Computer Science and Application 计算机科学与应用, 2024, 14(3), 178-185
Published Online March 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.143068

Hans汉斯

融合背景知识的异构图虚假新闻检测方法研究

何　迈，肖克晶*，曹少中，张　寒，姜　丹

北京印刷学院信息工程学院，北京

收稿日期：2024年2月20日；录用日期：2024年3月21日；发布日期：2024年3月29日

摘　要

如今虚假新闻检测任务越来越受人们重视。本文考虑到不同的新闻具有涉及领域众多、隐含背景信息丰富的特点，提出利用新闻中的实体链接到领域广、信息全的维基百科，挖掘新闻潜在的背景信息与结构化三元组信息组成异构图，丰富新闻的表示。为了学习并更新建模后新闻异构图的特征向量，在图卷积网络的基础上，提出了一个基于语义距离的图卷积网络注意力模型DGAT (Distance Graph Attention Network, DGAT)。具体的，通过赋予异构图中不同类型节点不同的变化矩阵，将不同类型的节点映射到相同的公共空间中，解决了GCN模型不能直接应用在异构图上的局限。针对本文建模的新闻异构图特点，引入了基于新闻语义距离的注意力机制，以捕获融合了外部知识后，新闻与背景知识的语义一致性，最终输入分类器中进行虚假新闻检测。在公开数据集上进行的实验表明了本文方法的有效性。

关键词

虚假新闻检测，异构图，图卷积网络模型

Research on Fake News Detection Method Using Heterogeneous Graph Fusion with Background Knowledge

Mai He, Kejing Xiao*, Shaozhong Cao, Han Zhang, Dan Jiang

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: Feb. 20th, 2024; accepted: Mar. 21st, 2024; published: Mar. 29th, 2024

Abstract

Nowadays, the task of fake news detection is receiving more and more attention. This article takes

*通讯作者。

文章引用: 何迈, 肖克晶, 曹少中, 张寒, 姜丹. 融合背景知识的异构图虚假新闻检测方法研究[J]. 计算机科学与应用, 2024, 14(3): 178-185. DOI: 10.12677/csa.2024.143068

into account that different news has the characteristics of covering many fields and rich hidden background information. It is proposed to use the entities in the news to link to Wikipedia, which has a wide range of fields and complete information, to mine the potential background information of the news and form a heterogeneous graph with structured triplet information to enrich the representation of the news. In order to learn and update the modeled news heterogeneous graph feature vectors, an improved graph convolutional network model (GCN) and a Distance Graph Attention Network (DGAT) model are proposed. Specifically, by assigning different types of heterogeneous graphs to Different change matrices of nodes map different types of nodes into the same common space, solving the limitation that the GCN model cannot be directly applied to heterogeneous graphs. In view of the characteristics of the news heterogeneous graph modeled in this article, an attention mechanism based on news semantic distance is introduced to capture the semantic consistency of news and background knowledge after fusing external knowledge, and finally input it into the classifier to perform false recognition and news detection. Experiments on public datasets demonstrate the effectiveness of our method.

Keywords

Fake News Detection, Heterogeneous Graph, GCN

1. 引言

随着互联网与社交媒体的快速发展，每天会产生大量的信息。大多数人是从社交媒体接收信息和新闻，并且虚假的信息往往比真实的信息传播更快[1]。虚假新闻的广泛传播会误导大众，还可能引起社会恐慌，扰乱社会治安[2]。因此迫切需要采用自动虚假新闻检测方法，以避免造成严重的负面影响。

虚假新闻检测与文本分类都可以看成是将文章划分为某一类，但是存在差异，因为在很多情况下，虚假新闻检测任务可能在从未见过的各个领域中进行虚假新闻检测[3]。此外，现有的深度学习方法通常侧重于依赖新闻内容的语言语义特征和社交上下文信息，而不能有效利用外部知识帮助确定新闻的真假，当人们在判断一个新闻是否可信时，通常会根据新闻中引申出的背景知识进行对比，来判断新闻的真假。因此新闻中潜在的背景知识，是判定新闻真假的一个重要因素。而维基百科等包含大量高质量的结构化主体、谓词、对象三元组和非结构化实体描述[4]，作为检测虚假新闻的证据，且涉及历史、科学、文化等众多领域，可以作为新闻的背景知识，融合进新闻中，丰富新闻的表示。

因此本文提出利用异构图，融合新闻的背景知识，来建模新闻的表示。首先将新闻中的每个句子拆分成句子节点，并获得新闻句子中的实体。再将实体链接到维基百科获得非结构化的背景摘要信息，作为非结构化背景节点。接着利用实体搜索维基百科的结构化三元组数据库 DBpedia，将获得的结构化的三元组信息构建成知识图谱，使用 TransE [5]模型训练三元组的特征向量表示，作为结构化的三元组节点。最后将这三种类型的节点为每一篇新闻构建一个异构图，即融合了有用的外部知识来丰富新闻的表示。利用基于语义距离的图卷积网络注意力模型 DGAT，对建模后的新闻进行特征向量更新，以捕获融合了外部知识的新闻中的内容一致性。最终输入分类器中进行虚假新闻检测。

2. 虚假新闻检测相关研究

与虚假新闻检测相关的任务有很多，如谣言检测、事实核查等。总体的工作都是从新闻的内容及相

关信息中提取特征判断新闻的真假，可以将现有的模型方法分为两类：基于社交上下文信息与基于新闻内容的虚假新闻检测。

2.1. 基于社交上下文信息的方法

社交上下文信息指的是新闻的发布者，新闻的传播网络，以及其他用户对新闻的评论和转发等信息。虚假新闻的发布者往往也会按照真实新闻的写法撰写虚假新闻，高可信度的用户发表的新闻内容更有可能是真实新闻，且社会学研究表明真实新闻和虚假新闻在社交网络的传播情况往往是不同的。因此利用社交上下文的信息对虚假新闻进行检测是一个有效的途径。

Lu [6]提出 GCAN 模型构建用户图，将用户的简介作为图的初始化节点信息，并采用 GCN 模型学习用户的特征向量表示，最后利用得到的特征向量进行虚假新闻检测。Jiang [7]用异构图建模了新闻的传播网络和用户的社交网络，并将新闻信息和用户信息拼接到一起进行虚假新闻检测。Khoo [8]利用 Transformer 建模时间序列，将源新闻以及其他转发句子作为 Transformer 的输入，将 Transformer 输出的新闻嵌入向量输入分类器中，得到新闻分类结果。Bian [9]提出将新闻的传播过程建模为两个同质图，利用 GCN 模型融合图中的节点信息，获得节点表示。最后将节点表示进行池化、拼接，输入分类器中，得到最终的分类结果。Kang [10]提出了利用新闻与领域、新闻与转发贴子、新闻与发布者之间的关系构建新闻异质信息网络，并使用异质图卷积网络获得新闻节点的特征向量，输入分类器中，得到分类结果。

2.2. 基于新闻内容的方法

基于新闻内容的方法是利用文章中所包含的文本信息以及图片视频等多模态信息作为模型的输入，进行虚假新闻检测。

Ma [11]首次将深度学习技术应用到虚假新闻检测中，将新闻的每个句子输入循环神经网络 RNN，LSTM 或者 GRU 中，利用循环神经网络的隐藏层向量表示新闻信息，将隐藏层信息输入分类器中，得到分类结果。R.Shah [12]利用 VGG19 提取视觉信息，BERT 提取文本信息，将视觉信息和文本信息拼接，输入分类器中，对新闻进行分类。Zhou [13]提出利用新闻的图片信息与文本信息的相似度，区别新闻真伪。首先利用 Image2text 模型将视觉信息转化为文本信息，并通过全连接层将文本信息和视觉信息映射到同一向量空间中，并对比视觉信息和文本信息之间的相似度。如果相似度较高，为真实新闻，反之为虚假新闻。这些方法往往忽略利用隐藏在新闻内容中，可挖掘的新闻背景知识。本文提出了利用异构图融合新闻外部背景知识丰富新闻的表示，并利用改进的 GCN 模型进行虚假新闻检测。

3. 本文提出的虚假新闻检测方法

本节将详细介绍本文提出的虚假新闻检测方法，包含异构图的构建和改进 GCN 模型的思路。

3.1. 新闻异构图的构建

对于数据集中的每篇新闻，都建立一个无向的异构图 $G = (V, E)$。其中 V 是节点的集合，E 为边的集合。节点的集合是由三种类型的节点构成：句子类型节点 $S = \{s_1, s_2, \ldots, s_i\}$，非结构化背景节点 $B = \{b_1, b_2, \ldots, s_j\}$，结构化三元组节点 $T = \{t_1, t_2, \ldots, t_k\}$，即 $V = S \bigcup T \bigcup E$。对于数据集中的一篇新闻，按照句子将该篇新闻划分成若干个句子节点。

如图 1(a)所示，取第一个句子节点“Trump came through… on the upcoming Republican convention and the position of Donald Trump”为例。识别出句子节点中有两个实体“Republican”和“Donald Trump”，利用 TagMe 工具[14]将两个实体链接到维基百科便可以获得这两个实体的背景描述摘要。DBpedia 是维

基百科的结构化三元组数据库，用实体“Republican”和“Donald Trump”搜索 DBpedia，可获得大量关于该实体的相关三元组信息，再将三元组构成如图 1(b)所示的知识图谱。为了对知识图谱进行特征向量的学习，本文选用高效的 TransE 模型，对于每个三元组实例(*head*, *relation*, *tail*)，TransE 模型是将 *relation* 看作为 *head* 到 *tail* 的翻译，通过调整它们的向量表示使得 *head* + *relation* = *tail*。最后将参与知识图谱构建的实体学习到的特征表示作为结构化三元组节点参与新闻异构图的构建。

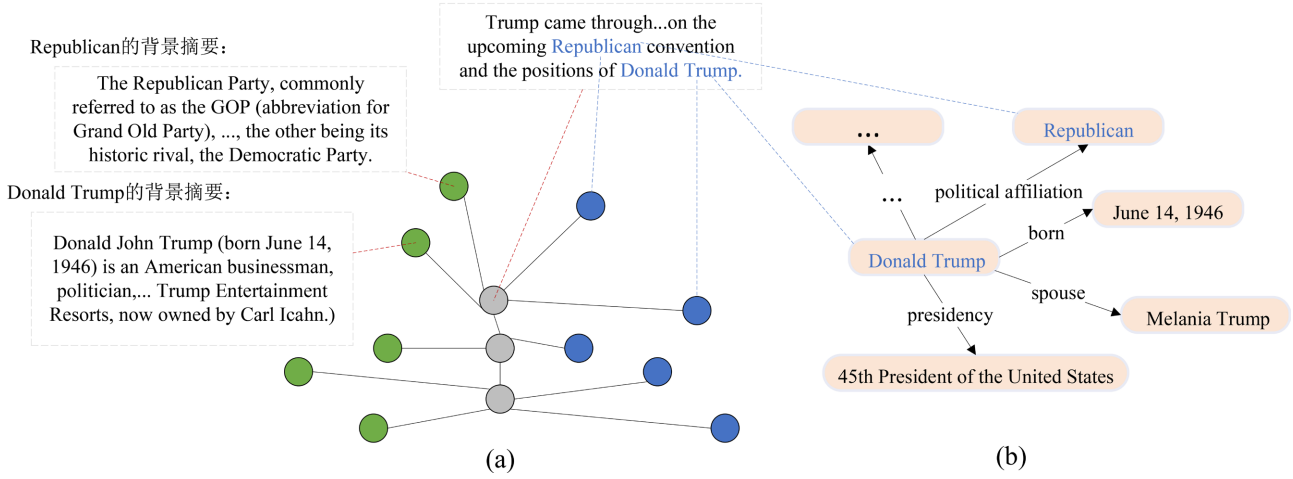

Figure 1. Diagram of news heterogeneous graph and knowledge graph
图 1. 新闻异构图与知识图谱示意图

3.2. 异构图卷积模型 DGAT

GCN 模型[15]是一个可应用在同构图上的图神经网络模型，对于一个给定的图 $G=(V, E)$，其中 V 为图中节点的集合，用 $n\left(n=|V|\right)$ 表示图中节点的个数，E 为图中边的集合。用矩阵 $X \in \mathbb{R}^{n \times l^{(0)}}$ 表示图中所有节点的特征向量，即矩阵 X 的第 i 行表示第 i 个节点的特征向量，$l^{(0)}$ 为初始特征向量的维度。再引入邻接矩阵 $A \in \mathbb{R}^{n \times n}$ 表示边的集合 E，若 A_{ij} 为 1 则表示第 i 个节点与第 j 个节点是邻居节点，也表示两个节点之间的权重值为 1，若为 0 则表示两个节点不相邻。由于图中节点在学习中更新自身向量表示时，不仅利用到邻居节点的向量表示，也需要利用到自身的向量表示，因此加入单位矩阵 I 使得邻接矩阵的表示为$(I+A)$。由于这样会让有更多邻居的节点具有更大的特征，因此再引入度矩阵 D，对邻接矩阵$(I+A)$的行和列进行归一化处理，且 $D_{ij}=\sum A_{ij}$ 为对角矩阵。这样每一个 GCN 层的输入为特征矩阵 $Z^{(j)} \in \mathbb{R}^{n \times l^{(j)}}$，输出特征矩阵为 $Z^{(j+1)} \in \mathbb{R}^{n \times l^{(j+1)}}$，可得不同 GCN 层间的传播规则如公式(1)所示：

$$Z^{(j+1)}=\sigma\left(D^{-\frac{1}{2}}(I+A)D^{-\frac{1}{2}}Z^{(j)}W^{(j)}\right) \tag{1}$$

其中，$w(j) \in \mathbb{R}^{l^{(j)} \times l^{(j+1)}}$ 为可学习的参数矩阵，$\sigma(\cdot)$ 为激活函数，$Z^{(0)}=X$。

本文构建的图模型为异构图，节点由几种不同类型的节点构成(句子节点、非结构化背景节点、结构化三元组节点)，但 GCN 模型不适合直接应用在异构图上。为了将 GCN 模型适用于本文提出的异构图，首先将不同类型的节点投影到一个相同的公共的空间中，如图 2 所示。

在公共空间中能够保持语义上的相似性以及去除掉更多的冗余信息。对于节点集合 V 中的不同类型的节点，$V=\{v_s, v_b, v_t\}$，考虑到不同类型节点的差异性，采用给不同类型的节点乘以不同的变换矩阵，最终将不同类型的节点映射到相同的空间 $\mathbb{R}^{l^{(j+1)}}$ 上，如公式(2)所示：

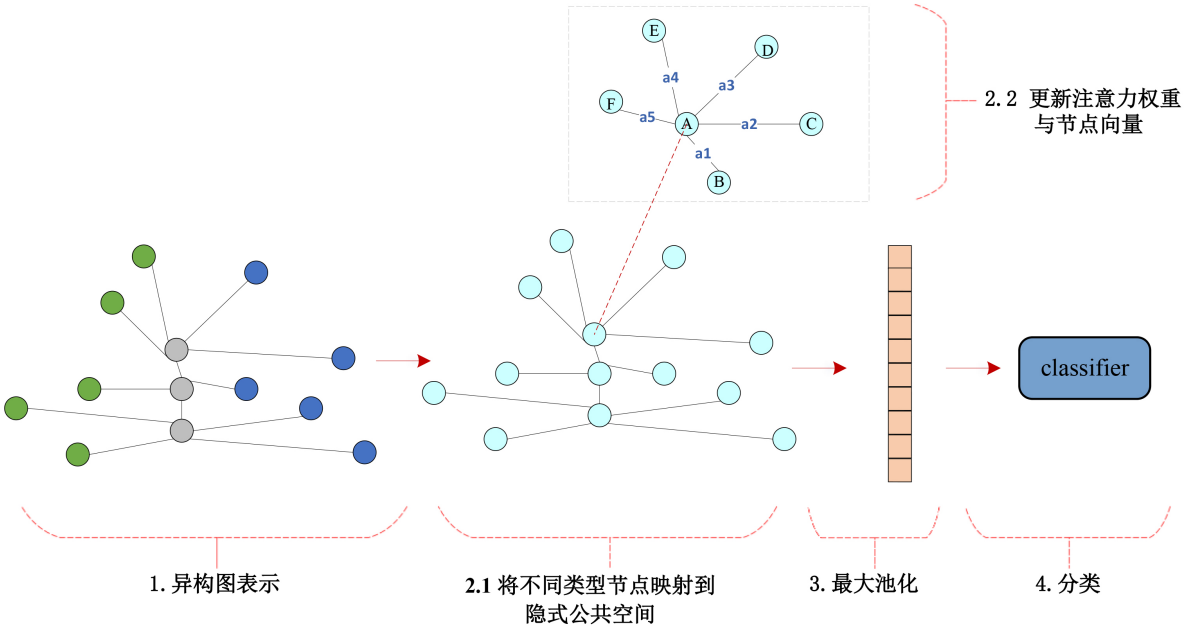

Figure 2. Diagram of news heterogeneous graph and knowledge graph
图 2. DGAT 模型的整体示意图

$$Z^{(j+1)} = \sigma\left(\max\left\{A_{v_s} Z^{(j)} W_{v_s}, A_{v_b} Z^{(j)} W_{v_b}, A_{v_t} Z^{(j)} W_{v_t}\right\}\right) \tag{2}$$

其中，A_v 是对应不同类型节点的选择矩阵，矩阵大小与 $Z^{(j)}$ 相同，A_v 的不同类型选择矩阵对应 $Z^{(j)}$ 所在位置的值为 1，其它位置值为 0，$Z^{(0)} = X$ 。最终将异构图的不同类型节点映射到相同的公共空间。

对于异构图中的一个节点，不同的邻居节点对自身特征向量的更新会有不同的权重，因此引入注意力机制赋予不同类型邻居节点不同的权重。上文构建的异构图融入了非结构化的背景信息与结构化的三元组信息，为了能够在训练中学习到语义冲突或者相符的地方帮助进行虚假新闻检测，采取根据给定节点与邻居节点之间的距离(如余弦相似度距离)计算注意力的权重值。令距离矩阵为 $S \in \mathbb{R}^{n\times n}$，注意力权重矩阵为 $\alpha \in \mathbb{R}^{n\times n}$，其中 s_{ij} 表示第 i 个节点与第 j 个节点之间的距离，α_{ij} 表示第 i 个节点与第 j 个节点之间的权重。

对于给定的第 i 个节点，利用余弦相似度计算节点之间的距离，即第 i 个节点与其他节点之间的距离表示为 $s_i = \{s_{i1}, si_2, \ldots, s_{in}\}$，最后将距离输入 *Soft*max 函数，以获得对应的权重值 α_i，如公式(3)所示：

$$\alpha_i = \frac{\exp\left(s_{ij}\right)}{\sum_{j=i}^{n} \exp\left(s_{ij}\right)} \tag{3}$$

对于异构图中的节点，不同的邻居信息会对自身的特征向量更新产生不同程度的影响。考虑语义相差较大的句子，在向量距离计算上，会有更远的距离。因此对于建模后的新闻异构图，在语义相差较大的地方，通过训练学习，赋予更加合理的权重，进行虚假新闻检测。使用最大池化作为分类前的最后一步操作。对于损失值的计算，采用交叉熵损失函数，如公式(4)所示：

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{P} Y_{ij} \cdot \log Z_{ij} \tag{4}$$

其中，N 为训练集中新闻的篇数，P 为新闻标签的种类数，Y 为新闻标签类别矩阵。

4. 实验

本节在数据集 Labeled Unreliable News Dataset (LUN) [16]上进行了大量实验。该数据集的标签有四种分类，分别为 Trusted、Satire、Hoax、Propaganda。将数据集按照 8:2 的比例划分成 LUN-train 训练集与 LUN-test 测试集，具体统计信息如表 1 所示。

Table 1. LUN dataset statistics used in the experiment
表 1. 实验中使用的 LUN 数据集统计信息

Dataset	Trusted (#Docs)	Satire (#Docs)	Hoax (#Docs)	Propaganda (#Docs)
LUN-train	9995	14,047	6942	17,870
LUN-test	750	750	750	750

4.1. 基线模型

选用神经网络模型 CNN [17]、LSTM [18]、BERT [19]，以及图神经网络模型 GCN [20]和 GAT [21]作为基线模型与本文的 DGAT 模型进行对比，基线模型的介绍设置如下：

- CNN 模型：使用一层 CNN 卷积层并将过滤器的尺寸设置为 3，通过最大池化层获得每个新闻的向量表示，再将新闻的向量表示传递给一个全连接的投影层，最终获得预测分类。
- LSTM 模型：使用一个 LSTM 层编码新闻的表示，并利用最后一个时间步的隐藏状态作为新闻特征向量，然后将其传递到一个全连接的投影层，最终获得预测分类。
- BERT 模型：使用 BERT 获取新闻中每个句子的向量表示，并在句子嵌入上应用 LSTM 层，最后使用投影层完成对每篇新闻的分类。
- GCN 模型：利用本文构建的新闻异构图，由于使用固定的邻接矩阵，在更新节点的向量表示时，利用的是邻居节点向量的加权求和。
- GAT 模型：在 GCN 模型的基础上，区别在更新节点向量表示时，使用注意力机制计算每个节点对于邻居节点的重要性，并根据重要性加权求和邻居节点的特征，从而更新节点的特征表示。

4.2. 超参设置

对于超参设置，模型中使用的所有隐藏维度都设置为 100，节点嵌入维数为 32。对于 GCN、GAT 和本文的 DGAT 模型，设置激活函数为 LeakyReLU，斜率为 0.2，并使用学习率为 0.001 的 Adam 优化器。此外所有模型的池化操作都采用最大池化。

4.3. 实验结果

选用 Micro-F1 与 Macro-F1 衡量本文的 DGAT 模型与基线模型的结果，如表 2 所示。

Table 2. Test results on the dataset
表 2. 在数据集上的测试结果

模型	Micro-F1	Macro-F1
CNN	53.92	51.36
LSTM	54.44	51.50
BERT	54.74	53.12
GCN	62.60	61.92
GAT	62.61	61.79
DGAT	**64.25**	**63.58**

可以看出本文提出的 DGAT 模型在 Micro-F1 与 Macro-F1 指标上都明显优于基线模型。与最佳基线模型相比，DGAT 模型将 Micro-F1 与 Macro-F1 都提高了近 2%，还可以发现基于图卷积网络的模型 GCN 和 GAT 的性能都优于包括 CNN、LSTM 和 BERT 在内的深度神经网络模型。由于图神经网络模型可以利用图中的节点进行信息的交流传递，而这对于虚假新闻检测很重要。本文的 DGAT 模型，能够获取到建模后新闻的语义相关性，根据语义的距离赋予合适的权重进行信息传递，为虚假新闻检测提供了检测依据。

5. 结论

在本文中，提出的融合新闻潜在的背景知识构建包含新闻句子节点、非结构化背景节点、结构化三元组节点的异构图，能够很好地建模新闻的表示。并在此基础上改进的图卷积网络模型 DGAT 模型表明了在异构图上能够将不同类型的节点映射到相同公共空间的可行性。此外基于距离的注意力机制能够捕获图中新闻内容与背景信息的一致性，用于虚假新闻检测。在公开的数据集上的实验证明了本文提出的方法的有效性。

基金项目

北京印刷学院博士启动资金－基于深度学习的虚假新闻检测关键技术研究(27170123034)；北京市教委科技计划一般项目(KM202110015003)。

参考文献

[1] Vosoughi, S., Roy, D. and Aral, S. (2018) The Spread of True and False News Online. *Science*, **359**, 1146-1151. https://doi.org/10.1126/science.aap9559

[2] Chen, Q., Zheng, Z. and Zhang (2020) Clinical Characteristics of 145 Patients with Corona Virus Disease 2019 (COVID-19) in Taizhou, Zhejiang, China. *Infection*, **48**, 543-551. https://doi.org/10.1007/s15010-020-01432-5

[3] Wang, Y., Qian, S., Hu, J., *et al*. (2020) Fake News Detection via Knowledge-Driven Multimodal Graph Convolutional Networks. *Proceedings of the* 2020 *International Conference on Multimedia Retrieval*, Dublin, 8-11 June 2020, 540-547. https://doi.org/10.1145/3372278.3390713

[4] Hu, L.M., Yang, T.C., Zhang, L.H., Zhong, W.J., Tang, D.Y., Shi, C., Duan, N. and Zhou, M. (2021) Compare to the Knowledge: Graph Neural Fake News Detection with External Knowledge. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing*, Volume 1, 754-763.

[5] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. and Yakhnenko, O. (2013) Translating Embeddings for Modeling Multi-Relational Data. *NIPS Conference*, Lake Tahoe, 5-8 December 2013, 2787-2795.

[6] Lu, Y.-J. and Li, C.-T. (2020) GCAN: Graph-Aware Co-Attention Networks for Explainable Fake News Detection on Social Media. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, 505-514.

[7] Jiang, S.Y., *et al*. (2019) User-Characteristic Enhanced Model for Fake News Detection in Social Media. 8*th CCF International Conference*, *NLPCC* 2019, Dunhuang, 9-14 October 2019, 634-646. https://doi.org/10.1007/978-3-030-32233-5_49

[8] Khoo, L.M.S., Chieu, H.L., Qian, Z. and Jiang, J. (2020) Interpretable Rumor Detection in Microblogs by Attending to User Interactions. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 8783-8790. https://doi.org/10.1609/aaai.v34i05.6405

[9] Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y. and Huang, J. (2020) Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 549-556. https://doi.org/10.1609/aaai.v34i01.5393

[10] Kang, Z.Z., *et al*. (2021) Fake News Detection with Heterogenous Deep Graph Convolutional Network. 25*th Pacific-Asia Conference*, *PAKDD* 2021, 11-14 May 2021, 408-420. https://doi.org/10.1007/978-3-030-75762-5_33

[11] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F. and Cha, M. (2016) Detecting Rumors from Microblogs with Recurrent Neural Networks.

[12] Shah, R., *et al*. (2019) SpotFake: A Multi-Modal Framework for Fake News Detection. *IEEE* 5*th International Conference on Multimedia Big Data* (*BigMM*), Singapore, 11-13 September 2019, 39-47.

[13] Zhou, X., Wu, J. and Zafarani, R. (2020) Similarity-Aware Multi-Modal Fake News Detection. 24*th Pacific-Asia Conference, PAKDD* 2020, Singapore, 11-14 May 2020, 354-367. https://doi.org/10.1007/978-3-030-47436-2_27

[14] Ferragina, P. and Scaiella, U. (2010) Tagme: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities) *Proceedings of the* 19*th ACM International Conference on Information and Knowledge Management*, Toronto, 26-30 October 2010, 1625-1628. https://doi.org/10.1145/1871437.1871689

[15] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks.

[16] Rashkin, H., Choi, E., Jang, J.Y., Volkova, S. and Choi, Y. (2017) Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 7-11 September 2017, 2931-2937. https://doi.org/10.18653/v1/D17-1317

[17] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, 25-29 October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[18] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[19] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 4171-4186.

[20] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations* (*ICLR*), Toulon, 24-26 April 2017, 1121-1129.

[21] Aswani, A.V., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 5998-6008.