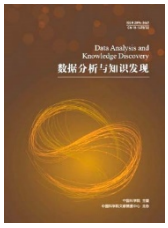


数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：一种融合知识图谱的图注意力神经网络谣言实时检测方法
作者：王根生，朱奕，李胜
网络首发日期：2023-06-07
引用格式：王根生，朱奕，李胜．一种融合知识图谱的图注意力神经网络谣言实时检测方法[J/OL]．数据分析与知识发现.
https://kns.cnki.net/kcms2/detail/10.1478.G2.20230606.1719.008.html

网络首发时间：2023-06-07 15:56:09
网络首发地址：https://kns.cnki.net/kcms2/detail/10.1478.G2.20230606.1719.008.html

一种融合知识图谱的图注意力神经网络谣言实时检测方法

王根生[1,2]，朱奕[3]，李胜[3]
[1](江西财经大学国际经贸学院　南昌　330013)
[2](江西财经大学人文学院　南昌　330013)
[3](江西财经大学财税与公共管理学院　南昌　330013)

摘要：
[目的]提高社交媒体中谣言实时检测的准确率，降低谣言传播危害。
[方法]提出一种融合知识图谱的图注意力神经网络谣言实时检测方法。首先，通过知识蒸馏从外部知识图谱中获取文本内容的背景知识；其次，通过点互信息把文本和背景知识转化为加权图结构表示，利用一种考虑边权重的图注意力神经网络从加权图中学习文本的非连续语义特征；然后，通过预训练语言模型 BERT 学习文本的连续语义特征，利用嵌入方法把用户和内容统计特征转化为连续向量表示；最后，把所有特征进行融合，输入全连接神经网络中进行谣言检测。
[结果]在两个公开的社交媒体谣言数据集 PHEME 和 WEIBO 上的实验结果表明，该方法的准确率分别达到了 92.1%和 84.0%，高于最先进的对比基线方法。
[局限]方法没有融合帖子中可能附加的图片或视频信息，不能进行多模态融合的谣言检测。
[结论]融合背景知识可以补充短文本的语义表示，融合用户和内容统计特征可以辅助文本语义特征做决策，提高模型的准确率。
关键词:谣言实时检测；图注意力神经网络；知识图谱；语义特征；统计特征；用户特征

中图分类号: TP393，G250　　**DOI:** 10.11925/infotech.2023-0405.

A Real-time Rumor Detection Method Based on Graph Attention Neural Network Integrated with Knowledge Graph

Wang Gensheng[1,2],　Zhu Yi[3],　Li Sheng[3]
[1](School of International Economics and Trade, Jiangxi University of Finance and Economics, Nanjing 330013, China)
[2] (School of Humanities, Jiangxi University of Finance and Economics, Nanchang 330013)
[3] (School of Finance, Taxation and Public Administration, Jiangxi University of Finance and Economics, Nanchang 330013)

Abstract:
[**Objective**] Improve the accuracy of real-time detection of rumors in social media and reduce the harm caused by rumors.
[**Methods**] A real-time rumor detection method based on graph attention neural network integrated with knowledge graph is proposed. First, obtaining the background knowledge of the text from the external knowledge graph through knowledge distillation. Second, transforming the text and background knowledge into a weighted graph structure representation through point mutual information, and a weighted graph attention neural network is used to learn the discontinuous semantic features of the text from the weighted graph. Then, learning the continuous semantic features of the text through the pre-trained language model BERT, and the statistical features of users and content are converted into continuous vector representations using the embedding method. Finally, all the features are fused and input into the fully connected neural network for rumor detection.
[**Results**] Experimental results on two public social media rumor datasets, PHEME and WEIBO, show that the method's accuracy reaches 92.1% and 84.0%, respectively, higher than the state-of-the-art baseline methods.
[**Limitations**] The method does not fuse the image or video information that may be attached to the

post and cannot perform multi-modal fusion rumor detection.
[**Conclusions**] Fusion of background knowledge can supplement the semantic representation of short texts. Fusing user and content statistical features can assist semantic features in making decisions and improve the model's accuracy.
Keywords: Rumor real-time detection; Graph attention neural network; Knowledge graph; Semantic features; Statistical features; User features

1 引言

随着互联网和移动客户端的快速发展，社交媒体成为人们获取和分享信息的重要平台。社交媒体方便人们信息交流的同时，也为谣言的传播提供了新的温床。谣言的传播不仅妨碍了人们对社交媒体的有效利用，而且可能造成民众的误解、引发负面情绪和扰乱社会秩序[1]。例如，在新冠疫情期初社交媒体上出现了许多“板蓝根+熏醋可预防新型肺炎”、“5G 网络导致新冠疫情传播”、“国家不再对新冠肺炎病人提供免费治疗”等谣言，导致民众的误解和恐慌。为了控制谣言的传播，相关机构建立了一些谣言查证平台，例如微博社区管理中心、中国互联网联合辟谣平台和美国的事实核查平台 Snopes。然而，这些平台主要依赖于人工验证的方式，不但需要耗费大量的人力和物力，而且存在较大的时间滞后问题。因此，研究网络谣言实时检测方法，对遏制谣言的传播具有重要的意义。

谣言检测主要有基于传播特征的方法、基于内容特征的方法和基于混合特征的方法。基于传播特征的方法，需要一定的谣言传播数据才能获得较好的检测效果，不能及时地识别谣言。基于内容特征的方法不依赖于谣言的传播数据，把谣言检测视为文本分类问题，可以实现谣言的实时检测。然后，有些谣言不具有典型的文本分类特征，导致单纯基于内容特征的方法不能取得较好的效果。基于混合特征的谣言检测方法融合传播特征、内容特征和用户特征，降低模型对传播特征的依赖，提高模型检测的时效性。人类对谣言进行实时判断时，不仅需要结合背景知识对谣言的内容进行分析，而且要结合发布人的信息进行判断。所以，本文提出一种融合内容特征、用户特征和背景知识的图注意力神经网络谣言实时检测方法（GAGK-RD），把离散化和碎片化的帖子文本转换成图结构形式，通过知识蒸馏从外部知识图谱中获取相关的背景知识，作为节点补充到图结构中。然后，利用图注意力神经网络对图结构进行处理，学习文本的深层语义特征。在两个公开的社交媒体谣言数据集 PHEME 和 WEIBO 上的实验结果表明，本文方法优于所有对比的实时谣言检测基线方法。

2 相关研究

2.1 基于传播特征的网络谣言检测

基于传播特征的谣言检测方法，主要利用谣言传播过程中的评论、转发和传播结构等信息进行谣言检测[2]。例如，Ma 等人[3]基于谣言的传播结构，提出一种传播树内核方法（Propagation Tree Kernel, PTK）来捕获谣言传播的高阶模式；文献[4]基于谣言传播过程中的评论时序数据，构建基于循环神经网络（Recurrent Neural Network, RNN）的谣言检测模型；文献[5]提出一种基于树结构递归神经网络的（Tree-structured Recursive Neural Networks, RvNN）谣言检测模型，同时学习谣言的传播结构特征和评论时序数据的语义特征。为了使模型更加关注于转发时序数据中具有谣言特征的部分，Chen 等人[6]提出一种注意力机制与循环神经网络相结合的谣言检测模型。胡斗等人[7]针对目前大部分基于传播特征的方法仅考虑了传播过程中的显式交互关系，忽略了对潜在关系的建模，提出一种基于多关系传播树的谣言检测方法。Song 等人[8]针对现有方法通常学习所有转发数据的语义表示，不能尽早识别谣言的问题，提出一种可信早期检测（Credible Early

Detection, CED）模型，在转发序列中寻找一个早期的时间点做出可靠的预测。随着图神经网络的发展，文献[9,10]提出一种基于图神经网络（Graph Neural Network, GNN）的谣言检测模型，通过 GNN 学习传播图中每个节点的表示。研究发现，基于传播特征的检测方法，需要谣言得到一定范围的传播后才能获得较好的检测效果，不能及时的识别谣言。

2.2 基于内容特征的网络谣言检测

基于内容特征的谣言检测方法不依赖于谣言的传播数据，把谣言检测视为文本分类问题，可以实现谣言的实时检测。目前主流的方式是通过深度学习方法挖掘谣言内容的深层语义特征。例如， Kaliyar 等人[11]提出一种基于卷积神经网络（Convolutional Neural Network, CNN）的谣言检测方法，通过 CNN 自动从文本中提取对谣言分类有帮助的特征；Ajao 等人[12]提出 CNN 和长短时记忆网络(Long-Short Term Memory, LSTM)相结合的模型，自动提取谣言的语义特征；刘赏等人[13]基于异构图网络（Heterogeneous Graph Network, HGN）获取新闻标题和正文差异性特征，实现虚假新闻检测； Alkhodair 等人[14]提出一种联合 Word2vec 和 LSTM 的谣言检测模型，识别新兴主题的突发新闻谣言； Cheng 等人[15]提出一种基于生成对抗网络的谣言检测，通过判别器和生成器的相互促进作用，强化谣言语义特征学习。为了提高检测的准确率，一些学者在语义特征的基础上，挖掘其它辅助特征，例如 Xu 等人[16]提出一种主题驱动的谣言检测框架 TDRD，利用 CNN 挖掘内容的主题信息，结合主题信息和文本词嵌入进行谣言判断。 Ma 等人[17]提出一种基于实体识别增强谣言文本语义理解的谣言检测方法，通过知识图谱得到实体解释，从而扩充原文内容，增强语义理解。研究发现，有些谣言不具有典型的文本分类特征，导致单纯基于内容特征的方法不能取得较好的效果。

2.3 基于混合特征的网络谣言检测

基于混合特征的谣言检测方法融合传播特征、内容特征和用户特征[18]，例如 Castillo 等人[19]首次针对推特上的谣言进行分析，提出基于传播特征、内容特征和用户特征的决策树谣言分类模型。Ma 等人[20]针对传统的谣言检测模型只利用了传播模式、文本内容和用户信息的静态统计特征，忽略了这些特征随时间的变化，提出一种基于动态序列时间结构（Dynamic Series-Time Structure , DSTS）的谣言检测模型；Ruchansky 等人[21]结合新闻的文本信息、用户的反馈和源作者信息进行虚假新闻检测，设计了三个模块： Capture 模块基于 RNN 提取原文和用户反馈信息，Score 模块根据用户的历史信息对用户可信度进行打分，Integrate 模块结合前两个模块的输出进行虚假新闻判别。为了提高实时谣言检测的准确率，Singh 等人[22]在内容特征的基础上加入用户特征，首先利用基于 Attention 的 LSTM 提取文本内容特征，然后结合用户特征进行谣言判断；黄学坚等人[23]提出一种融合多元用户特征和内容特征的微博谣言实时检测模型，利用用户的历史行为数据，挖掘用户理性值和专业度两个深层次特征；Kaliyar 等人[24]融合新闻内容、用户-新闻关联关系、用户-用户关系构建多元张量矩阵，对张量矩阵进行分解得到内容和用户的融合特征，基于融合特征进行谣言检测。为了降低模型对传播特征的依赖，提高模型早期谣言检测的性能， Tu 等人[25]提出一种联合文本和传播结构表示学习的谣言检测框架 Rumor2vec，实验表明该方法对谣言的识别提前了至少 12 小时；Lotfi 等人[27]提出一种联合用户特征和传播特征的谣言检测方法，通过图卷积神经网络（Graph Convolutional Neural Networks, GCN）学习谣言传播图和用户交互图的信息； Sun 等人[28]提出一种双动态图卷积神经网络

（Dual Dynamic Graph Convolutional Networks, DDGCN），对传播中的消息动态以及来自知识图谱的背景知识进行建模。

通过研究发现，在内容特征的基础上融入更多的辅助特征是提升实时谣言检测性能的重要手段。所以，本文基于已有的研究基础，提出一种融合内容特征、用户特征和背景知识的网络谣言实时检测方法，通过知识蒸馏从外部知识图谱中获取内容描述的背景知识，增强文本的语义理解。通过用户特征，例如是否认证、粉丝量、关注量等，挖掘用户的可信度。

3 模型构建

3.1 问题定义

谣言有广义和狭义两种定义，广义的谣言定义是指没有经过验证、不确定真实性的消息，狭义的谣言定义是指和事实不符的虚假信息。目前大部分研究把谣言检测视为二分类的虚假信息检测。根据检测粒度的粗细，分为帖子级别和事件级别两种谣言检测。事件级别的谣言检测基于构成事件的一组帖子对整个事件的真实性进行检测，帖子级别的谣言检测主要关注单个帖子的真实性。本文的研究目标是实时验证单个帖子的真实性，即是否是谣言。给定一组社交媒体上的帖子 $D=\{p_1,p_2,\cdots,p_N\}$，学习一个模型 $f:\mathrm{D}\rightarrow Y$，把每个帖子 p_i 分类到预定义的类别 $Y=\{0,1\}$，0 表示非谣言，1 表示谣言。

3.2 总体架构

本文提出的一种融合知识图谱的图注意力神经网络谣言实时检测方法（GAGK-RD）的总体框架如图 1 所示。首先，通过知识蒸馏从外部知识图谱中获得文本描述实体的背景知识。其次，利用点互信息把文本和背景知识转化为图结构表示。再次，利用一种考虑边权重的图注意力神经网络（Weighted Graph Attention Network, W-GAT）对图结构进行学习，得到文本的非连续语义特征。然后，基于预训练语言模型 BERT 得到文本的连续语义特征，基于 Embeding 得到用户特征和内容统计特征的连续向量表示。最后，把非连续语义特征、连续语义特征、用户特征和内容统计特征进行拼接，输入全连接神经网络（Fully Connected Neural Network, FCNN）中进行二分类。

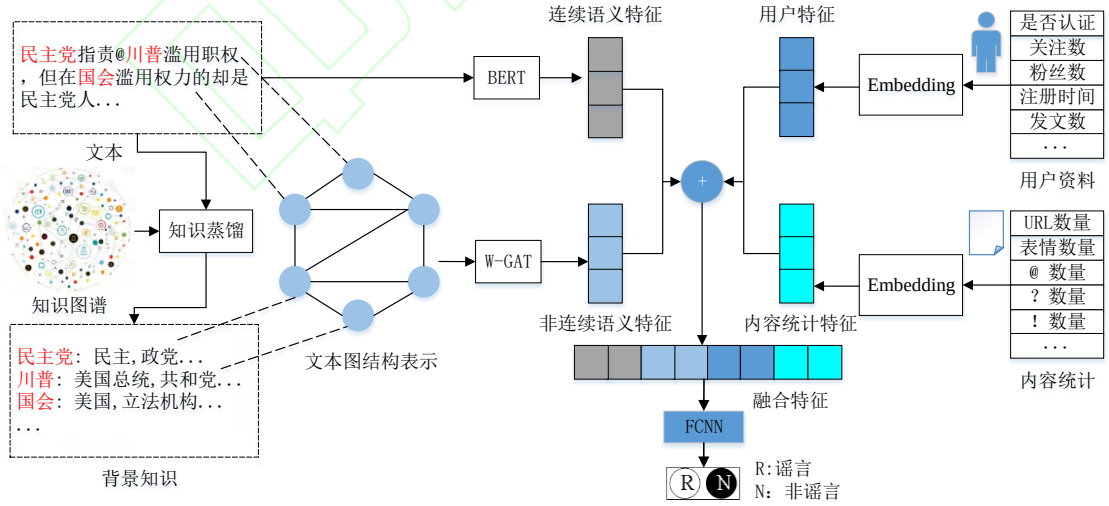


图 1 GAGK-RD 总体架构

Fig.1 GAGK-RD overall framework

3.2.1 知识蒸馏

在知识蒸馏的早期研究中，主要关注于如何使用一个大型的模型来指导小型

模型的训练，以达到更好的泛化性能和响应速度。在知识图谱中，由于图谱规模通常非常庞大，因此在实际应用中，处理和存储大型知识图谱的成本非常高。因此，文本借助知识蒸馏的理念，从大型的知识图谱中抽取出和我们问题领域相关的知识，即社交短文本中包含的实体的概念解释。这些概念解释可以充当为背景知识，很好地补充短文本的语义表示。知识蒸馏的具体流程如图 2 所示。首先，通过实体链接方法，把文本中模棱两可的实体提及 M 链接到知识图谱中的实体 T。然后，对于每个已识别的实体 $t \in T$，通过概念化从知识图谱中获取其概念信息。例如，对于一段短文本“民主党指责@川普滥用职权，但在国会滥用权力的却是民主党人…”，通过实体链接获得实体集合 T={民主党,川普,国会}。然后，基于知识图谱对 T 中的实体进行概念化，获得其概念集 $C_{民主党}$ ={民主,政党}、$C_{川普}$ ={美国总统,共和党}和 $C_{国会}$ ={美国,立法机构}。对于一个帖子 p_i，从知识图谱中进行知识蒸馏，得到 p_i 中每个实体 E_n 的一组概念 $C_{E_n}=\{c_1,c_2,\cdots c_m\}$，把所有概念集合进行合并得到帖子 p_i 的背景知识 $BK_{p_i}=\{C_{E_1},C_{E_2},\cdots C_{E_N}\}$。

图 2 知识蒸馏

Fig.2 Knowledge distillation

3.2.2 文本图结构表示

社交文本经常呈现离散化和碎片化的特点，为了更好的挖掘文本的非连续的语义特征，把文本转换成图结构表示形式 $G=(V,E)$，其中 V 表示节点集合，E 表示边集合。单词作为节点，单词之间的相关性作为边，单词间的关联程度作为边的权重。本文采用点互信息（Point Mutual Information, PMI）计算单词间的相关联程度，计算如公式(1)-(3)所示[29]。

$$p(w_i)=\frac{|W(w_i)|}{|W|} \tag{1}$$

$$p(w_i,w_j)=\frac{|W(w_i,w_j)|}{|W|} \tag{2}$$

$$PMI(w_i,w_j)=\log\frac{p(w_i,w_j)}{p(w_i)p(w_j)} \tag{3}$$

其中，$|W|$ 是滑动窗口的总数量，$|W(w_i)|$ 是包含单词 w_i 的滑动窗口的数量，$|W(w_i,w_j)|$ 是同时包含单词 w_i 和 w_j 的滑动窗口的数量。统计数据是基于全局的语料库而非一个特定帖子内容。PMI 值反映了单词之间的相关性，正值意味着语义相关性较高。因此，本文只保留 PMI 为正值的边，如公式(4)所示。

$$A_{i,j}=\begin{cases}PMI(w_i,w_j) & PMI(w_i,w_j)>0\\ 0 & PMI(w_i,w_j)<=0\end{cases} \tag{4}$$

本文把通过知识蒸馏获取的背景知识 BK_{p_i} 中的所有实体概念作为共现词，添加到帖子 p_i 的文本内容中。最终，通过 PMI 为每个帖子 p_i 构建一个包含文本词和实体概念的图结构表示 G 和邻接权重矩阵 **A**。

3.2.4 非连续语义特征学习

得到融合背景知识的文本图结构表示 G 和邻接权重矩阵 **A** 后，本文利用一种考虑边权重的图注意力神经网络（W-GAT）[30]进行非连续语义特征学习。W-GAT 的构建过程如图 3 所示。

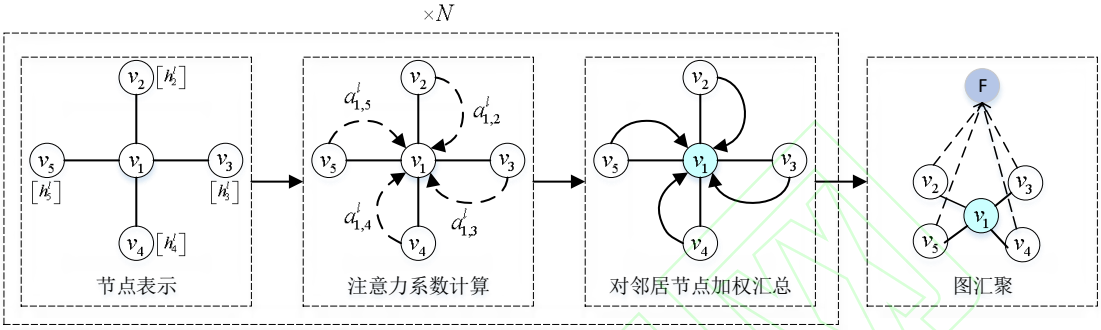


图 3 W-GAT 构造过程

Fig.3 Construction process of W-GAT

首先，将每个节点表示为向量，并将其作为网络的输入。其次，计算每个邻居节点对于中心节点的注意力系数。然后，对邻居节点的表示进行加权汇总，更新每个节点的表示，并作为下一层网络的输入。最后，把最后一层的所有节点的表示进行汇聚，得到整个图的表示。总体来看，W-GAT 的构建过程和 GAT 相似，但 W-GAT 在计算注意力系数时，不仅考虑节点之间的特征相似性，还将边权重作为一个额外的输入因素，从而得到更准确的全局信息。在 W-GAT 的第 l 层输入特征矩阵 $h^l \in \Re^{n\times d^l}$，$n$ 为 G 中的节点数量，d^l 为节点的 l 阶特征维度，输出为节点的 $l+1$ 阶特征矩阵 $h^{(l+1)} \in \Re^{n\times d^{(l+1)}}$，计算如公式(5)所示[30]。

$$h_i^{(l+1)} = \sigma\Big(\sum_{j\in N(i)} a_{i,j}^l W^l h_j^l\Big) \qquad (5)$$

其中，$h_i^{(l+1)}$ 代表节点 v_i 的 $l+1$ 阶特征向量，σ 代表激活函数，$N(i)$ 代表节点 v_i 的邻居节点集合，W^l 代表 W-GAT 第 l 层的学习参数，h_j^l 代表邻居节点 v_j 的 l 阶特征向量，$a_{i,j}^l$ 代表在 l 层中邻居节点 v_j 对节点 v_i 的权重。本文利用 Word2vec 作为每个单词节点的初始特征表示，即一阶特征表示。$a_{i,j}^l$ 的计算如公式(6)所示[30]。

$$a_{i,j}^l = \frac{\exp(e_{i,j}^l)}{\sum_{j\in N_i}\exp(e_{i,j}^l)} \qquad (6)$$

$$e_{i,j}^l = \alpha(h_i^l, h_j^l, A_{i,j}) \qquad (7)$$

其中，$e_{i,j}^l$ 表示在 W-GAT 第 l 层中节点 v_i 和节点 v_j 之间的注意力系数，其计算如公式(7)所示[30]。其中，α 是一个可学习的函数，$A_{i,j}$ 为连接节点 v_i 和节点 v_j 的边的权重，即单词 w_i 和 w_j 的 PMI 值。经过 l 层的 W-GAT 学习后，文本利用全局平均池化来聚合图中每个节点 v_i 的 $l+1$ 阶特征向量 $h_i^{(l+1)}$，得到文本的非连续语义特征 F_{NCSF}，计算如公式(8)所示，其中 V 表示节点集合，$|V|$ 表示节点集合大小。

$$F_{NCSF} = \frac{1}{|V|}\sum_{i \in V} h_i^{(l+1)} \quad (8)$$

3.2.5 连续语义特征学习

社交文本除了呈现离散化和碎片化的特点外，还具有一定的前后逻辑关联。所以，本文采用预训练语言模型 BERT 学习文本的连续语义特征表示。BERT 有通过全局池化和特殊标记 [CLS]两种方式来获取整个文本的语义表示。本文采用相对简单的[CLS]标记的方式，在每个文本词序列前插入一个[CLS]符号，将该符号对应的最后一层的输出向量 h_{CLS} 作为文本续语义特征表示 F_{CSF}，如公式(9)所示。

$$F_{CSF} = BERT\text{-}CLS(w_1, w_2, \cdots w_n) \quad (9)$$

3.2.6 用户特征学习

社交网络上的用户特征，例如是否认证、粉丝量、关注量、注册时长和发文量等信息一定程度上可以反映该用户的可信赖程度。相关研究也表明[31]，非认证用户比认证用户散布谣言的可能性更大；粉丝数远小于关注数的用户更可能散布谣言信息；用户注册的时间越长、发布的信息越多其谣言识别能力越强，散布谣言的概率越小。因此，本文把用户特征作为辅助特征，增强模型对谣言的识别能力，具体用户特征如表 1 所示。通过 Embedding 将这些离散型特征转化为连续型向量表示，并进行拼接得到用户特征向量表示 F_{User}，如公式(10)所示。

$$F_{User} = Emb(FU_1) \oplus \cdots \oplus Emb(FU_5) \quad (10)$$

表 1 用户特征

Table 1 User features

编号	特征描述
FU1	是否认证
FU2	粉丝数
FU3	关注数
FU4	注册时长
FU5	历史发文量

3.2.7 内容统计特征学习

在进行文本语义特征表示时，通常忽略文本中的表情、符号、URL 等信息，这些信息对谣言的判断也具有一定的辅助作用。例如，一些谣言为了获得用户流量通常会在文本中加入 URL 链接。因此，本文把内容统计特征也作为辅助特征，具体内容统计特征如表 2 所示。同理，通过 Embedding 进行处理得到内容统计特征向量表示 F_{SC}，如公式(11)所示。

$$F_{SC} = Emb(FC_1) \oplus \cdots \oplus Emb(FC_7) \quad (11)$$

表 2 内容统计特征

Table 2 Statistical features of content

编号	特征描述
FC1	是否有#标识的话题
FC2	包含的 URL 数量
FC3	包含的表情符号数量
FC4	包含的@符号数量
FC5	包含的？号数量
FC6	包含的！号数量
FC7	是否附有图片或视频

3.2.8 结果预测

获得非连续语义特征 F_{NCSF}、连续语义特征 F_{CSF}、用户特征 F_{User} 和内容统计特

征 F_{SC} 后，本文对其进行拼接得到最终的融合特征 F，如公式(12)所示。然后，文本把融合特征 F 输入全连接神经网络 FCNN 中，最后连接 *softmax* 得出分类结果，如公式(13)所示。

$$F = F_{NCSF} \oplus F_{CSF} \oplus F_{User} \oplus F_{SC} \tag{12}$$

$$y = softmax(W * FCNN(F) + b) \tag{13}$$

其中， W 和 b 分别表示线性层参数和偏置项。本文基于最小化交叉熵损失函数对模型的参数进行训练，交叉熵损失函数计算如公式(14)所示[32]。

$$L = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log y_i + (1-y_i)\log(1-y_i)) + \frac{\lambda}{2}\|W\|_2^2 \tag{14}$$

其中， y_i 表示样本 i 的真实标签值， y_i 表示模型的预测值， $\frac{\lambda}{2}\|W\|_2^2$ 为 L_2 正则化，降低模型的过拟合程度。

4 实验及分析

4.1 实验设置

4.1.1 实验数据和评价指标

实验使用两个公开的真实社交媒体谣言数据集：PHEME 和 WEIBO，来验证本文提出的 GAGK-RD 方法的有效性。PHEME 数据集包含查理周刊枪击案、弗格森骚乱、德国之翼飞机失事和渥太华枪击案五个突发新闻期间发布在推特上的谣言和非谣言的集合。WEIBO 数据集是从新浪微博不实信息举报平台抓取的中文谣言数据集，包含了各种领域的谣言和非谣言，如健康、政治、金融等。PHEME 和 WEIBO 两个数据集都包含原始帖子的文本信息、发文用户信息和传播信息。GAGK-RD 方法只利用文本信息和发文用户信息进行谣言的实时检测。PHEME 和 WEIBO 的统计数据如表 3 所示。

表 3 数据集统计

Table 3 Dataset statistics

Datasets	Rumors	Non-rumours	Total
PHEME	1972	3830	5802
WEIBO	2313	2351	4664

实验按照 3:1:1 的比例对数据集进行分层随机抽样，得到训练集、验证集和测试集，使用 K 折交叉验证的平均准确率（Accuracy）、谣言查准率（Precision）、谣言查全率（Recall）和 F1-score 作为方法性能评价指标。

4.1.2 实现细节

在知识蒸馏中，本文采用文献[33]提出的短文本实体链接方法，利用 Probase 和 YAGO 两个知识图谱获取实体的概念知识。PMI 计算时中文和英文文本的滑动窗口大小分别设置为 9 和 6。W-GAT 的层数设置为 2 层，第 1 层和第 2 层的输出维度分别为 128 和 64。使用预训练的词向量库 Chinese-World-Vectors 和 GoogleNews-Vectors 分别提取中文和英文的词向量表示，它们的词向量维度都是 300。利用基于多语言预训练的 BERT(bert-base-multilingual-uncased)获取文本的连续语义特征表示。用户特征和内容统计特征的 Embedding 维度都为 8。在结果预测中，采用 3 层的全连接神经网络，第 1 层、第 2 层和第 3 层的输出维度分别为 128、64 和 32。使用 Adam 优化器进行训练，学习率设置为 0.001，batch_size 为 16，交叉验证的 K-fold 为 5。实验基于 Pyhon3.8 和 Pytorch 深度学习框架进行算法实现，服务器配置为单卡 NVIDIA Ge Force RTX 2080 Ti 图形处理器、32 核 CPU、128GB 内存。模型和训练的具体超参数设置如表 4 所示。

表 4 超参数设置

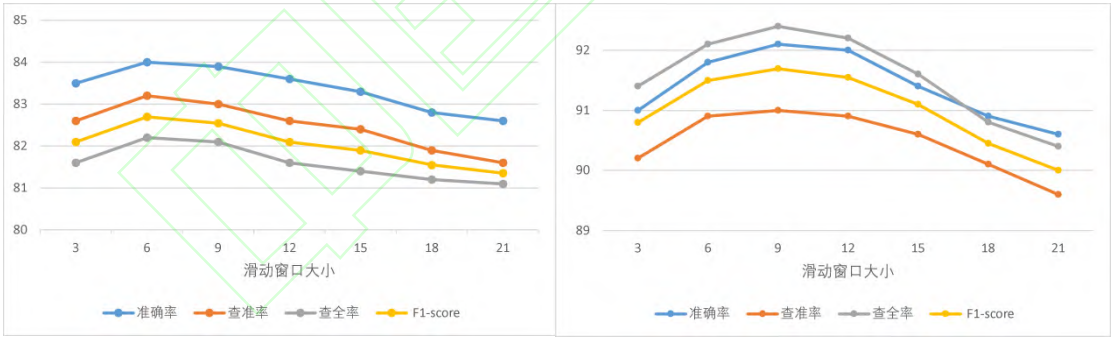
Table 4 Hyperparameter settings

参数类型	参数名称	参数值
模型参数	W-GAT 层数	2
	W-GAT 的输出维度	128,64
	中文文本滑动窗口大小	9
	英文文本滑动窗口大小	6
	词向量维度	300
	全连接神经网络的层数	3
	全连接神经网络的输出维度	128,64,32
	用户特征的 Embedding 维度	8
	内容统计特征的 Embedding 维度	8
训练参数	学习率	1e-3
	正则化参数	1e-5
	最大训练轮次	30
	停止训练的等待次数	8
	批量训练的 batch size	16
	Dropout 比例	0.1
	交叉验证的 K-fold	5
	优化器 Optimizer	Adam

4.2 实验结果和分析

4.2.1 不同滑动窗口大小的实验对比

PMI 计算两个词之间的相关性时，需要使用一个滑动窗口来确定上下文范围。本文采用网格搜索的方式来确定滑动窗口的大小，设定搜索空间为{3, 6, 9, 12, 15, 18, 21}。GAGK-RD 在不同滑动窗口下的准确率、查准率、查全率和 F1-score 如图 4 所示。



(a)PHEME (b)WEIBO

图 4 不同滑动窗口大小的实验对比

Fig.4 Comparison of experimental results with different sliding window sizes

从图 4(a)和图 4(b)中可以看出，滑动窗口设置的太小或太大都会影响模型的性能。当滑动窗口太小时，导致上下文信息不足，无法捕捉到两个词之间的关联性。相反，当滑动窗口太大时，会包含大量的无关信息，可能淹没与两个词有关的信息。数据集 PHEME 和 WEIBO 的最优滑动窗口大小分别是 6 和 9，前者小于后者的原因可能是 PHEME 的英文文本通常比 WEIBO 的中文文本的语法结构更为紧凑，表达方式更加简洁。

4.2.2 方法对比分析

为了验证本文提出的 GAGK-RD 方法的有效性，选择以下谣言检测方法作为基线：

GRU-2[4]（2016）：一种基于传播特征的谣言检测方法，通过门控循环神经网络 GRU 学习帖子的转发评论信息随时间的变化。

SVM-TS[20]（2018）：一种基于人工特征工程的谣言检测方法，设计了 13 个内容统计特征、11 个用户统计特征和 3 个传播统计特征，并考虑了这些统计特征随时间的变化。

LSTM-CNN[12]（2018）：一种基于内容特征的谣言检测方法，结合 LSTM 和 CNN 提取谣言文本语义特征。

TDRD[16]（2020）：一种基于内容特征的谣言检测方法，首先利用 CNN 挖掘文本主题信息，然后结合主题信息和文本词嵌入进行谣言判断。

LSTM_Word2vec[14]（2020）：一种基于内容特征的谣言检测方法，通过联合 LSTM 和 Word2vec 挖掘文本的谣言语义特征。

DGCN[10]（2021）：一种基于传播特征的谣言检测方法，通过动态图卷积神经网络 GCN 捕获谣言传播的结构和时间信息。

Rumor2vec[25]（2021）：一种联合内容特征和传播特征的谣言检测方法，减轻谣言级联传播结构的稀疏性问题。

User-Reply-GCN[27]（2021）：一种联合用户特征和传播特征的谣言检测方法，通过图卷积神经网络 GCN 学习谣言传播图和用户交互图的信息。

GAN_based[15]（2021）：一种基于内容特征的谣言检测方法，通过对抗网络的生成器和判别器的相互促进作用，强化谣言文本特征的学习。

LSTM-Attention[22]（2022）：一种融合内容特征和用户特征的谣言检测方法，利用基于 Attention 的 LSTM 提取文本语义特征，并结合用户和内容的统计特征进行谣言判断。

UMLARD[18]（2022）：一种融合传播特征、内容特征和用户特征的谣言检测方法，通过多层扩散图卷积网络 M-DGCN 和时间衰减 LSTM 学习这些特征的表示。

DDGCN[28]（2022）：一种在传播特征中融合背景知识的谣言检测方法，通过双动态图卷积网络 GCN 对传播中的消息动态以及来自知识图谱的背景知识进行建模。

以上 12 个基线方法分为考虑传播特征的非实时谣言检测方法和不考虑传播特征的实时谣言检测方法，具体实验结果如表 5 和表 6 所示。

表 5 在 WEIBO 上的实验结果

Table 5 Experimental results on the WEIBO

时效性	方法	准确率	查准率	查全率	F1-score
非实时检测	SVM-TS	84.4	86.0	85.3	85.6
	GRU-2	90.1	87.6	95.6	91.4
	User-Reply-GCN	91.6	90.2	92.3	91.2
	UMLARD	92.8	94.2	96.5	95.3
	DGCN	93.2	92.3	94.2	93.2
	DDGCN	94.8	95.3	94.8	95.0
	Rumor2vec	95.1	95.8	94.8	95.3
实时检测	LSTM-Word2vec	84.1	83.7	82.8	83.2
	LSTM-CNN	85.2	85.7	86.1	85.9
	TDRD	86.7	87.0	86.0	86.5
	GAN-based	86.8	86.5	89.4	87.9
	LSTM-Attention	89.6	89.9	91.6	90.7
	GAGK-RD	**92.1**	**91.0**	**92.4**	**91.7**

表 6 在 PHEME 上的实验结果

Table 6 Experimental results on the PHEME

时效性	方法	准确率	查准率	查全率	F1-score
非实时检测	SVM-TS	78.3	69.2	73.1	71.1
	GRU-2	82.6	82.1	81.0	81.5
	User-Reply-GCN	83.6	81.2	83.3	82.2
	UMLARD	84.2	83.8	83.5	83.6
	DGCN	84.0	82.9	82.1	82.5
	DDGCN	85.5	84.6	84.1	84.3
	Rumor2vec	86.7	84.2	85.3	84.7
实时检测	LSTM-Word2vec	79.5	72.8	70.6	71.7
	LSTM-CNN	80.4	80.1	81.1	80.6
	TDRD	82.7	81.3	78.6	79.9
	GAN-based	82.7	81.6	79.1	80.3
	LSTM-Attention	83.0	82.3	81.6	81.9
	GAGK-RD	**84.0**	**83.2**	**82.2**	**82.7**

从表 5 和表 6 中的实验结果可以发现，大部分基于深度学习自动学习谣言特征的方法的性能好于基于人工特征工程的统计机器学习方法 SVM-TS，这是因为人工设计的特征缺乏全面性和灵活性，无法表征谣言的深层语义特征。深度学习模型通过数据驱动的方式自动学习适合于任务的高层次抽象特征，具有更强的自适应性和更全面的多样性。基于传播特征的非实时的谣言检测方法的性能普遍优于基于内容特征的实时的谣言检测方法，因为基于传播特征的方法能够更好地反映谣言传播的路径和过程，捕捉到社交网络中的社交关系和交互行为，从而能够更准确地刻画谣言传播的特点。但基于传播特征的方法需要谣言在社交网络得到一定范围的传播后才能获得较好的效果，无法做到实时的谣言检测。为了降低模型对传播特征的依赖，User-Reply-GCN、UMLARD、DDGCN 和 Rumor2vec 在传播特征的基础上融合用户特征、内容特征和外部知识等，提高方法的早期检测性能，实验结果也显示这些融合方法比仅基于传播特征的 GRU-2 具有更高的准确率。

在实时检测中，所有的对比方法都采用了深度学习模型学习谣言文本的语义特征。融合了用户特征的 LSTM-Attention 方法和融合了文本主题特征的 TDRD 方法的准确率高于只基于文本语义特征的 LSTM-Word2vec 和 LSTM-CNN 两种方法。本文提出的 GAGK-RD 方法通过融合内容特征、用户特征和背景知识，在所有的评价指标上获得了最高值，并且优于 GRU-2 和 User-Reply-GCN 这两种非实时检测方法，证明了本文方法的有效性。

4.2.3 消融实验分析

为了分析不同模块对 GAGK-RD 方法的贡献，本文设计了 8 组消融实验，方法①-⑤分别为去除非连续语义特征 F_{NCSF}、连续语义特征 F_{CSF}、用户特征 F_{User}、内容统计特征 F_{SC} 和背景知识 BK，方法⑥为同时去除非连续语义特征 F_{NCSF} 和连续语义特征 F_{CSF}，方法⑦和⑧把考虑边权重的图注意力神经网络 W-GAT 分别换成普通的图注意力神经网络 GAT 和图卷积神经网络 GCN。每个方法相比于 GAGK-RD 方法的准确率、查准率、查全率和 F1-score 的变化情况如表 7 所示。

表 7 消融实验

Table 7 Ablation experiments

方法	WEIBO				PHEME			
	ΔAcc	ΔPre	ΔRec	ΔF1	ΔAcc	ΔPre	ΔRe	ΔF1
① (-) F_{NCSF}	-1.8	-1.4	-1.3	-1.3	-1.7	-1.4	-1.0	-1.2
② (-) F_{CSF}	-1.1	-0.8	-0.6	-0.7	-0.8	-0.7	-0.7	-0.7
③ (-) F_{User}	-2.4	-2.0	-2.2	-2.1	-2.0	-1.7	-1.5	-1.6

④ (-) F_{SC}	-0.4	-0.3	-0.5	-0.4	-0.2	-0.3	-0.2	-0.3
⑤ (-)BK	-1.0	-1.0	-0.7	-0.8	-0.9	-0.9	-0.9	-0.8
⑥ (-) F_{NCSF} F_{CSF}	-20.6	-22.8	-20.6	-21.7	-18.7	-18.0	-20.0	-19.0
⑦ W-GAT→GAT	-0.6	-0.4	-0.4	-0.4	-0.5	-0.4	-0.6	-0.5
⑧ W-GAT→GCN	-0.8	-0.8	-0.7	-0.7	-0.7	-0.8	-1.0	-0.9

从表 7 中可以发现，每个模块都发挥着各自的作用，去除任何一个模块都会影响 GAGK-RD 的性能。非连续语义特征 F_{NCSF} 对 GAGK-RD 的影响大于连续语义特征 F_{CSF}，因为社交帖子内容经常呈现离散化和碎片化的特点，图神经网络能更好的捕获文本的非连续和远距离依赖的语义特征。F_{NCSF} 和 F_{CSF} 一定程度上具有互补性和冗余性，所以只去除其中之一个并不会对 GAGK-RD 带来太大的影响，但同时去除 F_{NCSF} 和 F_{CSF} 会给方法带来巨大的影响，因为不考虑文本的语义特征，仅基于用户特征和内容统计特征无法提供足够的信息来区分谣言和非谣言。去除用户特征会降低 GAGK-RD 的性能，因为用户的否认证、粉丝量、关注量、注册时长和发文量等特征能够反映用户的权威性、专业性和可信度。去除内容统计特征会略微降低 GAGK-RD 的性能，因为文本中的表情、符号、URL 等统计信息有时也能为谣言的判断提供一定的依据。通过知识蒸馏从外部知识图谱中提取的背景知识可以补充短文本的语义表示，所以去除背景知识也会影响 GAGK-RD 的性能。把 W-GAT 换成 GAT 后方法的性能下降，因为 W-GAT 在计算注意力系数时，不仅考虑节点之间的特征相似性，还将边权重作为一个额外的因素，可以得到更准确的全局信息。把 W-GAT 换成 GCN 对 GAGK-RD 的性能影响更大，因为相比于 GCN 的平均池化，GAT 中的注意力机制可以使得节点更加灵活地聚合邻居节点的信息，从而提高模型的性能。

5 结语

对谣言进行实时检测，是控制谣言传播最有效的方式，也是目前谣言检测研究的难点问题。本文提出一种融合非连续语义特征、连续语义特征、用户特征、内容统计特征和背景知识的实时谣言检测方法，利用考虑边权重的图注意力神经网络学习离散化文本的非连续语义特征，通过知识蒸馏从知识图谱中获取的背景知识增强文本的语义表示。实验结果表明，本文方法优于所有对比的实时谣言检测基准方法，证明了本文方法的有效性。虽然，本文方法提高了谣言实时检测的准确率，但方法没有对帖子中可能附加的图片或视频信息进行联合分析，如何进行多模态学习的谣言检测是下一步的研究计划。

参考文献

[1] 黄学坚,马廷淮,王根生.基于分层语义特征学习模型的微博谣言事件检测[J/OL].数据分析与知识发现:1-13, (2022-07-22).[2023-03-30].http://kns.cnki.net/kcms/detail/10.1478.G2.20220721.1739.002.html.
(Huang Xuejian, Ma Tinghuai, Wang Gensheng. Weibo rumor events detection based on hierarchical semantic feature learning model[J/OL]. Data Analysis and Knowledge Discovery:1-13, (2022-07-22).[2023-03-30]. http://kns.cnki.net/kcms/detail/10.1478.G2.20220721.1739.002.html.)

[2] Davoudi M, Moosavi M R, Sadreddini M H. DSS: A hybrid deep model for fake news detection using propagation tree and stance network[J]. Expert Systems with Applications, 2022, 198: 116635.

[3] Ma J, Gao W, Wong K F. Detect rumors in microblog posts using propagation structure via kernel learning[C]. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada: ACL, 2017: 780–717.

[4] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[C]. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, USA: AAAI Press, 2016: 3818-3824.

[5] 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.
[6] Chen T, Li X, Yin H, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[C]. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, Australia: Springer, 2018: 40-52.
[7] 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.
(Hu Dou, Wei Lingwei, Zhou Wei, et al. A rumor detection approach based on multi-relational propagation tree[J]. Journal of Computer Research and Development, 2021, 58(7): 1395-1411.)
[8] Song C, Yang C, Chen H, et al. CED: credible early detection of social media rumors[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(8): 3035-3047.
[9] Wu Z, Pi D, Chen J, et al. Rumor detection based on propagation graph neural network with attention mechanism[J]. Expert systems with applications, 2020, 158: 113595.
[10] Choi J, Ko T, Choi Y, et al. Dynamic graph convolutional networks with attention mechanism for rumor detection on social media[J]. Plos one, 2021, 16(8): e0256039.
[11] Kaliyar R K, Goswami A, Narang P, et al. FNDNet–a deep convolutional neural network for fake news detection[J]. Cognitive Systems Research, 2020, 61: 32-44.
[12] Ajao O, Bhowmik D, Zargari S. Fake News Identification on Twitter with Hybrid CNN and RNN Models[C]. In: Proceedings of the 9th International Conference on Social Media and Society, Copenhagen, Denmark: ACM, 2018: 226–230
[13] 刘赏,沈逸凡.基于新闻标题-正文差异性的虚假新闻检测方法[J].数据分析与知识发现, 2023, 7(2): 97-107.
(Liu Shang, Shen Yifan. Detecting fake news based on title-content difference[J]. Data Analysis and Knowledge Discovery, 2023, 7(2): 97-107.)
[14] 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.
[15] 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.
[16] Xu F, Sheng V S, Wang M. Near real-time topic-driven rumor detection in source microblogs[J]. Knowledge-Based Systems, 2020, 207: 106391.
[17] Ma T, Zhou H, Tian Y, et al. A novel rumor detection algorithm based on entity recognition, sentence reconfiguration, and ordinary differential equation network[J]. Neurocomputing, 2021, 447: 224-234.
[18] Chen X, Zhou F, Trajcevski G, et al. Multi-view learning with distinguishable feature fusion for rumor detection[J]. Knowledge-Based Systems, 2022, 240: 108085.
[19] Castillo C, Mendoza M, Poblete B. Information Credibility on Twitter[C]. In: Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India: ACM, 2011:675-684.
[20] Ma J, Gao W, Wei Z, et al. Detect Rumors Using Time Series of Social Context Information on Microblogging[M]. Social Media Content Analysis: Natural Language Processing and Beyond. 2018: 67-77.
[21] Ruchansky N, Seo S, Liu Y. Csi: A hybrid deep model for fake news detection[C]. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore: ACM, 2017: 797-806.
[22] 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: 459-474.
[23] 黄学坚,王根生,罗远胜,等. 融合多元用户特征和内容特征的微博谣言实时检测模型[J]. 小型微型计算机系统, 2022, 43(12): 2518-2527.
(Huang Xuejian, Wang Gensheng, Luo Yuansheng, et al. Weibo Rumors Real-time Detection Model Based on Fusion of Multi User Features and Content Features. Journal of Chinese Computer Systems, 2022, 43(12): 2518-2527.)
[24] Kaliyar R K, Goswami A, Narang P. DeepFakE: improving fake news detection using tensor decomposition-based deep neural network[J]. The Journal of Supercomputing, 2021, 77: 1015-1037.
[25] Tu K, Chen C, Hou C, et al. Rumor2vec: a rumor detection framework with joint text and propagation structure representation learning[J]. Information Sciences, 2021, 560: 137-151.
[26] Bai N, Meng F, Rui X, et al. Rumour detection based on graph convolutional neural net[J]. IEEE Access, 2021, 9: 21686-21693.
[27] Lotfi S, Mirzarezaee M, Hosseinzadeh M, et al. Detection of rumor conversations in Twitter using graph convolutional networks[J]. Applied Intelligence, 2021, 51: 4774-4787.
[28] Sun M, Zhang X, Zheng J, et al. Ddgcn: Dual dynamic graph convolutional networks for rumor detection on social media[C]. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, California, USA: AAAI Press, 2022: 4611-4619.

[29] Church K, Hanks P. Word association norms, mutual information, and lexicography[J]. Computational linguistics, 1990, 16(1): 22-29.
[30] Inan E. ZoKa: a fake news detection method using edge-weighted graph attention network with transfer models[J]. Neural Computing and Applications, 2022, 34(14): 11669-11677.
[31] 刘雅辉,靳小龙,沈华伟,等.社交媒体中的谣言识别研究综述[J].计算机学报, 2018, 41(7): 1536-1558. (Liu Yahui, Jin Xiaolong, Shen Huawei, et al. A Survey on Rumor Identification over Social Media[J]. Chinese Journal of Computers, 2018, 41(7): 1536-1558.)
[32] Ho Y, Wookey S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling[J]. IEEE access, 2019, 8: 4806-4813.
[33] Chen L, Liang J, Xie C, et al. Short text entity linking with fine-grained topics[C]. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy: ACM, 2018: 457-466.

通讯作者（Corresponding author）:王根生（Wang Gensheng），ORCID：0000-0002-2443-831X，E-mail:wgs74@126.com。

基金项目：本文系国家自然科学基金项目“社交媒体健康信息可信度评估及偏好推荐研究”（项目编号：72061015）的研究成果之一。

This work is supported by the National Natural Science Foundation of China (Grant No. 72061015).

作者贡献声明：

王根生：提出研究思路，设计研究方案，分析数据，起草、修改论文，完成最终版本修订；
朱奕：采集和清洗数据，负责实验；
李胜：提出修改建议。

利益冲突声明：

所有作者声明不存在利益冲突关系。