融合信息对抗及混合特征表示的社交网络谣言检测方法*

朱 贺

(河南师范大学图书与档案信息中心 新乡 453007)

摘 要:[**研究目的**]针对现实社交网络中广泛存在的不实评论对谣言检测的负面影响问题,提出对抗学习框架下的谣言检测方法,从而在提升谣言检测准确率的同时,增强模型对噪声信息的容抗性。[**研究方法**]以信息对抗机制为基础,搭建具有融合结构及时序特征表示的生成网络,利用部分网络结构的共享及加强具有自注意力机制的二次鉴别网络,实现将非监督的对抗生成网络向有监督学习任务上的成功拓展。[**研究结论**]在PHEMEv5和新浪微博两个数据集上,该研究提出的模型在谣言检测的准确率上,相较于9种较为先进的基准模型至少提升了3.1%和4.1%;同时,实验显示,该研究提出的模型对于噪声信息并不敏感。充分证明了该模型在跨平台不同语言环境数据集上较高的谣言检测效果及较强的噪声容抗性。

关键词:网络谣言;谣言检测;信息对抗;对抗生成网络;特征融合;自注意力机制

中图分类号:TP391 **文献标识码**:A **文章编号**:1002-1965(2024)02-0118-10

引用格式:朱 贺.融合信息对抗及混合特征表示的社交网络谣言检测方法[J].情报杂志,2024,43(2):118-127.

DOI:10.3969/j.issn.1002-1965.2024.02.017

Social Networks Rumor Detection Based on Fusion of Information Campaign and Hybrid Characteristic Representation

Zhu He

(Information Center of Library and Archives, Henan Normal University, Xinxiang 453007)

Abstract: [**Research purpose**] In order to cope with the negative impact of untrue comments on rumor detection, this paper proposes a rumor detection method under the adversarial learning framework, which can improve the accuracy of rumor detection and enhance the tolerance of the model to noise information. [**Research method**] Based on information campaign, a generative network is built with fusion of the structural and temporal characteristic representation networks. By means of the sharing of partial network structures, and the integration of the self-attention mechanism and secondary discriminative network, the unsupervised generative adversarial network is successfully extended to supervised learning tasks of rumor detection. [**Research conclusion**] On the two public datasets of PHEMEv5 and Weibo, compared with the nine typical baseline models, the accuracy of the proposed model is improved at 3.1% and 4.1% at least. Meanwhile, further experiments show that the proposed model is not sensitive to the noise messages on the task of rumor detection. The model proposed in this study improves the performance of rumor detection on cross platform datasets in different language environments, and shows a high tolerance for noise information.

Key words: networks rumor; rumor detection; information campaign; generative adversarial network; hybrid characteristic representation; self attention mechanism

0 引 言

线上社交网络为信息的传播提供了快速的传播通路,而谣言作为信息的一种特定存在形式,自然也包括其中。谣言传播造成了巨大的经济损失,同时也给社会的平稳运行提出了严峻的挑战。在新冠疫情的背景

收稿日期:2023-03-13 修回日期:2023-04-14

基金项目:2021年度河南省哲学社会科学规划年度项目“线上社交网络上舆情监测及引导策略研究”(编号:2021CZH021);2022年度河南省高校人文社会科学研究一般项目“互联网舆情信息监测及引导策略研究”(编号:2022-ZZJH-419)研究成果。

作者简介:朱 贺,男,1987年生,博士,馆员,研究方向:自然语言处理、谣言检测、舆情传播。

下,多种涉及民生事件的谣言被居心不良的个人或者团体捏造并传播,对社会和谐、民心安定和政府治理造成了巨大的负面影响。谣言的传播已经成为了一项社会问题,相关领域的研究必须给与足够的重视。

现有高性能的谣言检测方法大多建立在广泛的特征提取和大规模的数据分析上,此类方法在一定程度上提高了谣言检测的效率和精度,使得在大规模传播事件中实现对谣言的甄别成为了可能。然而,需要指出的是,网络环境并不是“一尘不染”的,在广泛的自由互动的背景下,舆情参与个体变得更加复杂,舆情事件中往往也会包含着一定量的虚假评论或恶意陈述。遗憾的是,数据驱动的谣言检测方法并未对舆情信息中充斥着的各种“噪声”做出应有的应对,这也就限制了谣言检测精度在当前日益复杂的舆情传播背景下进一步的提高。

基于此,本研究提出了一种融合信息对抗及混合特征表示的社交网络谣言检测模型,从而在现实情形中广泛存在不实表达的背景下,增强模型对于“噪声信息”的容抗性,提高谣言检测的准确度。本模型利用混合特征呈现的方法,从传播时序和扩散结构双重维度来解析舆情事件,提取抽象化的高维谣言鉴别变量,克服了单一考虑“树形拓扑”或者“时序依存”时特征呈现不充分的缺点。此外,借助于信息对抗,在网络构建及学习过程中,采用竞争机制,利用舆情评论数据生成对抗性的虚拟噪声声音,推动谣言鉴别器在成功识别提取的混合谣言特征的同时,不断对生成的对抗性声音做出有利于正确识别谣言方向的应答,达到同步提升模型谣言检测精度和噪声容抗性的目的。

1　相关研究

现代谣言检测研究的重点在于适用于大规模且自动化的舆情处理,受益于人工智能技术的发展,线上社交网络上舆情信息的即时识别变得不再遥不可及[1-2]。一部分学者认为,谣言同真实信息之间存在着一些显性的,诸如在语法、句法、词汇或者情感表达等特征标志位的不同,而这些标志位的特征差异正可被利用作为识别谣言的依据[3-5]。Gupta 等[6]从发布的舆情信息中提取了一个多达 45 个标志位的谣言特征集,建立了一个即时的评估线上舆情可信性的分析系统。Popat[7]提出了一类包含“断言性短语”“词语符号”“主观判断”等变量的谣言判定特征集,并比较了在不同特征集组合下的谣言检测效果。为了提高谣言的识别效果, Yang[8]和 Sun[9]在谣言特征标志集中进一步加入了对谣言发布个体特征的描述,实现了对谣言特征更加全面的呈现。这些基于特征的谣言鉴别方法为自动化谣言检测提供了可能性,然而在任务前期却需要耗费大量的人力进行特征的筛选,提高了谣言鉴别的成本。此外,在现实情形中,不同平台、不同兴趣群的社交群体之间的信息交互方式是不同的,这就要求基于特征的谣言鉴别针对不同的平台设计不同的特征集,而这也就限制了此类方法跨平台的泛化能力。

机器学习特别是深度学习技术的发展促进了数据驱动的谣言检测方法的研究。得益于智能化的信息分析流程,数据驱动的谣言检测不再需要依赖于前期大量的人工特征提取,真正实现了谣言检测方法的自动化,提升了其跨平台的适用性[10-13]。为了降低话题偏移对突发性事件中谣言检测精度的影响, Alkhodair 等[13]提出了一个基于 word2vec 和循环神经网络 RNN 的融合监督及非监督性学习过程的谣言检测模型。Ma 等[14]提出了两类“自上而下”和“自下而上”的树形结构递归神经网络来呈现谣言传播过程中相关信息之间的层级结构关系,同时,在其随后的研究工作中[15],注意力机制也被融入到了树形结构的构建过程中,提高了关键性谣言检测信息点的识别。刘勘等[16]基于双层 LSTM 及迁移网络,分析并提取了用户及传播特征,提出了一种在无标注数据情况下的跨领域谣言检测策略。上述数据驱动的方法提高了谣言检测的准确性,但却只考虑了谣言传播的时序或结构特征,缺少了更加全面的混合特征的呈现。同时,现有的数据驱动方法也没能对现实情形中广泛存在的“噪声”信息做出充分的应对,而这也就进一步限制了此类方法在谣言检测效果上的提升。

生成对抗网络 GAN 在其提出之初就引起了学术界和业界极大的关注,被广泛的应用于图像和视频生成等非监督性学习任务之中。在“生成器”和“鉴别器”之间的对抗学习机制的作用下,生成对抗网络变得有能力提取出原本不易被学习或者提取出的“非显性特征”。为了将适用于非监督学习的生成对抗网络移植到监督学习的谣言识别任务中,Ma 等[17]首次提出了信息对抗的概念,他们利用基于循环神经网络 RNN 的生成器产生争议性的对抗言论,从而给鉴别器更大的压力使其更好的识别谣言文本中具有指示性的辨别特征。孟佳娜等[18]基于对抗神经网络提出了一个混合文本信息以及图片信息的跨模态谣言检测模型,提高了谣言检测的特征迁移能力。Cheng 等[19]建立一个具有智能化自学习能力的谣言信息输入序列修正模型,他们提出的基于 GAN 的模型框架很好地解决了谣言识别过程中“谣言”和“非谣言”数据的不平衡性。基于对抗学习的谣言检测方法加强了对谣言数据中“噪声”信息的处理能力,然而,由于生成对抗网络对于非监督学习的特异性,其模型关注的重心落在生成器而不是鉴别器上,生成器的对抗机制必然会降

低鉴别器的鉴别效果，这就要求在针对有监督任务时对鉴别器应做额外的加强处理，但遗憾的是，上述研究却没能给出有效的解决。

2 融合信息对抗及混合特征表示的社交网络谣言检测方法

本研究提出的融合信息对抗及混合特征表示的谣言检测方法（简称 IHCR，代指 Information-campaign and Hybrid Characteristic Representation）在整体上可以看作是一个具有部分参数共享的“双步”模型：在“第一步”中，为了提升对噪声信息的容抗性，本模型借鉴了 Wasserstein GAN（WGAN）[20-21] 以及 Auxiliary Classifier GAN（ACGAN）[22] 的模型构建思想，提出了监督性学习任务背景下的生成对抗网络，实现了针对谣言信息流的信息对抗机制；此外，考虑到谣言特征识别的全面性，在本步中，生成对抗网络中的“生成器”模块将被特别加强，通过融合图卷积网络 GCN[23] 以及双向门循环网络 Bi-GRU，实现谣言信息中的传播结构特征和时序依赖特征的混合提取。考虑到生成对抗网络在模型构建思想上对于鉴别效果的抑制作用，直接使用第一步“鉴别器”的输出作为谣言推断的依据将会提高谣言误判的可能性，因此，本模型特别引入了“第二步”基于自注意机制 self-attention 的判别网络：此步中的网络一方面接受“第一步”中提取的混合特征，维持信息对抗机制；另一方面，对“鉴别器”做进一步的加强处理，从而在保证噪声容抗性的同时，从整体上提高谣言鉴别的准确性。模型的整体结构如图 1 所示。

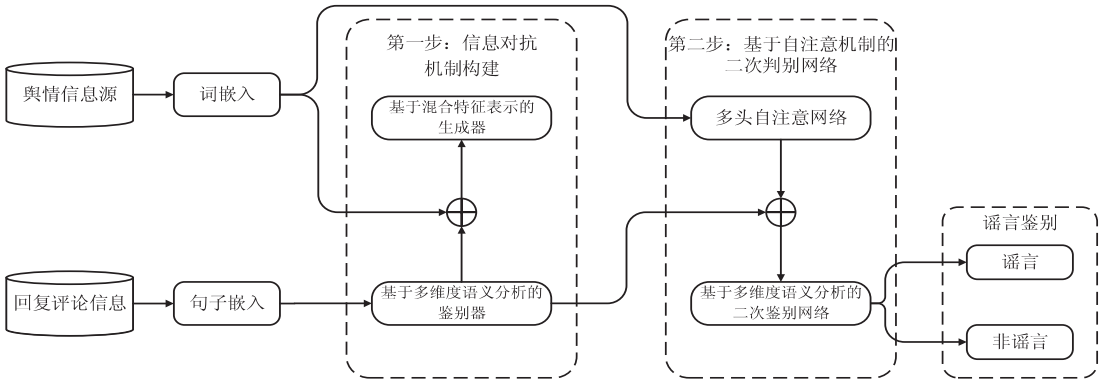

图 1 模型整体框架图

2.1 信息对抗机制的构建

本模块对应于本研究提出的“双步”模型的第一步。在本模块中，我们将结合 WGAN 和 ACGAN，提出适用于谣言检测任务的有监督学习背景下的信息对抗机制：利用加强的具备混合特征提取的生成器，产生对抗性的虚拟声音参与训练，从而提升模型对“噪声”的容抗性。需要强调的是，区别于其他基于 GAN 的模型，本研究提出的方法不再使用随机数据产生对抗性样本，而是使用真实舆情事件的回复评论数据。图 2 是基于一则从“医保缴费信息”引发的新浪微博舆情事件中提取的抽象化的传播网络示意图（截取部分信息）。图 2 中所有信息相对于时间轴的先后次序蕴含

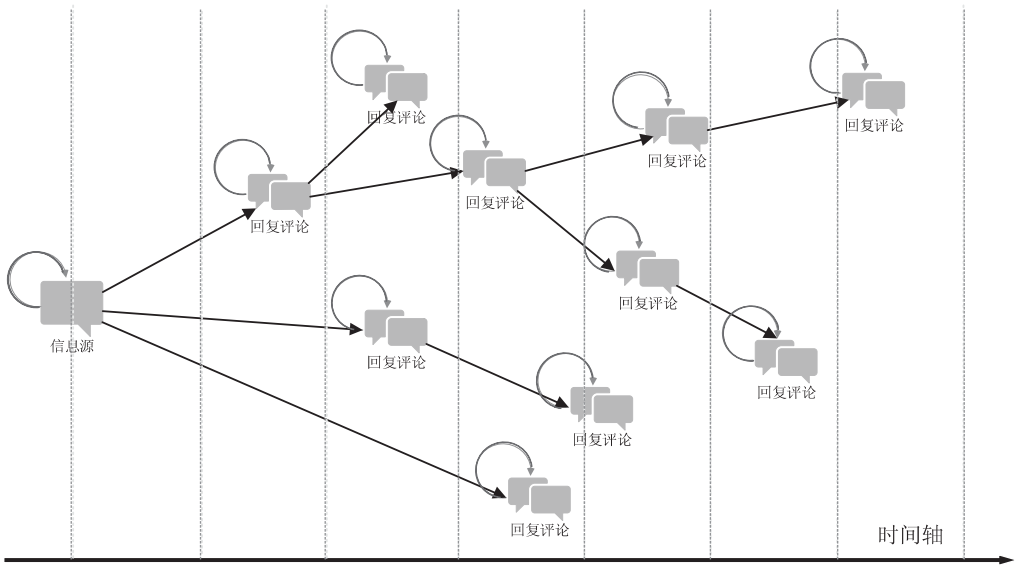

图 2 抽象化的传播网络示意图

了舆情传播网络的时序特征;同时,回复评论信息相对于信息源的不同位置构成了多层级的拓扑网络。图中在各个信息节点上人为加入了自连接环,目的在于后续结构特征的分析及提取。本模块将依据此抽象化的传播网络图提取舆情信息的结构特征以及时序特征,并加以融合。

2.1.1　基于混合特征表示的生成器

在本部分中,我们将利用回复评论数据生成对抗性的虚拟声音,相关网络结构如图 3 所示。为了表述的简洁性和清晰性,下述的回复评论数据均代指已经利用句子嵌入 Sentence Embedding 方法映射到句子表示空间的高维向量。

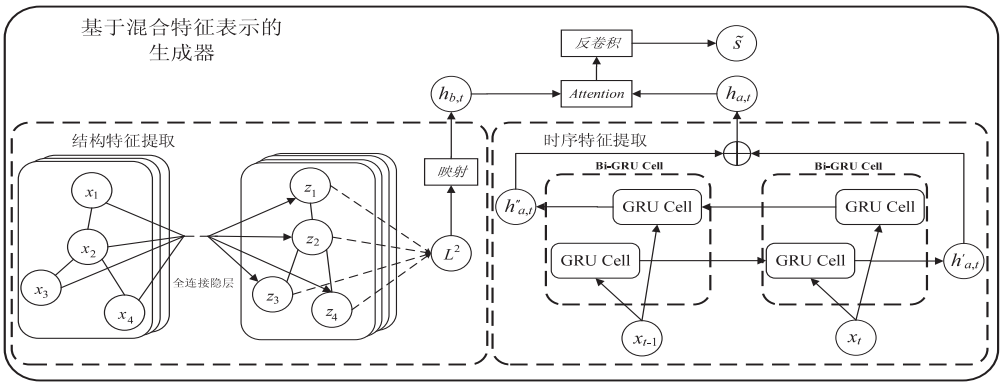

图 3　基于混合特征表示的生成器结构图

a.舆情信息结构特征提取:依据舆情事件相关的回复评论数据建立传播网络,网络中的节点为回复评论信息,连边根据回复评论之间的层级关系建立。再在生成的传播网络的节点上加入自连接环,得到最终关于回复评论信息的关系网 G 。

依据建立的信息关系网 G 生成邻接矩阵 A 和节点的度矩阵 D 。为了提取直接相邻节点之间的结构信息,我们需要采用单层的图卷积操作

$$L^1 = \delta(\hat{A}XW^1) \tag{1}$$

其中 $\hat{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, X 为待分析的回复评论数据矩阵, W^1 为本层的可学习权重, L^1 为图卷积层的输出, $\delta(\cdot)$ 为激励函数,通常选用 ReLU 函数,其形为 $\delta(x) = \max(0,x)$ 。考虑到现实情况回复评论数据的复杂性,仅仅提取直接相邻节点之间的结构特征不足以反映出整个网络的结构特点,因此,需要再加入更深一层的图卷积操作

$$L^2 = \delta(\hat{A}L^1W^2) \tag{2}$$

其中 W^2 为第二层的可学习权重, L^2 即为最终提取的舆情信息结构特征。

b.舆情信息时序特征提取:考虑到时序特征提取的效果以及网络结构的简单性,我们选择利用门循环网络 GRU 来分析回复评论数据之间的时间依赖关系。

令 x_t 表示一个回复评论数据流在 t 时刻对应的信息,那么,GRU 网络中相关变量的更新将遵循以下公式

$$z_{x_t} = \sigma(x_tU^z + h'_{a,t-1}W^z)$$
$$r_{x_t} = \sigma(x_tU^r + h'_{a,t-1}W^r)$$
$$\tilde{h}_{x_t} = \tanh(x_tU^{\tilde{h}} + (h'_{a,t-1} \otimes r_{x_t})W^{\tilde{h}})$$
$$h'_{a,t} = (1 - z_{x_t}) \otimes h'_{a,t-1} + z_{x_t} \otimes \tilde{h}_{x_t} \tag{3}$$

其中 U^z , W^z , U^r , W^r , $U^{\tilde{h}}$ 和 $W^{\tilde{h}}$ 为 GRU 网络中可自学习的权重参数, $\otimes$ 为逐元素相乘操作符, $h'_{a,t}$ 和 $h'_{a,t-1}$ 分别代表当前以及先前状态下的提取出的正向时序特征。 $\tilde{h}_{x_t}$ 为 $h'_{a,t}$ 的备选特征表示, r_{x_t} 和 z_{x_t} 分别代表重置门和更新门的状态。 $\sigma(\cdot)$ 为 sigmoid 函数,用来将 r_{x_t} 和 z_{x_t} 的值限制到 0~1 之间。

上述操作可以提取正向舆情信息流的时序特征,然而,在谣言检测的实际操作中仅仅依赖正向的时序特征往往是不够的。因此,我们进一步加入了反向舆情信息流的时序特征提取操作。通过堆叠正反向的 GRU 网络,最终可以得到双向的深度 Bi-GRU 网络,连接其输出的正向及反向时序特征,即可得到最终在整体上的时序特征表示 $h_{a,t}$ 。

c.结构特征和时序特征融合:采用公式(3)所述的更新规则,将结构特征 L^2 映射为 $h_{b,t}$,其维度和 $h_{a,t}$ 完全一致。为了实现融合结构特征和时序特征时权重的自动调整,我们利用 Attention 机制,其 Attention 权重 α 通过以下公式计算

$$\alpha = \text{softmax}\left(\frac{\gamma \cdot \tanh(hW^h)}{\sum_{h \in \{h_{a,t}, h_{b,t}\}} \gamma \cdot \tanh(hW^h)}\right) \tag{4}$$

其中, γ 和 W^h 是注意力网络中待学习的参数。那么得到的融合特征为

$$h_t = \sum_{h \in \{h_{a,t}, h_{b,t}\}} \alpha h \tag{5}$$

为了产生类似舆情信息源的对抗性声音,我们需要利用反卷积操作,将句子空间表示的 h_t 映射到词表示空间内

$$\tilde{s} = h_{pad} * f_{g,k} \tag{6}$$

其中，$*$ 代表卷积操作符，h_{pad} 为对 h_t 进行 zero padding 操作后的矩阵，$f_{g,k}$ 是 k 输出通道的卷积核，$\tilde{s}$ 即为生成器产生的在词空间表示的对抗性虚拟声音。

2.1.2 基于多维度语义分析的鉴别器

在本部分中，我们将对舆情信息源以及生成的对抗性虚拟声音进行鉴别，相关网络结构如图 4 所示。借助于 word2vec 方法将舆情信息源中的每个词映射到高维词空间，并使用符号 s 指代词嵌入 word embedding 后的舆情信息源。为了识别谣言内具有区别性的特异特征，我们构建了基于卷积神经网络 CNN 的鉴别器，从而对信息源 s 和生成信息 $\tilde{s}$ 进行细粒化的分析。CNN 网络包含 3 个平行放置的隐层，其卷积核的维度分别设置为[2，词嵌入维度]，[3，词嵌入维度]和[4，词嵌入维度]，目的是为了分析双词、三词和四词三个维度上的语义特征。

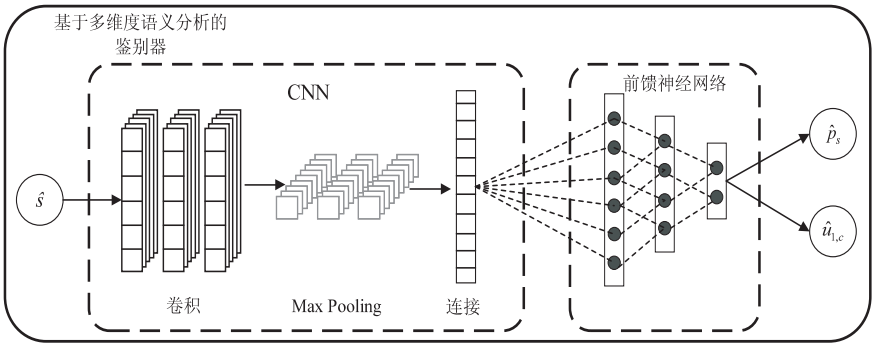

图 4 基于多维度语义分析的鉴别器结构图

对于三个卷积层，输出通道数被固定为相同的值，其卷积过程通过以下公式实现：

$$v^j = \text{Maxpooling}(\hat{s} * f^j_{d,k} + b^j_{d,k}) \tag{7}$$

其中，$j \in \{1,2,3\}$ 对应于三个卷积层，$f^j_{d,k}$ 为 j 层的具有 k 输出通道的卷积核，$b^j_{d,k}$ 为 j 层的偏置，$*$ 是卷积操作符，$\hat{s} \in \{s,\tilde{s}\}$ 为卷积网络的输入，v^j 代表 j 层的输出。为了使 CNN 网络学习到最显著的区别性特异特征，我们在卷积的基础上采用了 max-pooling 操作，连接三层的输出，得到

$$v = \text{Concatenate}(v^1, v^2, v^3) \tag{8}$$

再将其输入到一个多层的前馈神经网络中：

$$\begin{aligned} &l^j = \text{ReLU}(W^j l^{j-1} + b^j), \forall j \in [q] \\ &\hat{u}_{1,c} = W_c l^q + b_c \\ &\hat{p}_s = \sigma(W_s l^q + b_s) \end{aligned} \tag{9}$$

其中，l^j 为 j 层（$l^0 = v$）的输出，W^j 和 b^j 是 j 层的可学习权重和偏置，W_c，b_c，W_s 和 b_s 是输出层的可学习权重及偏置，q 代表隐层的数目。$\sigma(\cdot)$ 为 sigmoid 函数，$\hat{p}_s \in \{p_s, \tilde{p}_s\}$ 可以被视为当前输入被鉴别为信息源 s 而不是生成信息 $\tilde{s}$ 的概率，p_s 和 $\tilde{p}_s$ 分别对应于输入为 s 和 $\tilde{s}$ 的概率输出。$\hat{u}_{1,c} \in \{u_{1,c}, \tilde{u}_{1,c}\}$ 用于指代鉴别后输出的对于谣言分类的判定数值，$u_{1,c}$ 和 $\tilde{u}_{1,c}$ 分别对应于输入为 s 和 $\tilde{s}$ 的输出。这样，$y_{1,c} = \text{Softmax}(u_{1,c})$ 和 $\tilde{y}_{1,c} = \text{Softmax}(\tilde{u}_{1,c})$ 就可以表示第一步输出的谣言分类概率。

2.1.3 “第一步”的优化目标

鉴别器的输出包括两类，即判定输入为信息源而不是生成信息的概率，以及舆情信息的谣言分类概率。令 L_s 和 L_c 分别表示上述两类概率输出的损失函数，其形式为

$$\begin{aligned} L_c &= -\frac{1}{N_t}\sum_{j=1}^{N_t}\sum_{\{N,NR\}} y^j \log(y^j_{1,c}) - \frac{1}{N_t}\sum_{j=1}^{N_t}\sum_{\{N,NR\}} y^j \log(\tilde{y}^j_{1,c}) \\ L_s &= -p_s + \tilde{p}_s \end{aligned} \tag{10}$$

其中 N_t 为训练样本数，y 为训练样本的真实分类。为了提高训练的稳定性，我们借鉴了 WGAN 中 Lipschitz 约束的使用，在随机样本的梯度范数上添加了一个正则化项

$$GP = \lambda E[(\| \nabla_{\tilde{x}} D(\tilde{x}) \|_2 - 1)^2)] \tag{11}$$

其中，GP 为随机样本加权后的期望值，λ 为正则化项的相对权重调节参数，$\tilde{x}$ 是从真实信息分布和生成器学习的分布之间的数据对中提取的样本，$D(\cdot)$ 为判定输入为信息源而不是生成信息的概率输出。

那么，对于生成器其优化目标为最小化 $L_c - L_s$，对于鉴别器其优化目标为最小化 $L_c + L_s + GP$。在此优化目标的作用下，生成器将持续向迷惑鉴别器的方向优化，同时，鉴别器也将在保证谣言鉴别准确率的前提下，不断深挖非显性的谣言特异特征。

2.2 基于自注意机制的二次判别网络

本模块对应于本研究提出的“双步”模型的第二步，对于舆情信息谣言分类的判定将在本模块中最终给出。为了维持信息对抗机制，在“第一步”模块中习得的生成器及其相关权重参数将被传递到本模块中，从而产生对抗性的虚拟声音。类似于“第一步”模块，本模块接收的输入仍然包括两部分，即舆情信息源 s 和对抗性虚拟声音 $\tilde{s}$。具体网络结构如图 5 所示。

2.2.1 混合自注意机制的判别网络

自注意机制的引入目的是为了进一步提取舆情信息源 s 中隐藏的高维特征信息。对于一个具有 h 个 Head 的多头注意力（Multi-head Self-attention）网络，使用 Q^j，K^j 和 V^j 分别代表第 j 个 Head 的 Query，Key 和 Value 矩阵，他们的计算方法为

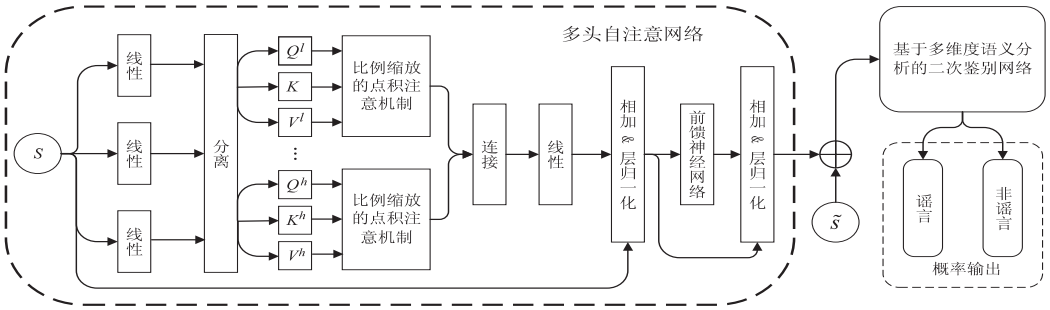

图 5　基于自注意机制的二次判别网络结构图

$$Q^j = sW_Q^j$$
$$K^j = sW_K^j \tag{12}$$
$$V^j = sW_V^j$$

其中 W_Q^j , W_K^j 和 W_V^j 分别是 Query,Key 和 Value 的线性 Linear 转换权重矩阵。那么,第 j 个 Head 输出的特征为

$$Z^j = \text{Attention}(Q^j,K^j,V^j) = \text{Softmax}(\frac{Q^jK^{jT}}{\sqrt{d}})V^j \tag{13}$$

其中, d 是比例缩放系数。连接所有 Head 输出的特征,可以得到整体的多头注意输出 Z'

$$Z' = \text{Concatenate}(Z^1,Z^2,\cdots,Z^j)W_Z \tag{14}$$

其中, W_Z 是自学习权重。为了防止原始输入信息过多的特征丢失,我们在自注意网络中使用了残差连接,进一步使用 layer normalization 后,输出变为

$$Z = \text{Layernorm}(s + Z') \tag{15}$$

参照 Transformer[27] 中的方法,我们将多头注意力网络的输出同一个全连接的前馈网络相连,同时保留残差连接和 layer normalization,得到的输出为

$$O_{att} = \text{Layernorm}(Z + ZW_f + b_f) \tag{16}$$

其中, W_f 和 b_f 分别是前馈网络的权重及偏置矩阵, O_{att} 即为从舆情信息源 s 提取的多维度特征矩阵。

将多维度特征矩阵 O_{att} 输入到一个基于“第一步”信息对抗模块中的“鉴别器”搭建的,并具有相同结构的二次鉴别网络中,得到对谣言分类的最终判定。区别于“第一步”中的鉴别器,此处的二次鉴别网络将接收多维度特征矩阵 O_{att} 以及对抗性虚拟声音 $\tilde{s}$ 作为输入,同时其中的前馈神经网络也将只输出对于谣言分类的判定 $\hat{u}_{2,c} \in \{u_{2,c},\tilde{u}_{2,c}\}$ ($u_{2,c}$ 和 $\tilde{u}_{2,c}$ 分别对应于输入为 O_{att} 和 $\tilde{s}$ 的输出)。

2.2.2　“第二步”的优化目标

利用“第二步”模块输出的谣言分类的判定值 $u_{2,c}$ 和 $\tilde{u}_{2,c}$,我们可以推断谣言信息的类别。最终输出的谣言分类概率的分布函数可以通过一个加权后的映射得到

$$\hat{y} = \text{Softmax}(u_{2,c} + \mu\tilde{u}_{2,c}) \tag{17}$$

其中, μ 是一个超参数,用于调整从原始信息源和生成的对抗性声音中习得知识的权重。

我们使用交叉熵作为“第二步”模块的优化目标,其包括两部分:对于原始信息源的损失评估以及对于生成的对抗性声音的损失评估

$$L = -\frac{1}{N_t}\sum_{j=1}^{N_t}\sum_{\{N,NR\}} y^j\log(y_{2,c}^j) - \mu\frac{1}{N_t}\sum_{j=1}^{N_t}\sum_{\{N,NR\}} y^j\log(\tilde{y}_{2,c}^j) \tag{18}$$

其中, y 是舆情的真实分类, $y_{2,c} = \text{Softmax}(u_{2,c})$, $\tilde{y}_{2,c} = \text{Softmax}(\tilde{u}_{2,c})$ 。

3　实验与分析

为了验证本研究提出的模型(简称 IHCR),在此部分中我们将基于公共数据集展开对比实验,证实本模型的谣言检测效果,检验在有噪声数据影响时,本模型对于干扰信息的容抗性。

3.1　实验数据

对比实验将利用在谣言鉴别领域被广泛使用的两个数据集展开,即 PHEMEv5[25] 和新浪微博[26] 数据集。PHEMEv5 数据集是在 Twitter 平台上爬取的关于 5 类话题事件发表的 5 802 条相关信息及其后续评论,该数据集信息的承载语言为英文,采集时间为 2016 年。新浪微博数据集是从新浪微博平台获取的包含多类话题事件的 4 664 条相关信息及其后续评论,该数据集信息的承载语言为中文,采集时间为 2016 年。两个数据集中所有的信息都别标记为“谣言”和“非谣言”两者中的一类。两个数据集具体的细节信息在表 1 中给出,选取这两个数据集的目的是为了验证我们提出的模型跨平台及跨语言的适用性。

表 1　PHEMEv5 和新浪微博数据集统计信息

数据集	PHEMEv5	新浪微博
用户数	49 345	2 746 818
帖文数	103 212	3 805 656
舆情事件数	5 802	4 664
谣言数	1 972	2 313
非谣言数	3 830	2 351

续表 1 PHEMEv5 和新浪微博数据集统计信息

数据集	PHEMEv5	新浪微博
平均事件持续时长/h	1	2 461
舆情事件的平均帖文数	18	816
舆情事件的最长贴文长度	228	59 318
舆情事件的最短贴文长度	3	10

3.2 评估指标及主要参数设置

为了全面地体现谣言检查效果,同时方便横向的模型比较,本研究选取了国内、外谣言检测学术界常用并被普遍认可的四个评估指标,即准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 得分四个指标,其计算方法为

$$Accuracy = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FP_c}$$

$$Precision = \frac{TP_c}{TP_c + FP_c} \tag{19}$$

$$Recall = \frac{TP_c}{TP_c + FN_c}$$

$$F_1 = \frac{2Precision \cdot Recall}{Precision + Recall}$$

其中,TP_c,FP_c 和 FN_c 分别表示正确预测的正例个数,错误预测的正例个数和错误预测的负例个数。

我们使用 PyTorch 来实现我们的模型。对于模型中相关参数的选择,本研究进行了大量的对比实验,比较了不同参数设置下谣言检测的效果,并找出了检测效果最优的一组作为最终的参数组合,具体为:词嵌入和句子嵌入维度设定为 300,“第一步”模块生成器中的 Bi-GRU 的隐层数为 2 并使用 Dropout,鉴别器中的 CNN 的输出通道数为 128,后续连接的前馈神经网络的隐层数为 4。使用 He initialization 方法来初始化“第一步”信息对抗模块中的可学习权重,并选用 Adam 算法来优化损失函数,其超参数为 $\beta_1 = 0.5$,$\beta_2 = 0.9$。在“第二步”二次鉴别模块中,设置自注意的多头数为 3,后续全连接的前馈网络的输出维度为 256,仍然选用 Adam 算法来优化第二步的损失函数,此处的超参数为 $\beta_1 = 0.9$,$\beta_2 = 0.999$。两步中的学习率都设置为 1e-3,并随着学习过程逐渐下降。实验中各结果通过五折交叉验证得到。

3.3 对比实验选取的参照模型

为了验证本模型的谣言鉴别效果,我们从现有研究中选取了 9 种具有代表性的模型来进行对比,包括一种传统的基于决策树构建的模型(DT-Rank[27])、一种基于树形传播网络构建的模型(BU-RvNN[14])、两种基于 RNN 的模型(DA-RNN[28] 和 GRU-R[26])、两种基于 CNN 的模型(Text-CNN[29] 和 TDRD[30])、一种基于混合特征提取的模型(GGNN[31])和两种基于 GAN 的模型(GAN-GRU[17] 和 RG-GAN[19]):

①DT-Rank:一种通过分析包含有争议事实的信息来对趋势消息进行分类的决策树模型。

②BU-RvNN:一种基于递归神经网络构建的,通过回溯信息传播路径推断谣言真实性的模型。

③DA-RNN:一种基于循环神经网络的方法,该方法通过识别潜在的时间敏感表征来捕获谣言信息特有的上下文变化。

④GRU-R:一种利用深度堆叠 GRU 单元构建的,通过在不同时间间隔内聚合判别性特征来鉴别谣言的模型。

⑤Text-CNN:一种基于卷积神经网络的文本分类器,其利用微调技术来学习文本分类任务中的指向性特征向量。

⑥TDRD:一种基于主题的 CNN 模型,其预测的主题特征被合并到源信息中,从而辅助谣言的检测。

⑦GGNN:一种混合时序和结构特征的谣言检测方法,该方法将 GRU 和 GCN 相结合,分析并融合了信息流和层次节点之间的非显性关系。

⑧GAN-GRU:一种基于 GAN 的谣言检测方法,该方法使用对抗性学习策略施压分类器,以获得更强的谣言指示性表示。

⑨RG-GAN:一种结合了 GAN 和强化学习的模型,通过有选择的在信息中插入生成的词向量来提高模型对于噪声信息的容抗性。

3.4 谣言检测结果及分析

表 2 中呈现的是本研究提出的 IHCR 模型和 9 种对比实验方法在谣言检测任务上的检测效果。每种检测指标的最优值我们使用粗体字来标示。从表 2 中可以看出,本研究提出的方法在两种典型的中文及英文语境中均有最好的谣言检测准确率,分别是 88.5%(PHEMEv5 数据集)和 95.7%(新浪微博数据集),相较于对比方法中在谣言检测准确率指标上最好的 GGNN 分别提升了 3.1% 和 4.1%。观察在“谣言”和“非谣言”分类的下 F1 得分,本研究提出的方法在两个数据集上相较于所有对比方法均得到了最高的得分,体现出 IHCR 有能力在谣言的各类别上挖掘并识别出区别于它类的指示性信息特征。

分析表 2 可知,相较于基于数据驱动的模型,基于特征的 DT-Rank 的检测效果明显要略逊一筹,在各个指标上的结果几乎都是最低的,反映出基于特征的方法在跨数据集泛化能力上的不足。基于 RNN 的方法目的在于提取时序特征,而基于 CNN 的方法分析的重点落在了信息结构特征的挖掘,从表 2 中可以看到,其识别效果是不及基于混合特征提取的模型 GGNN 的,而这也是本研究将混合特征表示融入到信息对抗的出

表 2　IHCR 和对比模型在 PHEMEv5 和新浪微博数据集上的谣言检测效果

模型	类别	准确率		精确率		召回率		F1 得分	
		PHEMEv5	微博	PHEMEv5	微博	PHEMEv5	微博	PHEMEv5	微博
DT-Rank	谣言	0.592	0.732	0.382	0.738	0.319	0.715	0.347	0.726
	非谣言			0.677	0.762	0.733	0.749	0.703	0.737
DA-RNN	谣言	0.774	0.788	0.638	0.819	0.814	0.746	0.715	0.780
	非谣言			0.884	0.762	0.753	0.831	0.813	0.795
GAN-GRU	谣言	0.783	0.765	0.761	0.746	0.825	0.781	0.792	0.762
	非谣言			0.809	0.786	0.741	0.750	0.773	0.767
BU-RvNN	谣言	0.789	0.857	0.782	0.842	0.793	0.881	0.788	0.861
	非谣言			0.795	0.873	0.784	0.831	0.790	0.851
TDRD	谣言	0.814	0.867	0.778	0.870	0.634	0.860	0.679	0.865
	非谣言			0.828	0.864	0.907	0.874	0.865	0.869
Text-CNN	谣言	0.839	0.730	0.757	0.731	0.789	0.739	0.773	0.731
	非谣言			0.885	0.739	0.865	0.723	0.875	0.726
RG-GAN	谣言	0.843	0.912	0.734	0.901	0.763	0.928	0.748	0.914
	非谣言			0.894	0.924	0.879	0.896	0.886	0.910
GRU-R	谣言	0.852	0.910	0.773	0.876	0.784	0.956	0.779	0.914
	非谣言			0.892	0.952	0.886	0.864	0.889	0.906
GGNN	谣言	0.854	0.916	0.771	0.896	0.740	0.945	0.755	0.920
	非谣言			0.888	0.940	0.903	0.887	0.896	0.913
IHCR	谣言	**0.885**	**0.957**	**0.793**	**0.946**	**0.842**	**0.970**	**0.816**	**0.958**
	非谣言			**0.928**	**0.969**	0.903	**0.944**	**0.916**	**0.956**

发点。GAN-GRU 和 RG-GAN 均是基于对抗生成网络构建,虽然提高了网络对噪声数据的容抗性,但其生成的对抗数据会影响鉴别器的鉴别效果,拉低了谣言分类的准确率,GAN-GRU 的准确率仅有 78.3%和 76.5%,而这也就是本研究构建二次判别网络的动因。

3.5　消融实验分析

为了验证本研究提出模型中各个组成模块存在的必要性,在本部分中我们设计了一系列消融实验,分析在特定组成部分缺失的情况下模型对于谣言检测的效果。各变体模型详述如下:

①w/o GCN:移除“第一步”中生成器中的结构特征提取网络,生成器仅提取时序特征。

②w/o GRU:移除“第一步”中生成器中的时序特征提取网络,生成器仅提取结构特征。

③Random-G:使用随机初始化而不是“第一步”学习得到的生成器作为“第二步”二次判别网络的输入。

④w/o ATT:移除“第二步”中的自注意力网络,仅使用后续连接的和“第一步”鉴别器同构的网络进行谣言判别。

⑤w/o GAN:完全移除“第一步”信息对抗模块,仅使用二次判别网络进行谣言检测。

⑥w/o SEC:完全移除“第二步”二次判别网络,仅使用信息对抗模块进行谣言检测。

表 3　消融实验结果

模型	类别	准确率		精确率		召回率		F1 得分	
		PHEMEv5	微博	PHEMEv5	微博	PHEMEv5	微博	PHEMEv5	微博
w/o GCN	谣言	0.835	0.942	0.717	0.934	0.757	0.953	0.736	0.943
	非谣言			0.891	0.951	0.869	0.931	0.880	0.941
w/o GRU	谣言	0.816	0.906	0.686	0.890	0.729	0.928	0.707	0.909
	非谣言			0.878	0.923	0.854	0.883	0.866	0.903
Random-G	谣言	0.807	0.790	0.668	0.797	0.729	0.784	0.697	0.791
	非谣言			0.876	0.783	0.842	0.797	0.859	0.790
w/o ATT	谣言	0.843	0.944	0.734	0.938	0.763	0.953	0.748	0.945
	非谣言			0.894	0.952	0.879	0.935	0.886	0.943
w/o GAN	谣言	0.852	0.859	0.751	0.854	0.768	0.869	0.760	0.861
	非谣言			0.898	0.863	0.889	0.848	0.893	0.856
w/o SEC	谣言	0.861	0.878	0.773	0.874	0.768	0.886	0.771	0.880
	非谣言			0.899	0.882	0.901	0.870	0.900	0.876
IHCR	谣言	**0.885**	**0.957**	**0.793**	**0.946**	**0.842**	**0.970**	**0.816**	**0.958**
	非谣言			**0.928**	**0.969**	**0.903**	**0.944**	**0.916**	**0.956**

消融实验的谣言鉴别效果如表 3 所示。可以看出,相较于本研究提出的模型完全体,各变体模型在谣

言识别的准确率和 F1 得分上均有了一定程度的降低，反映出所有组成模块都是必要的，任意一模块的缺失都会限制 IHCR 对谣言特征进行全面分析的能力。特别的，w/o GRU 下的谣言检测准确率要明显低于 w/o GCN，说明相较于结构特征，信息的时序依存性在谣言检测任务上具有更高的影响权重。Random-G 的谣言检测效果在两个数据集上都是最差的，随机初始化的生成器不仅没能提供任何谣言检测的线索，反而对模型产生了误导，这也就从侧面印证了“第二步”继承的从“第一步”习得的生成器的必要性以及正面的促进作用。w/o GAN 不尽如人意的谣言检测效果证实了单一的信息对抗机制对于鉴别器输出的抑制作用，而 w/o SEC 构型下较低的准确率则反映出了信息对抗的必要性。

3.6 噪声容抗性分析

在本部分的实验中，为了分析 IHCR 对噪声信息的容抗性，我们在 PHEMEv5 和新浪微博数据集的回复评论信息中人为插入了一定比例的噪声数据，从而研究在噪声信息影响情形下 IHCR 对谣言的检测效果。我们设计了两种模型训练策略进行比照：策略一是利用混合有噪声的数据集训练整个模型；策略二是利用原始数据训练“第一步”信息对抗模块，再使用混合有噪声的数据集训练“第二步”二次鉴别模块。所有的参数及初始化方法保持和原始设定一致。表 4 和表 5 中呈现的是两类训练策略下的谣言检测结果。在策略一下，随着噪声占比的提升，谣言检测效果并没有呈现出单调的下降趋势，谣言检测准确率在整体上比较稳定，这就说明我们提出的模型对于噪声信息并不敏感，具有较好的噪声容抗性。在策略二下，谣言检测效果随着噪声占比的提升而逐渐下降，但是下降的速率并不显著，噪声占比从 10% 提升到 50%，谣言检测准确率仅下降了 2.1%（PHEMEv5）和 4.1%（新浪微博）。相较于大部分对比实验中的模型，即使在训练策略二下，IHCR 仍是具有竞争力的，这就进一步反映出我们提出的信息对抗背景下混合特征表示的有效性。

表 4 按策略一训练时不同噪声占比情况下的谣言检测效果

噪声占比/%	类别	准确率		精确率		召回率		F1 得分	
		PHEMEv5	微博	PHEMEv5	微博	PHEMEv5	微博	PHEMEv5	微博
10	谣言	0.857	0.944	0.750	0.930	0.797	0.962	0.773	0.946
	非谣言			0.908	0.960	0.884	0.926	0.896	0.943
20	谣言	0.850	0.938	0.737	0.919	0.791	0.962	0.763	0.940
	非谣言			0.905	0.959	0.876	0.913	0.891	0.936
30	谣言	0.847	0.940	0.732	0.916	0.785	0.970	0.757	0.942
	非谣言			0.903	0.968	0.874	0.909	0.888	0.938
40	谣言	0.852	0.936	0.754	0.912	0.763	0.966	0.758	0.938
	非谣言			0.896	0.963	0.891	0.905	0.893	0.933
50	谣言	0.845	0.938	0.728	0.912	0.785	0.970	0.755	0.940
	非谣言			0.903	0.968	0.871	0.905	0.887	0.935

表 5 按策略二训练时不同噪声占比情况下的谣言检测效果

噪声占比/%	类别	准确率		精确率		召回率		F1 得分	
		PHEMEv5	微博	PHEMEv5	微博	PHEMEv5	微博	PHEMEv5	微博
10	谣言	0.852	0.908	0.741	0.891	0.791	0.932	0.765	0.911
	非谣言			0.906	0.927	0.879	0.883	0.892	0.905
20	谣言	0.845	0.891	0.728	0.878	0.785	0.911	0.755	0.894
	非谣言			0.903	0.905	0.871	0.870	0.887	0.887
30	谣言	0.840	0.880	0.721	0.833	0.774	0.953	0.747	0.889
	非谣言			0.898	0.944	0.869	0.805	0.883	0.869
40	谣言	0.836	0.872	0.759	0.846	0.678	0.911	0.716	0.878
	非谣言			0.865	0.901	0.906	0.831	0.885	0.865
50	谣言	0.831	0.867	0.739	0.845	0.689	0.903	0.713	0.873
	非谣言			0.868	0.893	0.894	0.831	0.880	0.861

4 结 语

本研究提出了一种融合信息对抗及混合特征表示的谣言检测模型。区别于现有的直接移植非监督性学习中的对抗生成网络 GAN 来搭建有监督的谣言检测模型，本研究提出了一种具有部分参数共享的“两步走”模型，从而克服了对抗机制对有监督学习造成的负面影响，实现了在提升检测效果的同时，增强模型对于噪声容抗性的目的。此外，为了进一步赋能信息对抗模块中生成器对于特征的表达，本研究搭建了有机混合时序特征及结构特征的深度网络，实现了对舆情信息在多种维度上的特征呈现。借助于真实的社交网上的舆情信息，本研究比较并分析了提出模型的谣言检测效果，结果表明，在中文及英文语言环境下，本研

究提出的模型均有能力在保证低噪声敏感性的同时提升谣言检测的准确率。

参 考 文 献

[1] 王友卫,童　爽,凤丽洲,等.基于图卷积网络的归纳式微博谣言检测新方法[J].浙江大学学报(工学版),2022,56(5):956-966.

[2] 胡　斗,卫玲蔚,周　薇,等.一种基于多关系传播树的谣言检测方法[J].计算机研究与发展,2021,58(7):1395-1411.

[3] Kwon S,Cha M,Jung K.Rumor detection over varying time windows[J].PloS One,2017,12(1): e0168344.

[4] Liu X,Nourbakhsh A,Li Q,et al.Real-time rumor debunking on twitter[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management,2015: 1867-1870.

[5] Luo Y,Ma J,Yeo C K.Exploiting user network topology and comment semantic for accurate rumour stance recognition on social media[J].Journal of Information Science,2022,48(5): 660-675.

[6] Gupta A,Kumaraguru P,Castillo C,et al.Tweetcred: Real-time credibility assessment of content on twitter[C]//Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain,November 11-13,2014.Proceedings 6.Springer International Publishing,2014: 228-243.

[7] Popat K.Assessing the credibility of claims on the web[C]//Proceedings of the 26th International Conference on World Wide Web Companion,2017: 735-739.

[8] Yang F,Liu Y,Yu X,et al.Automatic detection of rumor on sina weibo[C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics,2012: 1-7.

[9] Sun S,Liu H,He J,et al.Detecting event rumors on sina weibo automatically[C]//Web Technologies and Applications: 15th Asia-Pacific Web Conference,APWeb 2013,Sydney,Australia,April 4-6,2013.Proceedings 15.Springer Berlin Heidelberg,2013: 120-131.

[10] 王　繁,郭军军,余正涛.融合评论的多任务联合谣言检测方法[J].计算机工程与科学,2022,44(9):1702-1710.

[11] Torshizi A S,Ghazikhani A.Automatic Twitter rumor detection based on LSTM classifier[C]//High-Performance Computing and Big Data Analysis: Second International Congress,TopHPC 2019,Tehran,Iran,April 23 - 25,2019.Springer International Publishing,2019: 291-300.

[12] Ma J,Luo Y.The classification of rumour standpoints in online social network based on combinatorial classifiers[J].Journal of Information Science,2020,46(2): 191-204.

[13] Alkhodair S A,Ding S H H,Fung B C M,et al.Detecting breaking news rumors of emerging topics in social media[J].Information Processing & Management,2020,57(2): 102018.

[14] Ma J,Gao W,Wong K F.Rumor detection on Twitter with tree-structured recursive neural networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,2018: 1980-1989.

[15] Ma J,Gao W,Joty S,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(4): 1-28.

[16] 刘　勘,杜好宸.基于深度迁移网络的 Twitter 谣言检测研究[J].数据分析与知识发现,2019,3(10):47-55.

[17] Ma J,Gao W,Wong K F.Detect rumors on twitter by promoting information campaigns with generative adversarial learning[C]//The World Wide Web Conference,2019: 3049-3055.

[18] 孟佳娜,王晓培,李　婷,等.基于对抗神经网络的跨模态谣言检测[J].数据分析与知识发现,2022,6(12):32-42.

[19] 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): 5861.

[20] Arjovsky M,Chintala S,Bottou L.Wasserstein generative adversarial networks[C]//International Conference on Machine Learning.PMLR,2017: 214-223.

[21] Gulrajani I,Ahmed F,Arjovsky M,et al.Improved training of wasserstein GANs[J].arXiv e-prints,2017: arXiv: 1704.00028.

[22] Odena A,Olah C,Shlens J.Conditional image synthesis with auxiliary classifier gans[C]//International Conference on Machine Learning.PMLR,2017: 2642-2651.

[23] Kipf T N,Welling M.Semi-supervised classification with graph convolutional networks[J].arXiv preprint,2016: arXiv: 1609.02907.

[24] Vaswani A,Shazeer N,Parmar N,et al.Attention is all you need[J].Advances in Neural Information Processing Systems,2017: 30.

[25] Zubiaga A,Liakata M,Procter R.Learning reporting dynamics during breaking news for rumour detection in social media[J].arXiv preprint,2016: arXiv:1610.07363.

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

[27] Zhao Z,Resnick P,Mei Q.Enquiring minds: Early detection of rumors in social media from enquiry posts[C]//Proceedings of the 24th International Conference on World Wide Web,2015: 1395-1405.

[28] Chen T,Li X,Yin H,et al.Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[C]//Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops,BDASC,BDM,ML4Cyber,PAISI,DaMEMO,Melbourne,VIC,Australia,June 3,2018.Springer International Publishing,2018: 40-52.

[29] Kim Y.Convolutional neural networks for sentence classification[C]// Association for Computational Linguistics,Doha,Qatar,2014.

[30] Xu F,Sheng V S,Wang M.Near real-time topic-driven rumor detection in source microblogs[J].Knowledge-Based Systems,2020,207: 106391.

[31] Zhu H.Rumour detection based on deep hybrid structural and sequential representation networks[J].Journal of Information Science,2021: 01655515211056579.

(责编:王育英;校对:刘影梅)