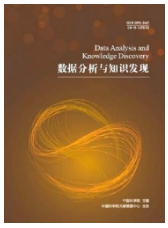

数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：基于循环生成对抗网络和 Wasserstein 损失的谣言检测研究
作者：张洪志，但志平，董方敏，高准，张岩珂
网络首发日期：2023-10-19
引用格式：张洪志，但志平，董方敏，高准，张岩珂. 基于循环生成对抗网络和 Wasserstein 损失的谣言检测研究[J/OL]. 数据分析与知识发现.
https://link.cnki.net/urlid/10.1478.G2.20231018.1732.010

网络首发时间：2023-10-19 15:45:06
网络首发地址：https://link.cnki.net/urlid/10.1478.G2.20231018.1732.010

基于循环生成对抗网络和 Wasserstein 损失的谣言检测研究

张洪志[1,2]，但志平[1,2]，董方敏[1,2]，高准[1,2]，张岩珂[1,2]
[1](三峡大学水电工程智能视觉监测湖北省重点实验室，湖北 宜昌 443000)
[2](三峡大学计算机与信息学院，湖北 宜昌 443000)

摘要：
[目的]通过循环生成对抗网络和 Wasserstein 距离改进的生成损失，利用对抗训练提高谣言检测模型在数据样本不平衡、非配对情况下的稳定性和精确度。**[方法]**利用生成器和判别器之间的对抗训练实现谣言判别模型的增强。在生成训练过程中引入循环一致性损失和识别损失以实现生成目标的可控性，并使用 Wasserstein 距离改进了模型生成损失，提高生成器的引导效果的同时避免对抗网络训练过程中可能出现的梯度爆炸的问题。**[结果]**在不平衡谣言数据集 PHEME 上，本文提出的方法准确率达到 0.8698，F1 值达到 0.855，与最新方法相比，分别提高了 0.0068 和 0.018。**[局限]**基于循环生成对抗网络的谣言检测模型只有两个生成器，因此只能实现两种类别样本的转换，只适用于二分类的谣言检测模型，对于多分类谣言检测任务则无法应用。**[结论]**使用 Wasserstein 距离改进生成损失的循环生成对抗网络可以有效提升谣言检测模型在数据不平衡情况下的谣言检测能力。

关键词：谣言检测；循环生成对抗网络；Wasserstein 损失
分类号：TP393，G250

Detect Rumors by CycleGAN with Wasserstein Distance

Zhang Hongzhi[1,2], Dan Zhiping[1,2], Dong Fangmin[1,2], Gao Zhun[1,2], Zhang Yanke[1,2]
[1] (Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443000, China)
[2](College of Computer and Information Technology, China Three Gorges University, Yichang 443000, China)

Abstract:
[Objective] By CycleGAN and improved generation loss through Wasserstein distance improving the stability and accuracy of the rumor detection model in cases of imbalanced and unpaired data samples. **[Methods]** The rumor discriminative model is enhanced by adversarial training between generator and discriminator. We use Wasserstein distance upgrade the cycle consistency loss and identity loss, and introduce both loss in generation training process to achieve controllability of the generated target, and improving the guidance effect of the generator while avoiding the problem of gradient explosion that may occur during adversarial network training. **[Results]**Our method achieves an accuracy of 0.8698 and an F1 value of 0.855 on the unbalanced rumor dataset PHEME, and compared with the latest method, it has increased by 0.0068 and 0.018 respectively. **[Limitations]** The rumor detection model based on CycleGAN only has two generators and can only achieve the conversion of two categories of samples, so it just suitable for binary classification rumor detection models, and cannot be applied to the multi classification rumor detection tasks. **[Conclusions]** Our proposed model can effectively enhance the rumor detection ability of the rumor detection model in cases of imbalanced data.

Keywords: Rumor detection; CycleGAN; Wasserstein loss

1 引言

网络谣言通常被定义为在互联网上制造并传播的看似真实的虚假故事[1]，这些虚假信息在社交网络上传播并与真实信息进行竞争。谣言制造者可以借助网络谣言吸引大量公众注意力、操纵公共舆论以实现其目的，给社会带来了极大的不稳定性。目前有效的谣言检测方法多使用事实核查方法，通过人工方式对社交网络上的信息进行鉴别，其时效性和鉴别范围都受到了极大的限制。而人工智能检测方法往往难以适用于现实的网络环境，重要原因之一是因为现实环境中的谣言与非谣言数量往往是不平衡的，社交网络上的谣言往往隐藏在大量的非谣言内容之中，这些谣言信息不断在社交网络上传播并伪装成真实信息，给谣言检测带来了极大的挑战。

社交网络上的信息数量庞大、内容复杂且包含大量不同用户之间的交互内容。为了从复杂的社交网络中准确地检测出谣言信息，国内外诸多学者利用深度神经网络来提取谣言特征，尝试构建能够理解谣言文本语义信息以及谣言传播结构特征的深度学习模型。如使用 RNN、CNN 以及注意力机制、Transformer 和预训练模型等方法，提取网络谣言中的语义特征和传播特征。但谣言在社交网络传播的过程中，信息序列会被不断加工，增加大量的噪声信息，具有很强的迷惑性[2]。这些噪声信息往往对谣言的判别没有贡献，反而会影响谣言检测的准确性。

为了提高谣言检测模型对谣言信息在社交网络传播的过程中序列特征变化情况的学习能力，从而提高训练模型对谣言序列核心特征的提取能力，部分学者[2][3][4]尝试利用生成对抗网络（Generative Adversarial Network, GAN）构建谣言检测模型，通过在贴子中插入争议内容、将谣言非谣言贴子相互转换等方式，利用对抗训练增强了判别器捕获谣言序列核心特征的能力。但是谣言样本与非谣言样本之间往往并不存在相互对应关系，因此利用生成器进行训练时，生成对象是否具有相应的谣言/非谣言特征难以有效评估；同时在样本数据不平衡的情况下，样本量少的类别的生成数据的多样性更低，导致判别器难以学习到这类数据更深层次的特征；而且随着判别器的训练效果增强，生成器将会面临梯度消失的问题，生成样本将逐渐失去多样性[5]，单一的生成效果反而会降低模型的判别能力。

为了解决上述问题，本文在原有谣言 GAN 网络模型的基础上，充分结合谣言信息在社交媒体传播过程中的特点，引入循环对抗网络（Cycle consistent Generative Adversarial Network，CycleGAN）中的循环一致性损失和识别损失，并使用 Wasserstein 距离[5]改进了生成损失，构建了基于 Wasserstein 生成损失的循环生成对抗谣言检测网络（WCGan），实现了 CycleGAN 网络在自然语言分类任务方面的应用。

WCGAN 使用生成模型对谣言特征与非谣言特征进行转换，使用循环一致性损失对生成器进行引导，可以使生成内容更加准确且更容易评估；使用 Wasserstein 距离作为生成损失，可以引导生成器生成更加多样化样本。与 GAN 谣言网络[2][3][4]相比，WCGAN 模型可以使谣言检测模型得到进一步的优化，减少数据不平衡对谣言检测结果的影响。在公开 PHEME 数据集上的实验证明，本文模型取得了更好的结果。

2 相关工作

近年来为了抑制谣言和假新闻在社交网络上的传播，诸多研究者针对如何使用人工智能检测谣言信息做了大量的研究与探索。基于文本内容的谣言检测研究以在社交网络上传播的文本内容为主要研究对象，使用监督学习的方式提取谣言内容特征，同时与用户转发行为、传播特征相结合提高谣言检测效率。

使用 RNN 可以很好的学习谣言文本的序列特征，Ma 等人[6]使用基于 RNN 的方法学习贴子上下文信息中隐藏的序列特征，通过长短期记忆网络（Long Short-Term Memory，LSTM）和门控循环单元（Gate Recurrent Unit，GRU）实现了对谣言更快更精准地检测。受分层任务和多任务学习的启发，Sujana[7]提出一种具有衰减因子的多损失分层双向长短期记忆（MHA）模型，利用贴子级和事件级的分层结构，可以更有效地从有限的文本中提取深层次的信息，有助于学习双边特征，并减少模型训练时间。Ma[8]基于递归神经网络提出一种树结构模型（RvNN），对非序列结构进行建模来学习贴子的判别特征，并生成更强大的表征实现谣言的识别。Chen[9]利用过滤后的标准化数据，充分考虑贴子的上下文关系，将双向门控循环单

元与强化学习方法相结合，提高了早期谣言检测的效率和准确性。Singh[10]提出了一种基于注意力机制的 LSTM 网络，通过融合文本特征和用户特征实现了网络谣言的鉴别。

GAN 网络[11]通过生成器与判别器的对抗训练，可以更好的学习样本潜在的概率分布。杨[12]使用 GAN 网络捕获双语文本的共享特征，并通过反向传播和优化转换矩阵相联合，降低了语言结构差异对自然语言分类任务带来的影响。Ma[3]使用 GRU 网络构建生成器和判别器，设计了一种基于 GAN 的谣言检测方法，通过生成器将谣言/非谣言转换为非谣言/谣言，可以更好地学习谣言的潜在特征，从而提高了谣言判别器的检测能力。考虑到 RNN 网络在长文本上会出现信息丢失的问题，李[2]使用 Transformer 代替了单一的 RNN 生成器，并在判别器中引入了注意力机制，提升了长文本序列谣言的检测能力。为了提高长文本训练的稳定性，Cheng[4]将对抗训练与强化学习的方法相结合，通过选择性的对原文中的某些词语进行替代，实现了有解释的文本级的谣言检测。

虽然通过引入更复杂的生成器和判别器可以提高基于 GAN 网络的谣言检测模型的精确度，但由于训练数据集中的谣言与非谣言样本并非成对存在，因此 GAN 网络难以学习不同数据域之间的映射关系，生成器难以生成足够有效的样本，Zhu[13]将循环一致性损失引入 GAN 网络，通过加强映射之间的约束关系，更好地实现了样本在两个数据域之间的映射。随着生成器与判别器在训练中不断优化，损失将无法有效引导训练过程，Arjovsky[5]将 Wasserstein 距离作为损失函数，通过最小化近似距离解决了 GAN 网络训练过程不稳定的问题。因此本文将在 GAN 网络的基础上，引入 CycleGAN 中的循环一致性损失以提升对生成器的引导效果，并使用 Wasserstein 距离改进生成损失，提高了模型的稳定性。

3 本文模型

信息在社交网络转发传播的过程中，不同的用户不断对信息进行转发和评论，信息传播链条不断增加，传播过程中会产生各种相互冲突和不确定的噪声信息，这些内容会与原始内容共存并一同传播，给谣言检测模型的特征学习过程带来了严重的负面影响。用户参与过程中带来的各种噪声信息，会使得原始内容在整个贴子中的频率不断降低，原始内容的特征被噪声信息隐藏，成为低频的隐藏特征，从而容易在特征学习的过程中被忽略；而传播过程中产生的噪声信息则可能更容易被谣言检测模型捕获并会成为关键特征。

利用对抗训练可以很好地模拟社交网络中围绕谣言核心内容不断传播变化的信息传播过程，进而增强谣言检测模型的特征学习能力。但是谣言样本与非谣言样本之间的不平衡性和非配对性极大地制约了生成器的训练效果，从而影响判别器的检测效果。本文的主要目的是利用改进的生成过程，提高生成器学习谣言和非谣言两个数据域之间文本内容、语义信息、传播特征的能力，迫使判别器可以更有效的学习谣言样本的低频隐藏特征，从而进一步提高谣言检测模型对社交网络谣言的检测能力。

本文将谣言的传播过程建模为信息的生成运动，以原贴子内容为起点，按照贴子传播期间评论产生的时间顺序，将原贴内容与评论内容拼接，将内容信息和传播信息结合。对于一个贴子，是由原贴内容以及后续的评论构成的一个集合$\{Y, p_1, p_2, \cdots p_n\}$，其中$Y$代表原贴内容，$p_i$代表后续的评论。不同的帖子拥有不同的传播结构，其中的各种评论会对帖子的真实性产生肯定或否定的指向，从而影响判别器的检测效果。生成器通过改变贴子中的内容信息以模糊原始样本的特征，从而产生与原始内容不同的指向性信息，如将支持原贴真实性的评论内容，转变为倾向于否定原贴真实性的内容。但在这个过程中，生成样本与原始样本在传播结构上是一致的，生成器仅改变内容而不创造新的传播过程。

3.1 基于循环生成对抗结构的谣言检测模型

生成对抗网络通过生成器与判别器之间的博弈实现两者的共同增强。生成器尽可能地生成符合目标特征的样本，以欺骗判别器；判别器通过使用原始数据和生成数据进行不断训练以实现最优化。通过反复多次的交替训练，判别器与生成器被不断增强，直到生成器可以近似生成真实样本，进而为判别器提供更丰富的训练样本，实现判别器判别能力的最大化。

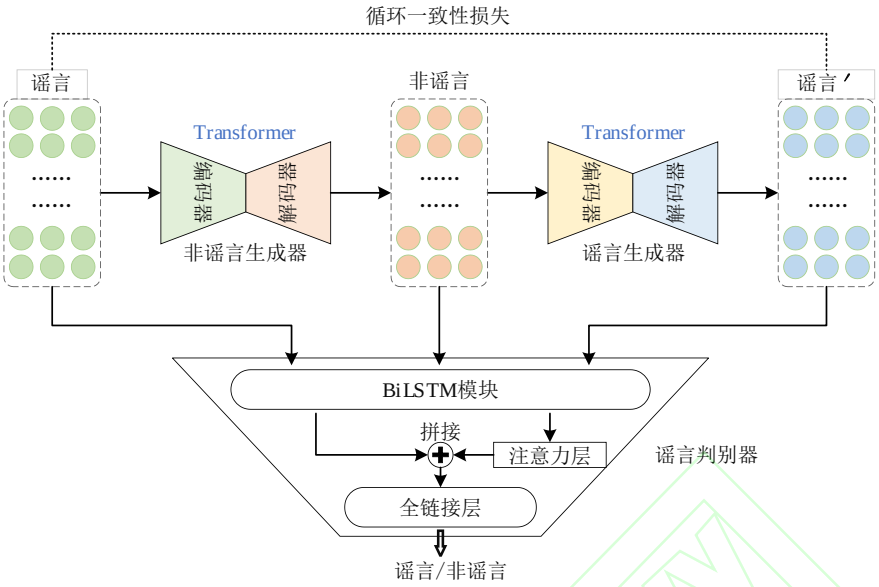

图 1 基于循环对抗结构的 WCGAN 谣言检测模型架构

Fig.1 WCGAN Architecture based on Cycle Generate Adversarial Framework

本文的目标是对谣言与非谣言序列的文本内容、语义信息、传播特征在两个数据域之间进行映射，使用生成器学习两个数据域之间的映射关系，对原始样本进行伪装，将谣言伪装成非谣言，将非谣言伪装成谣言，利用生成样本提高谣言检测模型对谣言深层特征的提取能力。

$$R \to G_N(R) \to N' \\ N \to G_R(N) \to R' \tag{1}$$

谣言检测问题作为一个二元分类任务，其目的是使用生成器最大可能地优化谣言判别器的训练效果。本文构造了两个生成器和一个谣言判别器，并通过对抗训练提升了谣言判别器的检测能力。如公式 1 所示，其中，G_N代表非谣言生成器，G_R代表谣言生成器，R代表谣言数据，N代表非谣言数据，N'代表由G_N生成的非谣言，R'代表由G_R生成的谣言。生成器的目标是实现$G_R(N) \to R'$和$G_N(R) \to N'$两个映射。即G_R的目标是使用非谣言数据生成具有谣言特征的数据，G_N的目标是使用谣言数据生成具有非谣言特征的数据。基于循环生成对抗网络结构的谣言检测模型架构如图 1 所示。

理论上通过对抗训练，G_N和G_R可以学习到谣言域与非谣言域之间的特征映射关系，从而输出与目标域相同分布的输出，但实际上仅依靠生成数据和目标域之间的对抗损失并不能保证模型将输入 x 映射到目标域的输出 y。

为了进一步减少可能的映射空间，从而确保映射函数可以稳定的映射至目标域，需要确保映射函数具有周期一致性。即对于每个谣言样本R，利用G_N将其映射为非谣言N'后，N'依然可以通过G_R映射回谣言样本。类似地，对于非谣言样本N，也具有相同的一致性。

对于判别器，需要有效地提取输入序列的特征并判断其类别，特别是低频的隐藏特征。GAN 网络的初始目标是获得一个可以生成真实数据的生成器，而 WCGAN 的目标则与之相反，是通过生成对抗训练利用生成器使判别器达到最优效果。对于训练生成器为主的 GAN 网络，目的是使生成器生成尽可能真实的数据，判别器的作用在于判断输入数据是否为生成数据，并为生成器提供反馈以提高生成器的生成能力。而本文模型的主要训练目标是判别器，生成器的任务是通过生成样本增强样本的多样性和突出样本的隐藏特征，以提高判别器对网络谣言的特征学习能力，此时 WCGAN 判别器的主要任务是学习数据的类别特征以实现对谣言信息的检测，而并不判断输入样本是否是真实数据。

3.2 生成模型与判别模型

本文的目的是通过对抗训练的方式提高判别器对网络谣言的检测能力，生成网络的主要作用是提取输入样本的核心内容特征，并对原始信息进行加工，模拟信息传播过程中无关噪音产生的过程，为判别器提供了更多得增强训练数据。

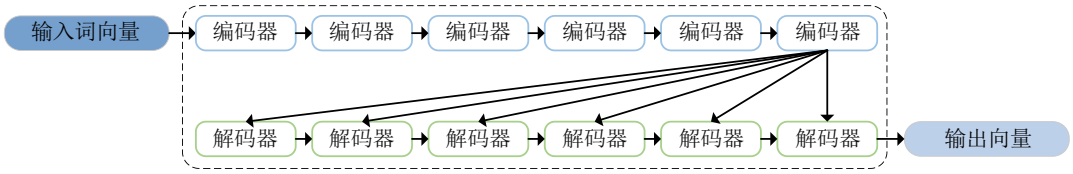

图 2　Transformer 模型架构

Fig.2 Framework of Transformer

Transformer 结构[14]可以很好地学习序列结构的特征，其拥有强大的语义特征提取能力和长距离特征捕获能力，可以很好地克服循环神经网络随着序列结构长度的增加而出现梯度爆炸或梯度消失的问题，同时拥有很高的计算效率，因此本文选择使用 Transformer 结构构建生成器。

Transformer 通过编码-解码架构的堆叠实现，每个堆叠结构都由一个编码器和一个解码器构成，其中编码器使用自注意力层与前馈神经网络构成，解码器在编码器的基础上增加了解码注意力模块，用于建立编码与解码关系。Transformer 模型架构如图 2 所示。编码器与解码器结构如图 3 所示。

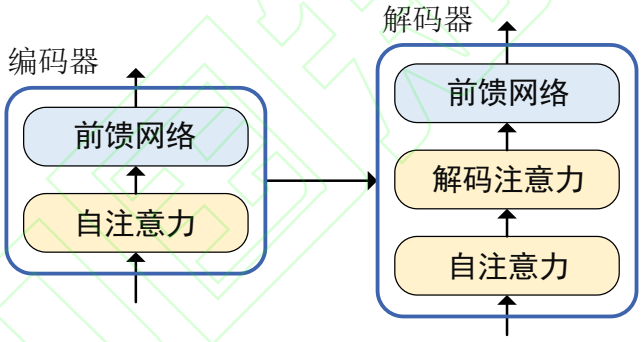

图 3　Transformer 编码器-解码器架构

Fig.3 Encoder-Decoder in Transformer

编码器中的注意力层为由自注意力层构成的多头注意力，对于一个长为n的输入序列$x=\{x_1,x_2,\cdots,x_n\}$，首先被编码器变换为具有不同线性投影的多个子空间，然后使用注意力函数计算其输出状态：

$$z=softmax\left(\frac{Q\cdot K^T}{\sqrt{d_k}}\right)\cdot V \tag{2}$$

其中，Q,K,V代表三个初始化状态矩阵，$\sqrt{d_k}$是一个缩放因子，d_k代表多头注意力中第k个头维度。最终的输出结果为所有头输出的拼接，最终通过归一化和前馈网络输出编码器结果。

解码器的目的是基于输入序列生成新的谣言/非谣言序列$x'=\left\{x_1',x_2',\cdots,x_n'\right\}$。编码器通过多头注意力层生成新的传播序列，并使用解码注意力模块来强化原始序列与生成序列之间的相关性。

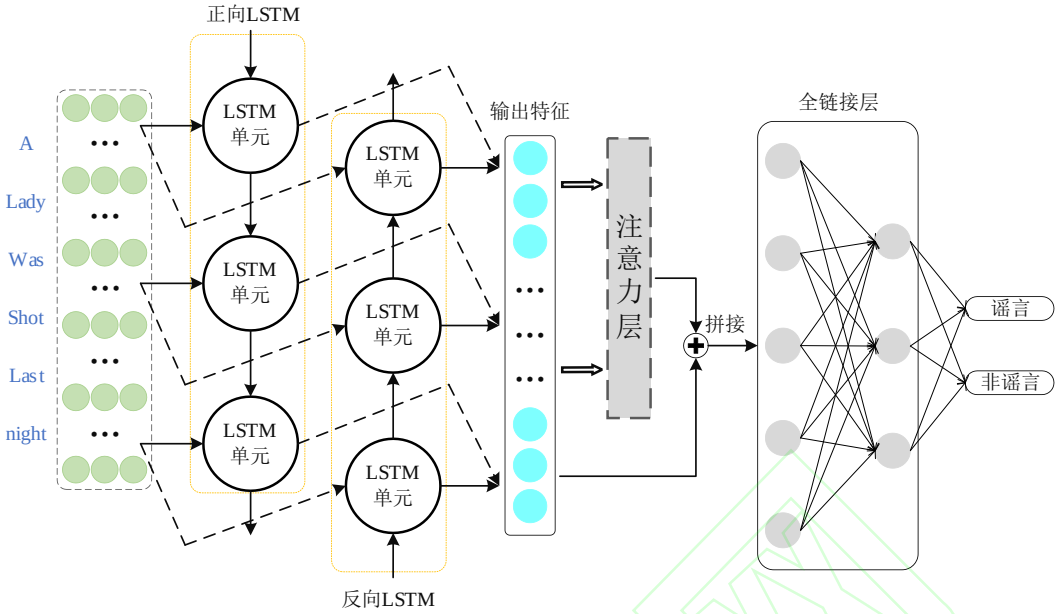

图 4 谣言判别器网络

Fig.4 Rumor Discriminator Network

谣言判别器网络架构如图 4 所示，判别器需要能够有效地捕获文本内容和传播过程中的关键特征，并过滤谣言序列中无关噪音的影响，本文判别器在 TextRNN-Att 模型[15] 改进的基础上实现，在保留了关键特征的基础上，融合了输入序列的原始特征。

$$s_t = BiLSTM(x_t, s_{t-1}; \theta_d) \tag{3}$$

如公式 3 所示，对于一个输入序列$x = \{x_1, x_2, \cdots, x_n\}$，首先将 t 时间步的特征向量x_t映射到隐藏向量s_t。s_{t-1}表示前一时间步的隐藏向量，θ_d代表 BiLSTM 的模型参数。

随后使用注意力机制计算每一时间步输出隐藏特征值的 attention 值，其代表了不同特征的重要程度，可以帮助判别器捕获更多关键特征，忽略无关特征的影响。attention 值的计算方法如公式 4 所示：

$$z = s \cdot soft\,max(\omega^T \cdot tanh(s)) \tag{4}$$

其中，s代表输入的特征向量，ω代表注意力参数，ω^T为ω的转置；z即经过注意力计算后的到的输出结果。注意力机制可以为序列中的每个特征赋予不同的重要程度，从而帮助模型将注意力集中在相对重要的信息上。为了增强对输入信息不同特征的利用，判别器通过对注意力输出特征与原始输出特征进行融合作为后续模型的输入。最终，使用全连接层对模型输出的融合特征进行线性运算，输出最终的二分类结果。

判别器的损失由真实样本的损失和生成样本的损失两部分构成，其中每个损失都定义为真实标签和输出标签的交叉熵：

$$L_D = l(y, \hat{y}) + \lambda l\left(y', \hat{y}'\right) \tag{5}$$

其中，l为交叉熵损失函数，y和$\hat{y}$分别表示原始样本的真实标签和判别器输出值，y'和$\hat{y}'$表示生成样本的真实标签和判别器输出值，生成样本的真实标签y'根据使用的生成器和原始样本的类别直接赋值，其值与y相反；若y代表真实谣言样本的标签，其值为“谣言”，则y'为使用G_N生成的非谣言样本标签，其值为“非谣言”；λ为权重系数。

3.3 基于 Wasserstein 距离的生成损失

随着判别器在训练过程中不断优化，生成器将倾向于生成更“安全”的样本，以尽可能地减小对抗损失，因此生成器将无法得到有效的信息，从而出现梯度消失的问题。因此需要对生成器进行进一步的引导，以帮助其生成更多样化且有效的样本。

Wasserstein 距离是两种概率分布之间的差值，可以反映生成器生成数据与实际数据之间的距离[5]。即便两个分布没有重叠，Wasserstein 距离仍然可以反应这两个分布之间的远近。因此采用 Wasserstein 距离[5]作为生成器的对抗损失（公式 6）和循环一致性损失(公式 7)，用以近似真实分布与生成分布之间的距离，当生成器的损失值越小，表明真实分布与生成分布之间的 Wasserstein 距离越小，生成对抗网络的训练效果越好。

本文模型的生成器损失由基于生成序列预测分类差值的对抗损失、识别映射损失以及循环一致性损失构成，其中对抗损失用于引导生成器向目标域进行映射，识别映射损失和循环一致性损失用于约束生成器的映射范围，以防止生成器生成非目标域分布的输出。

对抗损失定义为：

$$L_{GAN}(G) = E_{x\sim P_R}\left[D\left(G_N(x_R)\right)\right] - E_{x\sim P_R}[D(x_R)] \quad (6)$$

如图 5 所示，对于每个通过生成器生成的非谣言样本，在具有相应的非谣言特征的同时，应该可以通过谣言生成器映射回谣言，即生成器生成的样本应该与原始样本拥有周期一致性。

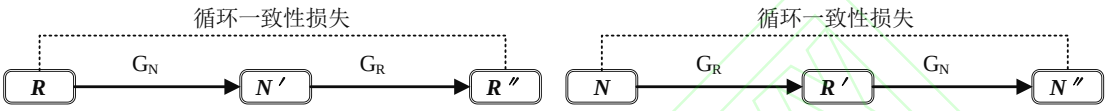

图 5 生成样本与原始样本之间的循环一致性

Fig.5 Cycle Consistency Between the Generated Sample and the Original Sample

其中，R和N分别代表原始的谣言和非谣言样本，N'和R'代表由生成器生成的非谣言和谣言，R''和N''代表基于生成样本N'和R'再次生成的谣言和非谣言。

本文使用循环一致性损失来实现生成器的周期一致性：

$$L_{cyc} = E_{x\sim P_R}\left[D\left(G_R\left(G_N(x_R)\right)\right)\right] - E_{x\sim P_R}[D(x_R)] \quad (7)$$

数据之间的传递性可以对数据结构进行调整，例如可以通过回译的方式提升文本翻译的效果[13]。由于原始样本中包含了文本信息与传播结构信息，为了保证生成样本与原始样本在文本内容与传播结构两方面都具有一致性，使用循环一致性损失实现生成器的周期一致性。利用谣言-非谣言-谣言的生成过程，采用类似回译的方式，通过数据在生成器之间的传递，可以提高生成样本在传播结构方面的合理性，保证生成样本具有与原始样本一致的传播结构。

循环一致性损失可以尽可能地保证谣言生成器生成的样本是谣言，但当生成器过于强大时，可能会对输入的谣言样本进行不必要的改动，以至于生成的样本与原始样本并不相关。因此对于谣言生成器，无论输入的是何种类型的样本，都应该能够输出相应的具有谣言特征的目标样本。对于谣言生成器，当输入样本为谣言时，输出的生成样本R'也应为谣言；对于非谣言生成器，当输入样本为非谣言时，输出的生成样本N'也应为非谣言。本文模型使用识别损失对生成器进行约束，识别损失定义为：

$$L_{id} = \sum y log\hat{y} \quad (8)$$

y是真实样本的标签，$\hat{y}$为生成器的生成样本经判别器判别后输出的预测值。识别损失可以避免生成器在进行迁移映射时迁移过多，从而生成非目标域的内容。当向生成器输入目标域的样本时，生成器的输出应该是输入样本的近似映射。使用识别映射损失可以实现这一目标。识别损失使用交叉熵损失函数而非 Wasserstein 距离作为损失函数，是因为生成器会改变输入序列的特征，输入的R和生成的R'只需要被判别为相同类别的样本，而非在序列特征上具有极大的相似性。

$$L_G = L_{GAN}(G) + \alpha L_{cyc} + \beta L_{id} \quad (9)$$

因此L_G即为最终的生成损失，其中α和β代表相应损失的权重系数。

4 实验结果与分析

4.1 实验数据集

现实世界中的社交平台上不同类型的内容的数量往往是极度不平衡的，通常真实信息的数量要远远多于虚假信息，不同事件之间的内容数量差距也极大。同时针对单一谣言事件，社交网络中往往缺乏事实核查内容却保有大量不同用户之间的交互内容。因此谣言检测模型往往在现实环境中表现较差。

本文选择使用公开数据集 PHEME9[16]进行谣言与非谣言的二元分类实验，PHEME9 数据集是一个不平衡数据集，由九个突发新闻相关的贴子构成，这些贴子是这些突发新闻发生期间在 Twitter 上传播的谣言与非谣言数据，不同事件之间的样本数量、每个事件下的谣言与非谣言样本数量的不平衡性都很大，如样本数量最大的 Charlie Hebdo 事件有 2070 条样本，其中谣言样本 456 条、非谣言样本 1614 条；而贴子样本数量最小的 Ebola Essien 事件只有 14 条样本，而且全部都是谣言样本。

表 1　数据集处理样例

Table1 Samples of processed dataset

序号	原始数据	处理后数据	标签
原贴	France: 10 people dead after shooting at HQ of satirical weekly newspaper #CharlieHebdo, according to witnesses http:\/\/t.co\/FkYxGmuS58	france 10 people dead after shooting at hq of satirical weekly newspaper according to witnesses [SEP] mt if zionists jews did this they d be nuking israel [SEP] they who stupid and partial opinions like this one only add noise to any debate [SEP] socialists antisemites anti zionists usual suspects [SEP] islamistas o taraos [SEP] a french crime of passion or another heathen moslem atrocity [SEP] lol 5 million muslims in france what a disgrace the french worm president and politicians killed them tine for croissants now	谣言（1）
评论 1	MT @euronews France: 10 dead after shooting at HQ of satirical weekly #CharlieHebdo. If Zionists/Jews did this they'd be nuking Israel		
评论 2	@j0nathandavis They who? Stupid and partial opinions like this one only add noise to any debate.		
评论 3	@nanoSpawn Socialists, Antisemites, anti zionists - usual suspects		
评论 4	@Nuno_Rodrigo5 @euronews islamistas o taraos?		
评论 5	@euronews @TradeDesk_Steve A French crime of passion or another heathen moslem atrocity?		
评论 6	@euronews LOL. 5 million Muslims in France, what a disgrace. the french worm president and politicians killed them. tine for croissants now		
原贴	Police car with bullet holes in front of Charlie hebdo headquarter. Photo @WilliamMolinie http://t.co/k6inhBmGsf	police car with bullet holes in front of charlie hebdo headquarter photo [SEP] charlie hebdo s last cartoon on twitter abu bakr al baghdadi saying and especially health new year s greeting [SEP] this was posted just 10 mins before attacks it s not in response to this specific cartoon pen is mightier than sword	非谣言（0）
评论 1	Charlie Hebdo's last cartoon on Twitter: Abu Bakr al-Baghdadi saying \"and especially, health!\" (New Year's greeting) http:\/\/t.co\/UMEWDeR6z4		
评论 2	@DavidKenner Fuck this animal and his #ISIS rats.		
评论 3	@DavidKenner @Ali_Gharib This was posted just 10 mins before attacks https:\/\/t.co\/iehJqib1H1 it's not in response to this specific cartoon		
评论 4	@DavidKenner @IrinaGalushkoRT Pen is mightier than SWORD		

由于初始数据集本部分事件的样本数量太少，难以有效利用，本文实验在初始数据集的基础上删除了部分无效样本，最终使用了六个事件，得到 5998 条样本作为本文的实验数据集。每条样本内容包含一个贴子的初始内容以及后续的评论信息，以内容和评论的时间先后顺序进行拼接，评论与评论之间使用分隔符进行标记，在最大限度利用贴子文本内容的同时尽可能地保留贴子一部分的传播过程信息。拼接后得到的样本，使用正则表达式对其中的话题、@用户字符、表情符号、网址信息和标点符号信息等内容进行数据清洗，得

到最终实验所用的数据集。数据集样本共分为两类，非谣言和谣言。数据集处理样例见表1，数据集最终统计信息如表2所示。

表2　PHEME数据集各事件信息统计表

Table2 Statistical Information for Event Information of PHEME Dataset

事件	贴子	谣言	非谣言	评论数
Charlie Hebdo	2070	456	1614	35303
Sydney siege	1210	519	691	16455
Ferguson	1131	283	848	22881
Ottawa shooting	884	469	415	9854
Germanwings-crash	466	238	228	3896
Putting missing	237	125	112	759
合计	5998	2090	3908	89148

4.2　实验设置

对于模型的评价标准，本文使用准确率和F1值作为评估指标。本文将数据集中的20%作为测试集，其余的作为训练集，六个新闻集中每个新闻集的谣言与非谣言样本中，皆等比例随机选取20%作为测试集，以保证实验的准确性。由于Wasserstein损失在使用Adam这类基于动量的优化算法进行优化时，损失的梯度会变得非常不稳定[5]，因此本文模型在训练过程中采用适合梯度不稳定情况下使用的RMSProp梯度优化算法。

在实验中选取了目前先进的谣言检测模型进行对比：

GRU、LSTM[6]：使用双层的GRU网络和LSTM网络捕获贴子上下文信息以及谣言序列的时间变化特征。

CNN[17]：基于内容特征，使用CNN网络提取输入序列中分散的关键特征。

GAN-GRU[3]：使用GRU网络作为判别器和生成器构建生成对抗网络的谣言检测模型。

GAN-Tran[18]：一种基于Transformer结构生成器的生成对抗网络谣言检测模型，同时考虑了谣言的内容特征与传播结构特征。

RvNN[8]：基于递归神经网络对谣言传播结构进行建模，利用文本信息和传播结构信息实现谣言的检测。

DAN-Tree[19]：用Transformer学习不同贴子间的语义关系，并利用注意力机制学习传播路径节点的重要度，利用传播树结构的表示向量实现谣言的判别检测；

gDART[20]：通过全局离散注意力捕捉不同词序列之间的相关性，利用特征融合网络对不同分支特征进行融合以提高谣言鉴别能力。

WCGAN：本文模型，使用LSTM和注意力机制构建生成器与判别器，基于CycleGAN网络和Wasserstein损失构建的谣言检测模型。

4.3　结果分析

实验结果表明，利用对抗训练可以有效地提升谣言检测模型的检测效果，而基于改进的循环对抗生成网络的谣言检测模型可以进一步提升检测效果。同时与最新的方法相比较，本文模型在各个指标上也获得了优秀的提升。

表3是不同基准模型和本文模型在PHEME数据集上的实验结果。可以看出，完全基于数据驱动的RNN、CNN网络，在一定程度上能够从内容中识别出谣言特征，但是由于主要依赖于上下文的文本内容，因此结果并不理想。在谣言判别器相同的条件下，利用对抗训练可以有效地提升谣言检测模型的效果。如GRU和GAN-GRU，使用对抗生成网络可以明显地增强检测模型的效果，这是因为生成器可以生成丰富样本帮助判别器捕获判别特征。GAN-Tran在与GAN-GRU使用相同的判别器的情况下，使用Transformer构建了生成器，在结果上取得了较大的提升，一方面是因为Transformer在长文本处理方面拥有更优秀的能力，另一方面同时考虑了谣言的传播特征。

表 3　PHEME 数据集谣言检测实验结果

Table3 Experimental Results of Rumor Detection in PHEME Dataset

模型	Acc	F1	Precision	Recall
GRU	0.742	0.739	0.7455	0.7415
LSTM	0.7541	0.7413	0.7543	0.7548
CNN	0.754	0.733	0.758	0.732
GAN-GRU	0.781	0.778	0.782	0.796
GAN-Tran	0.821	0.809	0.8217	0.8013
RvNN	0.855	0.8305	0.8445	0.817
DAN-Tree	0.845	0.830	0.8239	0.8362
gDART	0.863	0.837	0.8414	0.8325
WCGAN	**0.8698**	**0.855**	**0.8617**	**0.8484**

而基于传播特征与内容特征相结合的 RvNN、DAN-Tree、gDART，性能普遍优于基于内容特征的方法。这是因为谣言传播的过程是网络谣言形成的一个重要特征，社交网络中的交互行为与交互中产生的评论内容信息，都对网络谣言的形成具有重要影响。

虽然基于传播特征可的方法通过学习谣言传播的过程信息，可以更好地刻画谣言传播的特点。但是只有当谣言广泛传播并形成一定规模后，才形成足够有效的传播特征。PHEME 数据集中的贴子内容来源于社交平台上的数个突发新闻，因此在频繁的评论交互过程中，不同贴子之间的信息会有相互的重叠，同时每个贴子平均只有 15 条评论信息，无论贴子的文本内容还是每个贴子的传播途径，可以获取的有效信息都相对有限。

本文模型与基准模型相比，在各个指标上的得分都有很大程度的提升。因为利用循环生成对抗网络可以在原有数据的基础上，通过对抗训练生成新的样本，可以在一定程度上丰富样本特征，帮助谣言检测模型学习到更加复杂的隐藏特征。同时与其他模型相比，本文模型的召回率实现了明显的提升，进一步克服了数据样本类别不平衡对谣言检测结果的影响，因为通过生成样本，可以在一定程度上平衡不同类别数据的数量，从而使实验结果更为平衡。

表 4　WCGAN 在不同事件上的结果

Table4 Results of WCGAN on different events

事件	Acc	F1	Precision	Recall
Charlie Hebdo	0.8873	0.8251	0.8535	0,8045
Ferguson	0.8375	0.7755	0.7935	0.7624
Germanwings crash	0.825	0.825	0.8282	0.8271
Ottawa shooting	0.8744	0.8743	0.8758	0.878
Putin missing	0.7896	0.7891	0.7981	0.793
Sydney siege	0.8293	0.8257	0.8281	0.824

表 4 为本文提出的 WCGAN 模型在 PHEME 数据集上不同事件上的实验结果，其中在 Charlie Hebdo、Germanwings crash、Ottawa shooting、Sydney siege 四个事件上都取得了较高的指标得分，因为这些事件中的谣言与非谣言数量相对平均或是样本数量相对较多，模型可以获取到更多有效的交互信息，所以取得了远超其他事件的得分。不同种类样本数量的不均衡会导致召回率偏低，从表 4 可以看出，在多数事件上，本文模型实现了召回率与其他指标之间的平衡，极大地克服了数据不平衡对谣言检测带来的影响。在 Charlie Hebdo 和 Ferguson 两个事件中，召回率都偏低于其他指标的值，这在很大程度上都是因为这两个事件的样本不平衡性要更严重，特别是 Charlie Hebdo 事件中，谣言样本与非谣言样本之间的数量差达到一千二百多条。一是因为不同类别的样本数据量差别太大，仅通过生成样本无法完全弥补这种数据失衡带来的影响；二是本文模型仅通过对内容的调整实现样本生成过程，而并未改变样本的传播结构，也并未生成新的传播结构，虽然通过生成样本弥补了不同类别样本数量的失衡，但是不同类别样本的传播结构丰富度依然有差别。

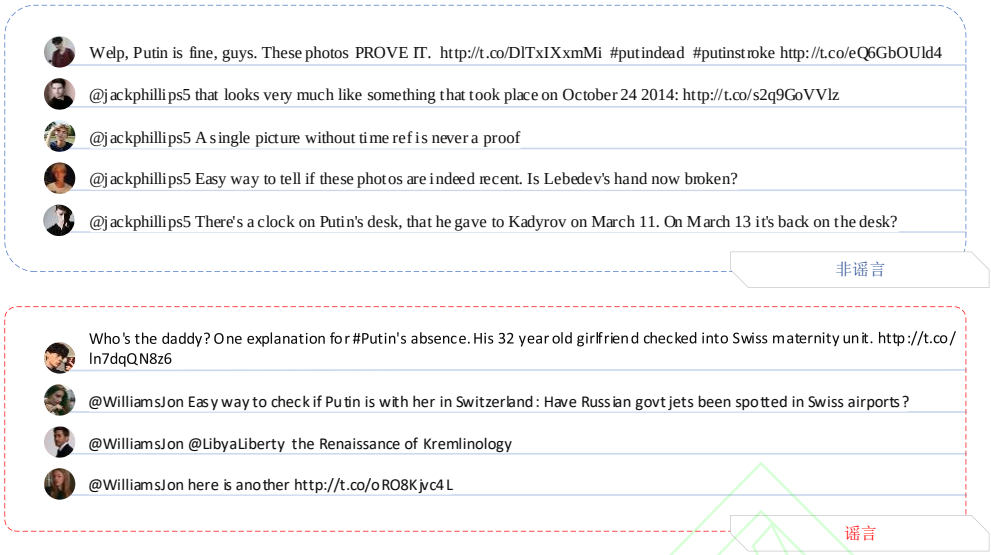

图 6　putin missing 事件中的谣言、非谣言样本

Fig.6 Rumors and Non-rumor Examples in the putin missing Event

在实验数据集中，Putin missing 事件在各个指标上得分都比较低且不同指标的值之间比较平衡。这可能是由于这个事件中的每个贴子下的评论数太少所导致，数据集中平均每个贴子有 15 条评论，而 Putin missing 事件的贴子下，每个贴子平均只有 4 条评论，因此每个贴子中的有效信息更少，但由于谣言与非谣言样本数量之间相对平衡，因此不同指标之间没有产生太大差距。

Putin missing 事件样本如图 6 所示，评论转发次数更少。同时每一相同事件中贴子的内容往往聚焦于本事件，因此这些事件之间具有很强大的关联性，分类模型可能会更聚焦于对不同事件进行分类，从而导致事件特征掩盖了谣言特征。同时这也是许多模型在 PHEME 数据集上检测效果较差的原因，而使用对抗网络进行训练，可以更好的帮助判别器关注样本的谣言特征而非事件特征，从而实现谣言检测效果的提高。

4.4　消融研究

为了验证在数据不平衡条件下，循环生成对抗网络的模型结构和改进的生成损失对谣言检测效果的提升，进行了消融研究。正负样本比例代表了每个事件中谣言样本与非谣言样本的比例，比例被设置为 1:1、1:2、1:3，训练集中的非谣言样本占比逐渐增加，数据失衡情况不断加剧。消融实验结果见表 5，其中 D-only 是仅使用本文模型中的判别器而不使用任何对抗训练的实验结果，GANbase 是基于本文模型的生成器与判别器构建的 GAN 结构谣言检测模型，WCGAN 是基于 CycleGAN 生成损失的检测模型，WCGAN-w 是基于 Wasserstein 损失改进生成损失的检测模型。

表 5　消融实验结果

Table5 Ablation Experiments

正负样本比例	1:1		1:2		1:3	
模型	Precision	Recall	Precision	Recall	Precision	Recall
D-only	0.7851	0.7843	0.7993	0.754	0.7894	0.6876
GANbase	0.821	0.8143	0.8185	0.8044	0.8127	0.7813
WCGAN	0.8494	0.8401	0.8379	0.8203	0.8381	0.8144
WCGAN-w	0.8693	0.8614	0.8634	0.8553	0.8572	0.8395

从表 5 中可以看出，与没有使用对抗训练，只使用判别模型进行训练相比，使用了对抗训练后，谣言检测模型的效果获得了极大的改进，说明对抗训练为判别器训练效果的提升提供了很大的帮助。WCGAN 在 GAN 谣言检测模型的基础上引入了循环一致性损失，这个损失可以很好的帮助生成器生成更加有效的样本，提高了生成样本的质量。在引入 Wasserstein 距离改进生成损失后，生成效果进一步改进，从而为模型的谣言检测结果带来了积极的影响。

采用不平衡的数据训练模型时，样本不均衡的程度往往会给分类指标带来明显的变化，特别是 Precision 值和 Recall 值会出现明显的失衡，这是由于模型偏向于预测样本数量大的类别所导致的。表 5 中实验结果表明，当不使用对抗训练时，D-only 在正负样本比例为 1 时，实现了 Precision 值与 Precision 值之间的平衡， 但随着非谣言样本占比增加，Recall 值不断降低，和 Precision 值逐渐失衡，表明模型在数据失衡的情况下，检测出正样本的能力迅速下降。但是通过对抗训练，即便数据失衡情况不断严重，模型也在两个指标之间取得了平衡，说明数据不平衡对模型带来的影响较小。证明本文模型极大程度地提升了谣言检测任务在数据不平衡情况下的稳定性和精确性。

5 结语

本文提出了一种基于循环对抗生成网络架构的谣言检测方法，并利用 Wasserstein 距离改进了模型的生成损失，提高了谣言检测任务在不平衡数据集中模型的稳定性和精确性。该方法以原始数据集为基础，利用对抗训练的方式生成更多的贴子数据，将同一事件下的谣言与非谣言相互转换，生成新的传播链与贴子内容，从而可以增强有效信息相对有限的条件下模型对贴子信息真实性的判别能力。基于 CycleGan 网络的生成架构可以提升谣言样本与非谣言样本非配对情况下的样本生成效果，使用 Wasserstein 距离改进的生成损失，极大限度地保证了生成数据的有效性多样性，减少了不同类别样本数据量不平衡对检测结果的影响。实验结果表明，本文模型在谣言检测在基准数据集上取得了出色的成绩。

本文的谣言检测模型主要依靠贴子的文本内容，虽然利用了转发评论过程中的传播结构信息，但也仅使用了评论的内容信息，而未能有效利用传播过程上的用户特征信息等；同时贴子中往往包含有与文本内容相关的图片、视频以及其它来源的引用信息，这些信息都未被应用，以至于很多贴子仅凭文本内容无法获取足够的有效信息，因此未来可以考虑在模型中加入多模态的内容。同时 CycleGAN 网络仅涉及两个数据域之间的相互迁移转换，因此本文模型仅适用于二分类的谣言检测任务，可以采用更加复杂多样化的生成对抗网络架构，以实现在多分类谣言检测任务中的应用。

参考文献：

[1] Varshney D, Vishwakarma D K. A review on rumour prediction and veracity assessment in online social network[J]. Expert Systems with Applications,2021,168: 114208.

[2] 李奥, 但志平, 董方敏, 等. 基于改进生成对抗网络的谣言检测方法[J]. 中文信息学报 ,2020,34(9): 78-88.(Li Ao, Dan Zhiping, Dong Fangmin, et al. An Improved Generative Adversarial Network for Rumor Detection[J]. JOURNAL OF CHINESE INFORMATION PROCESSING, 2020,34(9): 78-88.)

[3] Ma J, Gao W, Wong K-F. Detect rumors on twitter by promoting information campaigns with generative adversarial learning[C]. In: Proceedings of The world wide Web conference, 2019: 3049-3055.

[4] Cheng M, Li Y, Nazarian S, et al. From rumor to genetic mutation detection with explanations: a GAN approach[J]. Scientific Reports, 2021, 11(1): 1-14.

[5] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]. In: Proceedings of International conference on machine learning, 2017: 214-223.

[6] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[C]. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016: 3818–3824.

[7] Sujana Y, Li J, Kao H-Y. Rumor detection on twitter using multiloss hierarchical bilstm with an attenuation factor[J]. arXiv preprint arXiv:2011.00259

[8] Ma J, Gao W, Joty S R, et al. An Attention-based Rumor Detection Model with Tree-structured Recursive Neural Networks[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2020, 11: 1-28.

[9] Chen X, Wang C, Li Dong, et al. A New Early Rumor Detection Model Based on BiGRU Neural Network [J]. Discrete Dynamics in Nature and Society, 2021: 2296605.
[10] Singh J P, Kumar A, Rana N P, et al. Attention-Based LSTM Network for Rumor Veracity Estimation of Tweets[J]. Information Systems Frontiers, 2022, 24(2): 459-474.
[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020,63(11): 139-144.
[12] 杨文丽,李娜娜.基于对抗网络的文本对齐跨语言情感分类方法[J].数据分析与知识发现 ,2022,6(7):141-151.(Yang Wenli, Li Nana. A Text-Aligned Cross-Language Sentiment Classification Method Based on Adversarial Networks[j]. Data Analysis and Knowledge Discovery, 2022,6(7):141-151.)
[13] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. In: Proceedings of the IEEE international conference on computer vision, 2017: 2223-2232.
[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000–6010.
[15] Zhou P, Shi , Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016: 207-212.
[16] Kochkina E, Liakata M, Zubiaga A. All-in-one: Multi-task Learning for Rumour Verification[C]. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: ACL, 2018: 3402-3413.
[17] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, A convolutional approach for misinformation identification[C]. In: Proceedings of IJCAI, 2017, 3901–3907
[18] Ma J, Li J, Gao W, et al. Improving Rumor Detection by Promoting Information Campaigns with Transformer-based Generative Adversarial Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2021.
[19] 韩雪明, 贾彩燕, 李轩涯, 等. 传播树结构结点及路径双注意力谣言检测模型[J].计算机科学, 2022: 1-16. (Han Xueming, Jia Caiyan, Li Xuanya, et al. Dual-attention Network Model on Propagation Tree Structures for Rumor Detection[J/OL]. Computer Science, 2022: 1-16.)
[20] Roy S, Bhanu M, Saxena S, et al. gDART: Improving rumor verification in social media with Discrete Attention Representations[J]. Information Processing & Management, 2022, 59(3), 102927.

通讯作者（Corresponding author）:但志平（Dan Zhiping），ORCID：0000-0002-2616-5730，E-mail: zp_dan@ctgu.edu.cn。

基金项目：本文系国家自然科学基金—新疆联合基金项目(U1703261)的研究成果之一。
The work is supported by National Natural Science Foundation of China-Xin Jiang Joint Foundation (Grant No. U1703261).

作者贡献声明

张洪志：构建实验模型，进行实验及论文撰写；
但志平：提出研究思路和研究方案，论文修改和定稿；

董方敏：确定论文选题，讨论研究方案，提出修改意见；
高准：辅助实验，论文修改；
张岩珂：模型修订，论文修改。

利益冲突声明：
所有作者声明不存在利益冲突关系。