wei Lingwei, Hu Dou, Zhou Wei, *et al.* Towards propagation uncertainty: edge-enhanced bayesian graph convolutional networks for rumor detection [C]// Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA: ACL Press, 2021: 3845-3854.

[17] Zheng Jiaqi, Zhang Xi, Guo Sanchuan, *et al.* Mfan: multi-modal feature enhanced attention networks for rumor detection [C]// Proc of the 31st International Joint Conference on Artificial Intelligence. Freiburg: IJCAI Press, 2022: 2413-2419.

[18] 韩雪明, 贾彩燕, 李轩涯等. 传播树结构节点及路径双注意力谣言检测模型 [J]. 计算机科学, 2023, 50 (4): 22-31. (Han Xueming, Jia Caiyan, Li Xuanya, *et al.* Dual-attention network model on propagation tree structures for rumor detection [J]. Computer Science, 2023, 50 (4): 22-31.)

[19] Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks [EB/OL]. (2018-2-4) [2023-05-31]. https://arxiv. org/pdf/1710. 10903. pdf.

[20] Tian Tian, Liu Yudong, Yang Xiaoyu, *et al.* Qsan: a quantum-probability based signed attention network for explainable false information detection [C]// Proc of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2020:

1445-1454.

[21] Devlin J, Chang Mingwei, Lee K, *et al.* Bert: pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-5-24) [2023-05-31]. https://arxiv. org/pdf/1810. 04805. pdf.

[22] Radford A, Kim J W, Hallacy C, *et al.* Learning transferable visual models from natural language supervision [C]// Proc of International Conference on Machine Learning. Cambridge, MA: JMLR Press, 2021: 8748-8763.

[23] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. (2021-6-3) [2023-05-31]. https://arxiv. org/pdf/2010. 11929. pdf.

[24] Li Junnan, Li Dongxu, Xiong Caiming, *et al.* Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation [C]// Proc of the 39th International Conference on Machine Learning. New York: PMLR Press, 2022: 12888-12900.

[25] Li Junnan, Selvaraju R, Gotmare A, *et al.* Align before fuse: vision and language representation learning with momentum distillation [J]. Advances in Neural Information Processing Systems, 2021 (34): 9694-9705.

[26] Song Changhe, Yang Cheng, Chen Huimin, *et al.* Ced: credible early detection of social media rumors [J]. IEEE Trans on Knowledge and Data Engineering, 2019, 33 (8): 3035-3047.

[27] Zubiaga A, Liakata M, Procter R. Exploiting context for rumour detection in social media [C]// Proc of Social Informatics: International Conference, Cham, Switzerland: Springer, 2017: 109-123.

[28] Yuan Chunyuan, Ma Qianwen, Zhou Wei, *et al.* Jointly embedding the local and global relations of heterogeneous graph for rumor detection [C]// Proc of IEEE International Conference on Data Mining. Piscataway, NJ: IEEE Press, 2019: 796-805.

[29] 蒋保洋, 但志平, 董方敏等. 基于双预训练 Transformer 和交叉注意力的多模态谣言检测 [J]. 国外电子测量技术, 2023, 42 (4): 149-157. (Jiang Baoyang, Dan Zhiping, Dong Fangmin, *et al.* Multimodal rumor detection method based on dual pre-trained Transformer and cross attention mechanism [J]. Foreign Electronic Measurement Technology, 2023, 42 (4): 149-157.)