中国科学: 信息科学 2023 年 第 53 卷 第 4 期: 663–681
SCIENTIA SINICA Informationis
论文

# 知识图谱驱动的图卷积神经网络谣言检测模型

徐凡[1], 李明昊[1], 黄琪[1], 鄢克雨[1], 王明文[1], 周国栋[2*]

1. 江西师范大学计算机信息工程学院, 南昌 330022
2. 苏州大学计算机科学与技术学院, 苏州 215006
* 通信作者. E-mail: gdzhou@suda.edu.cn

收稿日期: 2022–04–29; 修回日期: 2022–06–30; 接受日期: 2022–09–16; 网络出版日期: 2023–04–12

国家自然科学基金 (批准号: 62162031, 62076175, 62066020, 62266023) 和江西省杰出青年基金项目 (批准号: 20192ACBL21030) 资助

**摘要** 社交媒体谣言以极低的成本在互联网中被快速扩散, 给社会带来显著的负面影响. 传统的谣言检测模型主要考虑传播模式、写作风格、用户信用和世界知识等信息. 然而, 谣言的传播模式通常难以被捕捉, 写作风格却容易被模仿, 由元数据 (如职业、家乡、学历、年龄等) 构成的用户信息也容易被伪造. 本文提出了一种新颖的知识驱动的图卷积神经网络谣言检测模型. 该模型首先将社交媒体文本表示成一种语义 – 实体无向图结构, 其中节点包含原社交媒体文本中的词语, 利用世界知识库扩展的实体词语, 以及利用语言知识库扩展的语义词语, 边包含三类节点的 6 种有效组合. 该语义 – 实体图可以有效地增强任意两种节点的共现性, 从而丰富了原社交媒体文本的表示, 从一定程度上缓解数据稀疏共现问题. 语言知识利用了 HowNet (义原和同义词) 以及 WordNet (上义词、下义词和同义词) 分别对中英文社交媒体文本的主题词进行扩充. 并成功地将语言知识和实体知识通过图卷积神经网络框架有效集成. 在 4 个国际基准中英文谣言语料库上的实验结果和可视化分析表明了本文模型的有效性.

**关键词** 语言知识, 世界知识, 主题模型, 图卷积神经网络, 谣言检测

## 1 引言

Reuters 发布的 2019 年数字新闻报告 [1] 提到 38 个国家的人民对互联网信息的真假仍然表现出不同程度的担忧, 其中巴西国家人民怀疑网络谣言率高达 85%, 南非国家达 70%, 墨西哥达 68%, 法国达 67%. 虽然德国和荷兰相对较低, 但也分别各占 38% 和 31%. 此外, 有文献研究表明人们往往容易受到谣言的影响, 并加速其传播 [2]. 由此可见, 社交媒体已成为滋生谣言的温床. 互联网中散布的大量谣言不仅加剧社会恐慌和引发社会信任危机, 而且会损害国家形象和扭曲人们的意识形态, 进而会危

**引用格式:** 徐凡, 李明昊, 黄琪, 等. 知识图谱驱动的图卷积神经网络谣言检测模型. 中国科学: 信息科学, 2023, 53: 663–681, doi: 10.1360/SSI-2022-0170
Xu F, Li M H, Huang Q, et al. Knowledge graph-driven graph neural network-based model for rumor detection (in Chinese). Sci Sin Inform, 2023, 53: 663–681, doi: 10.1360/SSI-2022-0170

 

害到国家利益. 例如, 与 2020 年新型冠状病毒肺炎 (新冠肺炎) 疫情相关的谣言: “99.9% 的新冠病毒十分钟被茶水消灭”, “抗生素能有效预防和治疗新型冠状病毒” 等给防疫工作的有序开展带来了极大的阻碍.

当前, 已有文献 [3, 4] 对谣言检测任务进行了较为全面的综述. 例如, Zannettou 等 [3] 针对谣言、虚假新闻、恶作剧等不同类型的虚假信息, 从感知、动机、传播和检测模型 4 个方面进行了综述. 此外, Xu 等 [4] 针对虚假信息检测和真值发现任务进行了统一视角下的分析, 深入介绍了具有代表性的单模态和多模态虚假信息和真值发现在传播、用户、写作风格等方面的检测模型. 一般而言, 谣言的传播模式通常难以被捕捉, 写作风格 (词汇和句法) 却相对容易被模仿, 而构成用户信用的元数据 (例如, 职业、家乡、学历、年龄等) 很容易被伪造. 因此, 其他文献 [5∼10] 着重研究谣言检测模型中如何引入知识. 实际上, 知识在谣言检测中起着非常重要的指导作用, 人们通常根据外部知识 (如, 世界知识、语言知识或常识) 来判断给定社交媒体文本的真实性. 然而, 现有知识驱动的谣言检测模型都聚焦在如何挖掘世界知识 (比如: 三元组) 的作用. 像 WordNet 和 HowNet 等语言知识却很少被使用. 实际上, 以 WordNet 和 HowNet 为代表的词法语言知识能够辅助人们判断给定信息的真假性. 具体而言, WordNet [11] 构建了一个涵盖范围广泛的英语词汇语义网, 包括各种表示隐含关系的上下义关系以及同义词集. WordNet 不把单词分解成更小的有意义的单位. 相比于英文, 在人类语言的言语或写作过程中, 汉语词汇是独特而有意义的元素. 汉语词的意义可以用一组语义单位来表达. 语言学家将人类语言最小的语义单位定义为义原, 这有助于我们更好地理解人类语言. HowNet [12] 以还原论机制为基础, 强调义原所代表的部分和属性的重要性. 实际上, 这些上下义、同义和义原等语言知识可以辅助我们进行社交媒体谣言检测. 此外, 现有模型也忽略了语言知识和世界知识 (如实体) 在谣言检测方面的联合作用.

为清晰起见, 图 1 说明了中英文语言知识构成的一种层次结构情况. 对于给定的中文例子 “假借‘共享经济’ 吸金 102 亿, ‘鑫圆系’ 崩盘再次警示: 共享经济、区块链等新词极易被骗子包装利用. (类别: 非谣言)” 中出现的词语 “区块链”, 根据图 1(a) 所示, 汉语词语 “区块链” 在 HowNet 中可以被定义为 “知识”, 其同义词是 “金融科技”, 其义原表示为修饰词和领域 “利用, 处理, 问题, 具体, 金融”. 类似地, 对于英文例子 “US Democratic nominee Joe Biden declared victory in the 2020 presidential race on Saturday night, while US President Donald Trump refused to concede defeat. (类别: 非谣言)” 中出现的词语 “victory (胜利)”, 根据图 1(b) 所示, 英文单词 “victory (胜利)” 在 WordNet 中上义词是 “success (成功) 和 ending (结束)”, 其下义词是 “win (赢)、slam (大满贯)、independence (独立) 和 Pyrrhic victory (皮洛士式胜利)”, 其同义词是 “victory (获胜) 和 triumph (巨大成功)”. 实际上, 这种层次化语言知识结构也极大丰富了社交媒体文本中原词语的表示, 从而有助于谣言检测.

因此, 本文提出了一种新颖的知识驱动的谣言检测模型 KDRD (knowledge-driven rumor detection). 具体而言, 我们首先通过 LDA (latent Dirichlet allocation) 提取给定社交媒体文本的特定主题词, 并分别基于 WordNet 和 HowNet 获得英汉主题词的语义知识 (例如, 上下义词、同义词、义原). 然后, 我们对社交媒体文本进行实体链接, 并通过外部世界知识库获取扩展的实体知识. 接下来, 我们根据给定社交媒体文本的词语、扩展的语义知识和实体知识构建一个功能强大的语义 – 实体图 (如图 2 所示). 构建图 2 的动机包含两个方面: 其一, 在添加语义知识和实体知识后, 可以对给定社交媒体文本中任意两个词语之间的词共现进行增强, 丰富了给定社交媒体文本的表示, 从而在一定程度上缓解了数据稀疏共现性问题, 进而有利于谣言检测. 其二, 实体知识信息有助于扩充谣言实体中的信息量, 而语义知识信息能够更有效挖掘词语潜在的语义信息, 为谣言判别做信息补充. 通过构建语义 – 实体无向图, 可以丰富谣言的语义信息. 此外, 在图卷积神经网络框架下, 层次化语言知识与世界知识可以被

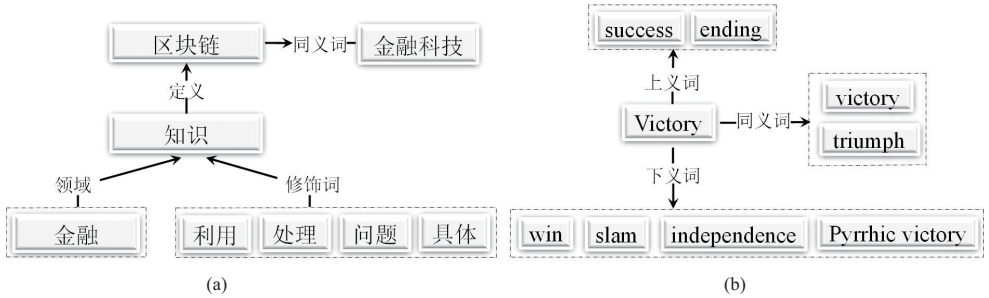


**图 1 给定词语对应的层次化语言知识实例. (a) 中文 HowNet 语言知识实例; (b) 英文 WordNet 语言知识实例**

**Figure 1** Samples of a given word with a hierarchical structure of language knowledge. (a) Sample of Chinese language knowledge in HowNet; (b) sample of English language knowledge in WordNet

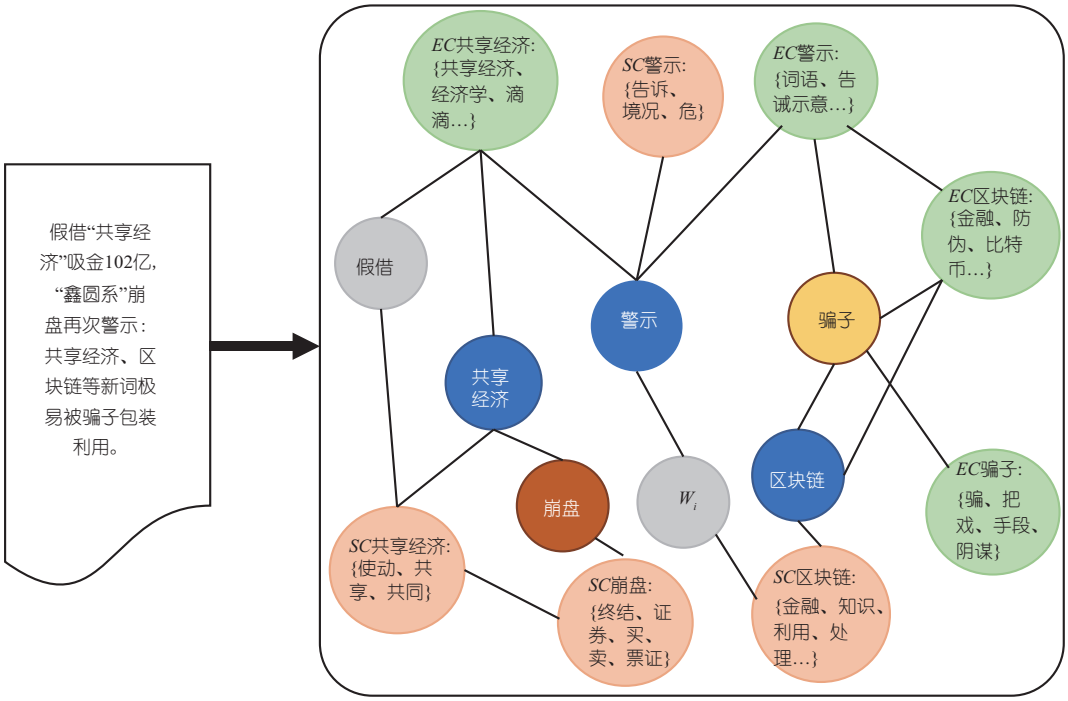


**图 2 语义 – 实体无向图. 备注: 节点包含对给定社交媒体文本删除停用词后的词语、浅绿色的扩展实体知识 (entity knowledge, 简称 ek) 和橙色的扩展语义知识 (semantic knowledge, 简称 sk). 灰色节点表示该词既不是实体词也不是主题词, 蓝色节点表示该词既是实体词又是主题词, 红色节点表示该词是主题词, 黄色节点表示该词是实体词. 各类边由词语共现信息构建**

**Figure 2** The semantic-entity undirected graph. Nodes involve the previous words for a given social media text after removing stop words, the extended entity knowledge (ek) in light green, and the extended semantic knowledge (sk) in orange. The gray node represents the word is non-entity and non-topic word, the blue node indicates the word is both an entity word and a topic word, the red node refers to the word is a topic word, and the yellow node denotes the word is an entity word. The edge is built with the word co-occurrence information

有机地集成. 在 4 个中英文国际基准谣言数据集上的实验结果表明了我们模型的有效性.

本文的贡献主要有以下两个方面.

(1) 本文构建了一个功能强大的语义 – 实体无向图, 其中节点包含社交媒体文本中的词语、扩充的实体知识和层次化语义知识, 边是 3 种类型节点的 6 种有效组合. 该同构类型的语义 – 实体无向图

极大丰富了给定社交媒体文本的表示, 因为可以有效地增强任意两种节点的词语共现性, 从而在一定程度上缓解了数据稀疏共现性问题.

(2) 在图卷积神经网络框架下, 层次化语言知识和世界知识被有效地结合起来, 发挥各自在谣言检测方面的互补作用. 此外, 本文将图卷积神经网络的隐层输出的最大池化拓展为平均池化, 发挥平均池化可以保留更多的全局信息作用. 本文在 4 个国际基准中英文谣言语料库上进行了大量实验, 实验结果表明本文模型显著优于现有代表性的谣言检测基准模型.

本文的其余部分组织如下. 第 2 节回顾了相关工作. 第 3 节详细地阐述了本文提出的知识驱动的谣言检测模型, 包含语义 – 实体无向图详细构建过程、图卷积神经网络建模等. 第 4 节详细介绍了实验设置及对比实验结果. 最后, 第 5 节对全文进行了总结.

## 2 相关工作

本节主要从代表性的谣言检测计算模型和图卷积神经网络两个方面介绍相关工作.

### 2.1 谣言检测模型

代表性的谣言检测模型可分为四类: 基于传播模式的方法、基于写作风格的方法、基于用户信用的方法和基于世界知识的方法.

**(1) 基于传播模式的方法.** 一般来说, 不同类型用户 (如普通用户和舆论领袖) 之间的互动模式或社交媒体源帖及其后续反馈等传播模式有助于谣言检测. 正如 Wu 等 [13] 中所述, 谣言传播机制表现为: 谣言通常由普通用户发布, 然后被一些意见领袖转发, 最后被大量普通用户转发. 然而, 非谣言信息通常由意见领袖发布, 然后由许多普通用户转发. 鉴于此, Wu 等提出了一个混合式核方法 [14] 下的虚假新闻检测模型. 相比之下, Ma 等 [15] 考虑了将社交媒体文本表示成树结构后, 着重分析了从根节点到子树的传播路径信息, 以便为虚假新闻检测融入更多的信息. Kumar 等 [16] 提出了一个多任务学习框架, 可以同时进行虚假新闻和立场检测. 此处, 立场是人们对个人、事物和事件的看法或态度, 如支持和反对等. Kumar 等基于 Tree-LSTM (long short-term memory, 长短时记忆) 模型, 有效融入了虚假新闻源帖及其后续反馈的传播关系. 类似的工作可以参考文献 [17~19]. 一般而言, 基于传播模式的谣言检测方法的难点在于如何观察并刻画这些不同的谣言传播模式.

**(2) 基于写作风格的方法.** 由于社交媒体谣言和非谣言往往存在着很多词汇和句法上差异, 一些学者从写作风格角度研究谣言检测模型. 具体来说, 有研究人员从文本的修辞结构角度进行虚假新闻检测 [5]. 修辞结构是语篇分析中广泛使用的一种篇章理论, 它描述了语篇的组成单元如何按照一定的关系组织成连贯完整的语篇. 根据修辞结构理论, 语篇单元之间的关系主要表现为核心 – 附属关系, 核心话语单元与附属话语单元之间存在着不同的话语关系. 这种话语关系可以有效融入谣言检测模型之中. 相比之下, 有些研究人员使用社交媒体文本中信号词的不确定性和熵比例来检测虚假新闻 [6]. Potthast 等 [8] 还设计了一些有效的手工提取特征 (例如, 字符 uni-gram, bi-gram 和 tri-gram, 停用词, 词性, 可读性, 词频, 引用词和外部链接的比例, 段落数以及文本的平均长度等) 来检测虚假新闻. Horne 等 [20] 采用了代表写作风格的其他手工提取特征 (例如, 名词数量、类型标记比率、字数和引号数量等) 进行虚假新闻检测. Li 等 [21] 利用 LIWC (linguistic inquiry and word count), POS (part-of-speech) 和 unigram 来进行垃圾邮件检测. 通常, 基于写作风格的方法缺点是这些手工创建的特征往往是一项比较耗时的工作.

**(3) 基于用户信用的方法.** 由于权威用户发布谣言的可能性比较低, 而普通用户则有很高的概率

发布或转发谣言. 因此, 有些学者研究在谣言检测中如何有效融入用户信用特征. Yang 等 [9] 提出了一种基于吉布斯抽样的方法来同时推断新闻的真实性和用户的可信度. Wang [22] 发布了一个基准虚假新闻数据集, 同时标注了用户的元数据 (例如, 党派关系、当前工作、国家和信用记录等), 并利用这些代表用户信用的元数据进行虚假新闻检测. Castillo 等 [23] 通过提取多种用户特征 (例如, 平均年龄、平均粉丝和朋友数、朋友是否有 URL 等) 来判断推文的真实性. 此外, Yuan 等 [24] 提出了一个联合模型来评估用户的可信度和进行虚假新闻检测, 该模型研究了如何整合发布人员和转载用户的可信度或声誉信息. Mukherjee 等 [25] 利用社区参与度 (例如, 答案数量、给出的评分、评论、收到的评分、分歧和评分者数量等)、用户间的一致性、典型观点或专业知识以及互动给用户可信度进行建模, 并进行虚假新闻检测. 类似的工作也可以参考文献 [26]. 一般而言, 基于用户信用的方法的缺点是代表用户信用的这些元数据很容易被伪造.

**(4) 基于世界知识的方法.** 众所周知, 外部知识在谣言检测中起到至关重要的作用. Hu 等 [10] 通过引入外部知识库, 将待检测新闻与知识库中信息进行比较, 并采用图神经网络结构进行虚假新闻检测. Ciampaglia 等 [27] 表明世界知识图谱中概念节点之间的最短路径信息可用于复杂的事实检查任务. Lao 等 [28] 提出融入世界知识图谱的路径排序算法 (path ranking algorithm) 下的虚假新闻检测模型. 相比之下, Shi 等 [29] 提出了一个基于链接预测的模型进行事实检查任务. 此外, Shiralkar 等 [30] 提出了一种无监督网络流方法, 用于预测以三元组 (例如, 主语、谓语和宾语) 形式表达的语句真实性. 类似地, Pan 等 [31] 提出了基于知识图谱嵌入的政治选举类虚假新闻检测模型, 他们提取政治选举类文本中的世界知识 (三元组), 然后通过比较三元组之间相似度值的大小关系进行虚假新闻检测. 通常, 外部世界知识库在谣言检测中的作用是显而易见的, 因为人们通常使用外部知识来判断给定社交媒体文本的真实性. 然而世界知识只是众多知识中的一种, 如何在谣言检测中有效融入语言知识的工作却比较鲜见.

### 2.2 图卷积神经网络

由于图卷积神经网络 [32] 能够有效捕获图结构中的高阶邻域信息, 因此其已被广泛应用于许多自然语言处理任务中, 如对话系统 [33]、文本分类 [34] 和虚假新闻检测 [10,35] 等. 例如, 文献 [34] 将文本表示成一个异构图, 节点是文档中的单词和文档本身, 边利用单词节点的互现和单词与文档的词频 – 逆文档频度 (term frequency-inverse document frequency, TF-IDF) 进行刻画, 从而获得单词和文档节点的表示. 文献 [10] 的作者为新闻文本构建了文档、实体、主题异构图, 并采用图卷积神经网络模型进行单模态 (文本) 虚假新闻检测. 而文献 [35] 也把新闻文本和图像表示成实体、图像异构图, 同样采用图卷积网络进行多模态 (文本和图像) 虚假新闻检测. 虽然异构图融入了更多类型的信息, 但异构图的计算量非常大.

因此, 与上述谣言检测模型不同, 我们提出了一种新颖的图神经网络谣言检测模型, 该模型除了能有效融入世界知识图谱信息外, 还能有效集成层次化语言知识信息 (上下义词、同义词和义原), 通过图卷积神经网络对我们构建的同构语义 – 实体无向图进行建模, 充分发挥语言知识和世界知识在语言和实体方面的互补优势.

## 3 模型

本节主要介绍谣言检测问题定义、语义 – 实体无向图构建和图卷积神经网络建模过程.

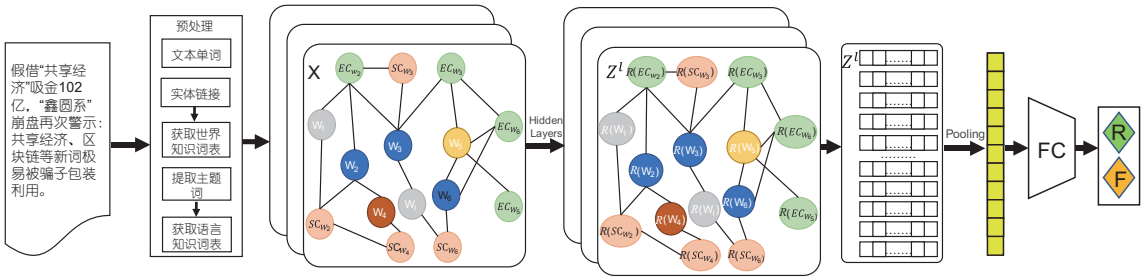


**图 3　本文提出的 KDRD 模型框架. FC 代表全连接, R 代表不是谣言, F 代表谣言**

**Figure 3**　The framework of our KDRD. FC refers to full connection; R denotes a non-rumor; F indicates a rumor

### 3.1　问题定义

我们将每篇社交媒体文本 $n$ 构建一个无向图 $G(n)=(V,E)$, 其中 $V=\{w_{\mathrm{up}},w_{\mathrm{ek}},w_{\mathrm{sk}}\}$ 和 $E=\{E_{w_{\mathrm{up}},w_{\mathrm{up}}},E_{w_{\mathrm{up}},w_{\mathrm{ek}}},E_{w_{\mathrm{up}},w_{\mathrm{sk}}},E_{w_{\mathrm{ek}},w_{\mathrm{ek}}},E_{w_{\mathrm{ek}},w_{\mathrm{sk}}},E_{w_{\mathrm{sk}},w_{\mathrm{sk}}}\}$ 分别表示图中的节点集合和边集合. $w_{\mathrm{up}}$ 表示对候选社交媒体文本通过分词和去除停用词后获得的词语集合. ek (entity knowledge) 表示对候选社交媒体文本通过实体链接到外部知识库而获得的实体词语集合. sk (semantic knowledge) 表示对候选社交媒体文本中主题词的层次化语义词语集合. 六种类型的边 ($E_{w_{\mathrm{up}},w_{\mathrm{up}}}$, $E_{w_{\mathrm{up}},w_{\mathrm{ek}}}$, $E_{w_{\mathrm{up}},w_{\mathrm{sk}}}$, $E_{w_{\mathrm{ek}},w_{\mathrm{ek}}}$, $E_{w_{\mathrm{ek}},w_{\mathrm{sk}}}$, $E_{w_{\mathrm{sk}},w_{\mathrm{sk}}}$) 是上述三种类型节点的有效组合. 因此, 社交媒体谣言检测问题可以归结为图分类问题. 也就是给定一组无向图 $G=\{G_1,G_2,\ldots,G_n\}$, 其中每篇候选社交媒体文本 $n$ 被形式化为语义 – 实体无向图, 我们的目标是学习函数 $f:G\to L$, 其中 $G$ 表示图的输入空间, $L$ 表示谣言标签集, 包括谣言和非谣言.

### 3.2　模型阐述

图 3 显示了本文提出的融合知识图谱的谣言检测模型 (限于篇幅, 图 3 中语义 – 实体无向图仅采用图 2 中的抽象表示形式), 模型包含语义 – 实体无向图构建和图卷积神经网络建模两个主要模块. 下面分别描述.

#### 3.2.1　语义 – 实体无向图构建

对于给定的中英文社交媒体文本 (如, 推特、微博或微信等), 我们首先对该文本去除停用词[1)], 并采用 Jieba 工具包[2)] 进行中文文本分词, 生成更新后的文档 $D_{\mathrm{up}}$, 其中该文档中的词为 $w_{\mathrm{up}}$. 由于主题信息可以帮助判别是否为谣言 [36], 因此我们还提取了文档 $D_{\mathrm{up}}$ 的主题词, 以丰富社交媒体文本的表示. 我们采用 LDA 工具[3)] 和 Gensim 工具[4)] 分别提取英语和汉语社交媒体文本中的主题词, 生成 top-k 个主题词集 $w_{\mathrm{to}}$. 边的构建主要通过 PMI 来计算词节点之间的共线关系, 针对三类不同的节点 $w_{\mathrm{up}}$, $w_{\mathrm{ek}}$ 和 $w_{\mathrm{sk}}$, 一共存在 6 种边关系. 我们把语义 – 实体无向图中的边分为一对一映射, 一对多映射, 多对多映射三种情形. 通过三种映射 PMI 的计算公式得到潜在边的权重. 在判定时, 如果 PMI 值大于 0, 则将 PMI 值作为边的权重; 如果 PMI 值小于等于 0, 则将边的权重判定为 0, 即不存在边.

**(1) 一对一映射情形.** 对于 $E_{w_{\mathrm{up}},w_{\mathrm{up}}}$, 我们直接采用 PMI (pointwise mutual information) [37] 来计

1) https://blog.csdn.net/u012661010/article/details/70880847/.
2) https://github.com/fxsjy/jieba.
3) https://lda.readthedocs.io/en/latest/index.html.
4) https://radimrehurek.com/gensim/.

算此类边的权值, 如下所示:

$$\mathrm{PMI}(i,j)=\log\frac{p(i,j)}{p(i)p(j)}, \tag{1}$$

$$p(i,j)=\frac{\#W(i,j)}{\#W}, \tag{2}$$

$$p(i)=\frac{\#W(i)}{\#W}, \tag{3}$$

其中 $W(i,j)$ 代表滑动窗口中同时包含词语 $i$ 和 $j$ 的个数, $W$ 代表滑动窗口中的词语个数, $W(i)$ 代表滑动窗口中包含词语 $i$ 的个数.

**(2) 一对多映射情形.** 对于 $E_{w_{\mathrm{up}},w_{\mathrm{sk}}}$ 和 $E_{w_{\mathrm{up}},w_{\mathrm{ek}}}$, 我们首先计算 $w_{\mathrm{up}_i}$ 和 $w_{\mathrm{sk}}$ 或 $w_{\mathrm{ek}}$ 中每个词语的 PMI, 然后求平均值. 以 $w_{\mathrm{up}_i}$ 和 $w_{\mathrm{sk}}$ 的计算过程为例, $w_{\mathrm{up}_i}$ 和 $w_{\mathrm{ek}}$ 的计算类似, 如下所示:

$$\mathrm{PMI}(w_{\mathrm{up}_i},w_{\mathrm{sk}_j})=\frac{\sum_{k=1}^{n}\mathrm{PMI}(w_{\mathrm{up}_i},w_{\mathrm{sk}_j^k})}{n}, \tag{4}$$

其中 $n$ 是 $w_{\mathrm{sk}_j}$ 中总单词个数.

**(3) 多对多映射情形.** 对于 $E_{w_{\mathrm{sk}},w_{\mathrm{ek}}}$, $E_{w_{\mathrm{ek}},w_{\mathrm{ek}}}$ 和 $E_{w_{\mathrm{sk}},w_{\mathrm{sk}}}$, 我们首先计算 $w_{\mathrm{sk}}$ 和 $w_{\mathrm{ek}}$ 中每个词语的 PMI, 然后求平均值, 如下所示:

$$\mathrm{PMI}(w_{\mathrm{sk}_i},w_{\mathrm{ek}_j})=\frac{\sum_{h=1}^{m}\sum_{k=1}^{n}\mathrm{PMI}(w_{\mathrm{sk}_i^h},w_{\mathrm{ek}_j^k})}{m\times n}, \tag{5}$$

其中, $m$ 是 $w_{\mathrm{sk}_i}$ 中词语个数, $n$ 是 $w_{\mathrm{ek}_j}$ 中词语个数.

**语义词语节点生成.** 与世界知识不同, 语言知识也构成了社交媒体文本表示的一个重要方面. 语义知识是 LDA 对去除停用词的文档获取主题词后, 通过语言知识库获取. 针对英语文本, 我们提取经过 LDA 操作后的 $w_{\mathrm{to}}$ (topic words) 中每个词语的上义词、下义词和同义词, 生成一个语义知识词语集 $w_{\mathrm{sk}}$. 类似地, 针对汉语文本, 我们也为 $w_{\mathrm{to}}$ 中每个汉语词语获取义原和同义词, 生成对应的 $w_{\mathrm{sk}}$. 为清晰起见, 图 4 显示了对给定社交媒体文本句子 “假借 ‘共享经济” 吸金 102 亿, ‘鑫圆系’ 崩盘再次警示: 共享经济、区块链等新词极易被骗子包装利用. ” 中汉语词 “区块链” 的语义词语集构造过程. 我们查询 “区块链” 在 HowNet 中的义原: “knowledge | 知识: domain = finance | 金融, use | 利用: purpose = handle | 处理, patient = problem | 问题, modifier = concrete | 具体” 以及同义词 “金融科技”, 得到 $w_{\mathrm{sk}}=\{$知识, 金融, 利用, 处理, 问题, 具体$\}$.

限于篇幅, 对于给定英文社交媒体文本句子 “US Democratic nominee Joe Biden declared victory in the 2020 presidential race on Saturday night, while US President Donald Trump refused to concede defeat.” 中词语 “Victory”, 利用 WordNet 可获得它的上义词 “ending, success”, 下义词 “win, independence, landside, slam, Pyrrhic victory” 和同义词 “Victory, trimpion”, 并得到 $w_{\mathrm{sk}}=\{$ending, success, win, independence, landside, slam, Pyrrhic victory, trimpion$\}$.

在为 $D_{\mathrm{up}}$ 中给定的词语 $w_i$ 提取语义知识词语 $w_{\mathrm{sk}}$ 之后, 我们使用 word2vec 工具来获得 $w_{\mathrm{sk}}$ 的向量 $\mathrm{Vector}_{\mathrm{sk}}$. 为了生成更强大的语义信息, 我们对 $\mathrm{Vector}_{\mathrm{sk}}$ 分别执行平均和最大池化操作, 以丰富特定主题词的表示, 如下所示:

$$\mathrm{Vector}_{\mathrm{sk\_avg(区块链)}}=\text{average pooling}\left(V_{\text{知识}},V_{\text{金融}},V_{\text{利用}},V_{\text{处理}},V_{\text{问题}},V_{\text{具体}}\right), \tag{6}$$

$$\mathrm{Vector}_{\mathrm{sk\_max(区块链)}}=\max\text{ pooling}\left(V_{\text{知识}},V_{\text{金融}},V_{\text{利用}},V_{\text{处理}},V_{\text{问题}},V_{\text{具体}}\right), \tag{7}$$

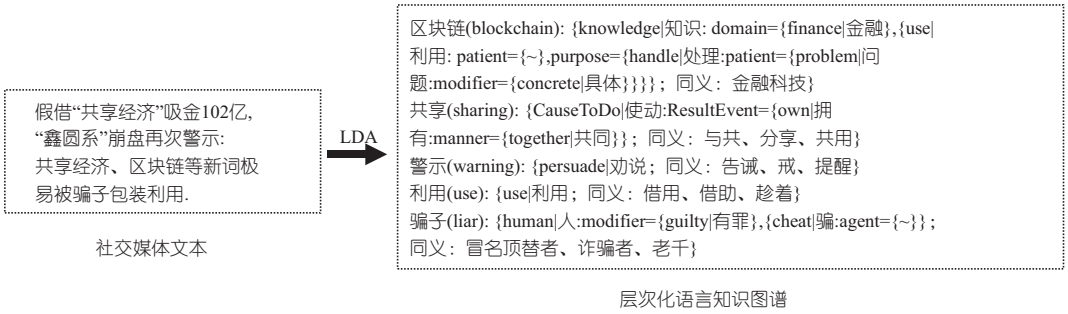


**图 4　中文层次化语言知识实例**

**Figure 4**　A Chinese sample of hierarchical language knowledge

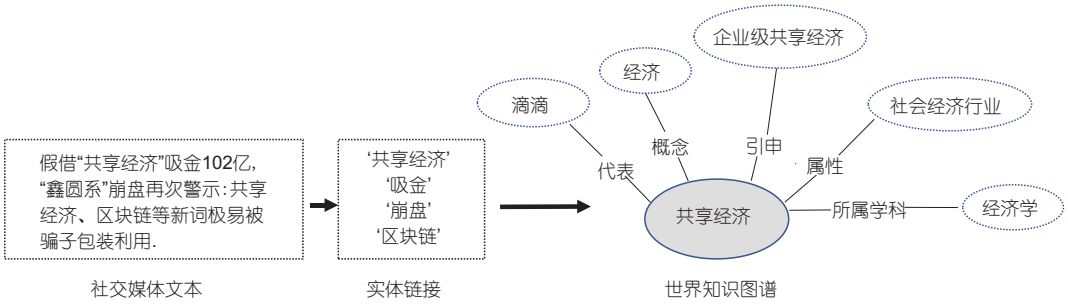


**图 5　(网络版彩图) 中文实体链接实例**

**Figure 5**　(Color online) A Chinese sample of entity linking

其中, $V$ 代表语义知识词语向量.

**实体词语节点生成.** 实体链接是将文本中出现的实体提及 (mention) 链接到其在知识库中对应的实体过程. 实体链接可以将社交媒体文本中已识别的提及 (mentions) 正确地指向知识库中的目标实体. 在英文数据集下, 本文采用实体链接工具 TAGME [5], 可以提取出社交媒体文本存在的提及 (mentions), 并将这些 mentions 链接到英文类型的 Wiki 世界知识库中已经存在的实体词. 在中文数据集下, 本文采用中文 CN_DBpedia [6] 世界知识库提供的实体链接的 API 直接将社交媒体文本中识别出的 mentions 和 CN_DBpedia 知识库中已经存在的实体词对应起来. 为清晰起见, 图 5 显示了对给定社交媒体文本句子 “假借 ‘共享经济’ 吸金 102 亿, ‘鑫圆系’ 崩盘再次警示: 共享经济、区块链等新词极易被骗子包装利用. ” 中汉语词语 “共享经济” 的实体词语生成过程. 我们可以获得 $w_{\mathrm{ek}} = \{$经济, 滴滴, 经济学, 社会经济行业, 企业级共享经济$\}$.

限于篇幅, 对于给定英文社交媒体文本句子 “US Democratic nominee Joe Biden declared victory in the 2020 presidential race on Saturday night, while US President Donald Trump refused to concede defeat.” 中词语英文单词 “Joe Biden”, 我们可以提取其实体词语 $w_{\mathrm{ek}} = \{$美国第 46 任总统约, 瑟夫 · 罗比内特 · 拜登, 锡拉丘兹大学, 政治家, 民主党等$\}$. 事实上, 这些扩展的实体词也极大丰富了社交媒体文本的表现形式. 事实上, 我们还尝试了两跳实体词. 与一跳实体词相比, 二跳实体词比较抽象, 从而带来更多的噪声.

类似地, 我们使用 word2vec 来获得 $w_{\mathrm{ek}}$ 的向量表示 $\mathrm{Vector}_{\mathrm{ek}}$. 为了生成更强大的实体信息, 与

5) https://pypi.org/project/tagme/.
6) http://kw.fudan.edu.cn/apis/cndbpedia/.

式 (6) 和 (7) 类似, 我们同样对 $\text{Vector}_{\text{ek}}$ 执行平均和最大池化操作, 以丰富特定实体的表示.

#### 3.2.2 图卷积网络模型

受文献 [35] 提出的多模态 (文本和图像) 卷积神经网络虚假新闻检测模型启发, 该文献作者首先将新闻文本中的单词和文本包含的图像构成一个异构图, 然后利用图卷积神经网络对该异构图进行建模, 接着对图神经网络隐层的输出进行最大池化操作, 最后利用 Softmax 对异构图进行分类. 由于异构图计算量较大, 而且本文主要聚焦在单模态 (文本) 谣言检测任务. 因此, 本文在文献 [35] 的单模态 (文本) 卷积神经网络虚假新闻检测模型 (KMGCN-NoKDVisual) 上进行改进. 主要改进体现在 3 个方面: 一是我们采用更丰富的 Wiki 知识库去替换文献 [35] 仅采用的 "isA" 关系知识库. 二是除扩展的实体库外, 本文融入更丰富的语义信息 (例如, 上下义词、同义词、义原), 并将社交媒体文本表示成同构形式的语义 – 实体图, 从而简化后续图卷积神经网络的计算量. 三是我们将图卷积神经网络的隐层输出分别采用平均和最大池化操作, 实际上文献 [35] 采用的最大池化仅保留了局部信息, 而平均池化可以保留更多的全局信息, 并通过实验验证了两种池化操作对谣言检测性能的影响.

一般而言, 图卷积神经网络是一种多层神经网络, 它直接对一个图进行卷积操作并产生一个嵌入向量. 对于上述构建的语义 – 实体无向图 $G=(V,E)$, 其中 $V$ ($|V|=m+n+k$) 和 $E$ 分别是 3 种类型的节点和 6 种类型的边集合. 设 $X\in R^{(m+n+k)\times d}$ 是一个包含所有节点及其特征的矩阵, 其中 $m$ 是删除停用词后社交媒体文本中的总单词数, $n$ 是语义知识 sk 的词集的总个数, $k$ 是实体知识 ek 的词集的总个数, $d$ 是特征向量的维数.

针对构建的语义 – 实体无向图, 我们首先采用两层图卷积神经网络来获得其隐藏状态 $Z^{l+1}$ (见式 (8)). 然后, 我们对 $l+1$ 层的输出 $Z^{l+1}$ 进行平均池化操作, 获得社交媒体文本的表示, 然后经过一个全连接层和一个 Softmax 激活函数, 最终得到一个社交媒体文本的二元分类器. 图卷积神经网络的第一层和第二层采用的激活函数均为 ReLU. 图卷积神经网络在每一层对所有节点的表示会进行全局更新.

$$Z^{l+1}=\sigma(D^{\frac{1}{2}}(I+A)D^{-\frac{1}{2}}Z^{l}W), \tag{8}$$

其中, $A$ 为邻接矩阵 (见式 (9)), $D$ 为度矩阵, $I$ 为单位矩阵, $\sigma$ 为激活函数, $W$ 为参数.

$$A(i,j)=\begin{cases}\text{PMI}(i,j), & \text{PMI}(i,j)>0,\\ 0, & \text{PMI}(i,j)\leqslant 0.\end{cases} \tag{9}$$

我们采用 $r_i$ 作为社交媒体文本 $i$ 的预测值:

$$r_i=\text{Softmax}(Z_{\text{mean}}), \tag{10}$$

其中, $Z_{\text{mean}}$ 是 $Z^l$ 进行平均池化操作后的输出.

我们采用交叉熵作为损失函数, 并采用 Adam 算法 [38] 进行优化. 与文献 [32,34,39] 的实验结果一样, 我们发现两层图卷积神经网络的性能优于一层图卷积神经网络, 而多层图卷积神经网络并不能提高谣言检测的性能.

## 4 实验

本节介绍实验所采用的数据集、基准模型、评测指标、实验设置以及中英文谣言检测实验结果分析.

表 1 语料库统计数据

**Table 1** Statistics of the benchmark corpora

| | Pheme | Twitter | WeChat | Weibo |
|---|---|---|---|---|
| #Real | 1972 | 7898 | 1990 | 4749 |
| #Fake | 3830 | 6026 | 1990 | 4779 |

### 4.1 数据集

为了验证本文模型的有效性, 我们采用 4 个国际基准谣言检测语料库. 其中, 包含英文 Pheme [7] 和 Twitter [8], 以及中文 WeChat [9]和 Weibo [40]. 表 1 显示了 4 个语料库的统计数据, 其中 #Real 代表不是谣言的数量, #Fake 代表谣言的数量.

### 4.2 基准模型

**(1) BOW_SVM.** 该基准模型将社交媒体文本表示成词袋形式, 并采用 SVM 作为分类器.

**(2) TextGCN [34].** TextGCN 是当前流行的将单词和文档表示进行联合学习的框架. 我们直接采用公开的源代码[10] 进行谣言检测.

**(3) KMGCN-NoKDVisual [35].** KMGCN-NoKDVisual 模型为每一篇虚假新闻构造一个无向图 (节点是文本中的单词, 边反映了单词间的共现性), 并采用图卷积神经网络检测虚假新闻. 我们重现了他们的工作, 并在 4 个数据集上进行谣言检测.

**(4) KCNN [41].** KCNN 是一种融入知识的卷积神经网络谣言检测模型. 该模型的输入包含三部分: 文本嵌入、实体嵌入和实体嵌入. 我们直接采用公开的源代码[11] 进行谣言检测.

**(5) CompareNet [10].** CompareNet 是一种端到端图神经网络谣言检测模型, 该模型用于将文本与外部知识库进行比较以检测谣言. 我们直接采用公开的源代码[12] 进行谣言检测.

**(6) BERT [42].** BERT 作为一种基于 Transformer 的预训练模型, 其在自然语言处理多种任务中均取得了很好的性能. 我们直接采用公开的源代码[13], 通过微调进行谣言检测.

### 4.3 实验设置

在本实验中, PMI 的滑动窗口大小设置为 20, 原文词、语义词语以及实体词语表示都利用 word2vec 初始化 (本文同时采用 BERT 词向量作为初始化), 词语向量维度设置为 100. 图卷积神经网络第一层的输出维度设置为 64, 其第二层的输出维度设置为 32. batch size 为 128, epoch 为 200 次, 学习率设置为 0.001, dropout 设置为 0.5. 我们采用随机梯度下降法对模型进行更新, 采用 Adam 算法对模型进行优化, 将英文和中文社交媒体文本的主题词总数分别设置为 10 和 5. 在 Pheme 数据集上, 我们采用 5- 倍交叉验证, WeChat 数据集采用默认的划分. 在 Weibo 和 Twitter 数据集上, 随机选取了 10% 的数据作为验证集, 剩下的数据以 4:1 随机划分训练集和测试集. 本文采用 4 个常用的评测指标进行谣言检测评测, 包含: Accuracy, Precision, Recall 和 F1 值.

7) https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619.
8) https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2016.
9) https://github.com/yaqingwang/WeFEND-AAAI20.
10) https://github.com/yao8839836/text_gcn.
11) https://github.com/hwwang55/DKN.
12) https://github.com/ytc272098215/FakeNewsDetection.
13) https://huggingface.co/bert-base-uncased.

表 2 英文谣言检测性能

**Table 2** Fake news detection results on the English datasets[a)]

| | Accuracy | Fake | | | Real | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 |
| Pheme | | | | | | | |
| BOW_SVM | 0.6437 | 0.5621 | 0.5433 | 0.5526 | 0.6832 | 0.6918 | 0.6875 |
| KCNN | 0.6725 | 0.6265 | 0.6491 | 0.6427 | 0.7481 | 0.7312 | 0.7396 |
| TextGCN | 0.8276 | 0.7652 | 0.7343 | 0.7494 | 0.8568 | 0.8607 | 0.8597 |
| KMGCN-NoKDVisual | 0.8395 | 0.7767 | 0.7566 | 0.7665 | 0.8604 | 0.8632 | 0.8618 |
| BERT | 0.8192 | 0.8091 | 0.7266 | 0.7560 | 0.8238 | 0.8599 | 0.8415 |
| CompareNet | 0.8734 | 0.8372 | 0.7916 | 0.8138 | 0.8864 | **0.9256** | 0.9056 |
| KDRD_word2vec (ours) | **0.8822*** | **0.8612** | 0.7918 | **0.8250*** | 0.8936 | 0.9217 | **0.9074*** |
| KDRD_BERT (ours) | 0.8764* | 0.8475 | **0.8021** | 0.8242* | **0.8997** | 0.9007 | 0.9001* |
| Twitter | | | | | | | |
| BOW_SVM | 0.5451 | 0.4980 | 0.4973 | 0.4976 | 0.5532 | 0.5564 | 0.5548 |
| KCNN | 0.6354 | 0.7492 | 0.6257 | 0.6819 | 0.5894 | 0.7194 | 0.6479 |
| TextGCN | 0.7128 | 0.8182 | 0.4259 | 0.5602 | 0.6680 | **0.9392** | 0.7807 |
| KMGCN-NoKDVisual | 0.7543 | 0.7821 | 0.7326 | 0.7565 | 0.7571 | 0.8165 | 0.7857 |
| BERT | 0.8055 | **0.8315** | 0.8162 | **0.8238** | 0.7657 | 0.7842 | 0.7748 |
| CompareNet | 0.8124 | 0.7874 | **0.8271** | 0.8068 | 0.8232 | 0.7865 | 0.8044 |
| KDRD_word2vec (ours) | 0.8231* | 0.8263 | 0.8088 | 0.8175 | 0.8184 | 0.8381 | **0.8281**** |
| KDRD_BERT (ours) | **0.8269*** | 0.8166 | 0.8000 | 0.8082 | **0.8414** | 0.8298 | **0.8356**** |

a) KDRD_word2vec indicates the vector initialization of the word using word2vec; KDRD_BERT stands for the vector initialization of the word using BERT; * refers to the p-value corresponding to the comparison significance test of the reference system <0.05; ** denotes the p-value corresponding to the comparison significance test of the reference system <0.01.

### 4.4 实验结果及分析

本小节分别介绍英文和中文谣言检测结果、实验数据分析以及模型可解释性分析, 通过实验结果主要验证: (1) 本文采用的层次化语言知识 (例如, 上义词、下义词、同义词、义原) 的有效性; (2) 层次化语言知识与世界知识消融及组合的有效性; (3) 扩展语义上下文和实体知识所采用的不同融合策略 (例如, 平均和最大池化) 的影响.

#### 4.4.1 英文谣言检测实验结果及分析

表 2 和 3 分别显示了英文谣言检测实验结果和详细的消融实验结果. 我们可以得出结论: 除了在 Twitter 数据集的 Fake 类型外, 本文模型在 Accuracy 和 F1 值方面表现最佳.

此外, 我们还可以得到如下结论.

(1) 在模型层面, BOW_SVM 在所有 4 个评测指标中表现都最差, 说明了简单的词袋模型在谣言检测中作用较小. 然而, 融入知识后, 谣言检测性能得以提升. 由于 KCNN 模型只包含浅层知识 (例如, 新闻嵌入和实体嵌入), 因此其性能提升有限. 由于 TextGCN 中文本的强大表示功能, 其性能优于 BOW_SVM 和 KCNN. 相比于采用异构图形式的 TextGCN 模型, 采用同构图形式的 KMGCN-

**表 3　英文数据集上消融实验结果**

**Table 3**　Results of the ablation experiment on the English datasets

| | Accuracy | Fake | | | Real | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 |
| Pheme | | | | | | | |
| entity knowledge | 0.8463 | 0.7825 | 0.7631 | 0.7727 | 0.8742 | 0.8863 | 0.8802 |
| entity knowledge+hypernym | 0.8526 | 0.7936 | 0.7742 | 0.7838 | 0.8786 | 0.8868 | 0.8827 |
| entity knowledge+hyponym | 0.8597 | 0.8021 | 0.7813 | 0.7916 | 0.8834 | 0.8912 | 0.8873 |
| entity knowledge+synset | 0.8686 | 0.8349 | 0.7884 | 0.8110 | 0.8837 | 0.9035 | 0.8935 |
| entity knowledge +hyponym+synset | 0.8712 | 0.8396 | 0.7887 | 0.8134 | 0.8872 | 0.9143 | 0.9005 |
| entity knowledge +hypernym+hyponym | 0.8648 | 0.8295 | 0.7792 | 0.8036 | 0.8792 | 0.9033 | 0.8911 |
| entity knowledge +hypernym+hyponym+synset | 0.8744 | 0.8408 | 0.7915 | 0.8154 | 0.8922 | 0.9098 | 0.9009 |
| entity knowledge +hyponym+synset | **0.8822** | **0.8612** | **0.7918** | **0.8250** | **0.8936** | **0.9217** | **0.9074** |
| Twitter | | | | | | | |
| entity knowledge | 0.7780 | 0.7991 | 0.7248 | 0.7601 | 0.7890 | 0.8312 | 0.8096 |
| entity knowledge+hypernym | 0.8086 | 0.8186 | 0.7843 | 0.8011 | 0.7934 | 0.8297 | 0.8111 |
| entity knowledge+hyponym | 0.8102 | 0.8206 | 0.7851 | 0.8025 | 0.7916 | 0.8298 | 0.8102 |
| entity knowledge+synset | 0.8100 | 0.8217 | 0.7955 | 0.8084 | 0.7983 | 0.8210 | 0.8095 |
| entity knowledge +hypernym+hyponym | 0.8122 | 0.8192 | 0.7851 | 0.8079 | 0.8013 | 0.8344 | 0.8175 |
| entity knowledge +hypernym+synset | 0.8170 | 0.8216 | 0.7867 | 0.8038 | 0.8004 | 0.8385 | 0.8190 |
| entity knowledge +hyponym+synset | 0.8218 | 0.8251 | 0.7997 | 0.8122 | 0.8183 | 0.8367 | 0.8274 |
| entity knowledge +hypernym+hyponym+synset | **0.8231** | **0.8263** | **0.8088** | **0.8175** | **0.8184** | **0.8381** | **0.8281** |

NoKDVisual 模型性能更好, 潜在原因可能在于谣言检测的文档一般比较短小, 单词和文档节点的表示不太适合采用 TF-IDF 计算. TextGCN 在 Twitter 数据集中真新闻类别上取得了不错的 Recall 值, 潜在原因在于 TextGCN 在语料库规模相对较大的 Twitter 中计算单词和文档节点的 TF-IDF 值相对精确. 但是, TextGCN 偏向于将社交媒体文本识别成真新闻类别. 由于 TextGCN 所取得的 Precision 值较低, 导致其整体谣言检测性能一般. 此外, CompareNet 在所有 4 个评测指标上的性能都优于 KCNN, 因为 CompareNet 采用了一种融合世界知识的图神经网络模型. 然而, 相比于 BOW_SVM 和 KCNN, BERT 模型也取得了不错的谣言检测性能.

(2) 在不同语义知识类型层面, 总体而言, WordNet 的下义词比上义词表现得更好. 因为下义词为特定的英语单词提供了详细的语义解释, 从而有助于谣言检测. 相比之下, 同义词则更为抽象, 它只解释了特定单词的公共属性. 然而, 上下义词和同义词的部分组合也能进一步提高谣言检测性能.

**表 4 向量平均池化下英文数据集上图卷积层数实验结果**

**Table 4** Experimental results with different convolution layers under the average pooling mode on English datasets

| | Accuracy | Fake | | | Real | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 |
| | | | | Pheme | | | |
| KDRD (1 layer) | 0.8024 | 0.7324 | 0.7113 | 0.7217 | 0.8317 | 0.8328 | 0.8322 |
| KDRD (2 layers) | **0.8822** | **0.8612** | **0.7918** | **0.8250** | **0.8936** | 0.9217 | **0.9074** |
| KDRD (3 layers) | 0.8701 | 0.8579 | 0.7804 | 0.8173 | 0.8732 | 0.9192 | 0.8956 |
| | | | | Twitter | | | |
| KDRD (1 layer) | 0.7332 | 0.7621 | 0.7037 | 0.7317 | 0.7194 | 0.7603 | 0.7393 |
| KDRD (2 layers) | **0.8231** | **0.8263** | **0.8088** | **0.8175** | **0.8184** | **0.8381** | **0.8281** |
| KDRD (3 layers) | 0.8156 | 0.8194 | 0.7901 | 0.8045 | 0.8019 | 0.8299 | 0.8157 |

(3) 在标签类别层面, 总体而言, Real (不是谣言) 性能要优于 Fake (谣言). 原因在于, 相比于非谣言文本, 谣言制造者在发布谣言时, 在文本的措辞上面花费了精力, 导致模型在谣言检测性能上低于非谣言. 相比于其他基准模型, 我们模型由于融入了更丰富的语义知识和实体知识, 在谣言类别检测上面也体现了一定的优势.

(4) 在数据集层面, Pheme 数据集主要采集了 5 个突发事件对应的新闻, 而 Twitter 数据集收集了 52 个谣言相关的事件, 从而 Pheme 数据集上的性能要优于 Twitter 数据集.

(5) 在词向量初始化层面, 总体而言, 采用 word2vec 作为初始化词嵌入取得的谣言检测性能相对比较稳定. 其潜在的原因在于 word2vec 采用了 Wiki 知识图谱进行训练, 而 BERT 采用了普通文本进行训练.

表 4 显示了平均池化下图卷积神经网络的层数 (1 层, 2 层和 3 层) 对英文谣言检测性能的影响, 我们可以明确: 采用 2 层的图卷积神经网络性能要优于 1 层, 原因在于 2 层网络融合了更多的图全局信息, 但是随着层数增加, 并没有带来性能的提升, 潜在原因可能是产生了过平衡现象. 此外, 我们对最优的 2 层图卷积神经网络情况, 进一步比较了最大池化和平均池化两种方式. 其中, 平均池化下谣言检测性能比最大池化要高 2%. 其原因是平均池化可以包含更多的全局信息, 而最大池化仅融入了更多的局部信息.

### 4.4.2 中文谣言检测实验结果及分析

表 5 和 6 分别显示了中文谣言检测实验结果和详细的消融实验结果. 我们同样可以得出如下结论.

(1) 在模型层面, 本文模型在 Accuracy 和 F1 值方面表现优异. 正如我们的预期, 义原为特定的汉语词语产生了详细的层次化语义解释, 这表明义原的层次结构是一个词语的有效表征, 一个词语可能有不同的意义, 而每个意义都被定义为一个义原的层次化结构. 实验结果表明义原确实有助于谣言检测. 同样, 在所有基准模型中, BOW_SVM 在所有 4 个评测指标中表现最差. 由于知识的融合, KCNN 优于 BOW_SVM. 采用同构图形式的 KMGCN-NoKDVisual 模型性能也优于异构图形式的 TextGCN. TextGCN 在 Weibo 数据集中真新闻类别上取得了不错的 Recall 值, 潜在原因在于 TextGCN 在语料库规模相对较大的 Weibo 中计算单词和文档节点的 TF-IDF 值相对精确. 但是, TextGCN 偏向于将社交媒体文本识别成真新闻类别. 由于 TextGCN 所取得的 Precision 值较低, 导致其整体谣言检测

**表 5　中文谣言检测性能**

**Table 5**　Fake news detection results on the Chinese datasets[a)]

| | Accuracy | Fake | | | Real | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 |
| WeChat | | | | | | | |
| BOW_SVM | 0.6197 | 0.6638 | 0.5822 | 0.6204 | 0.6104 | 0.6733 | 0.6403 |
| KCNN | 0.7398 | 0.8684 | 0.5934 | 0.7050 | 0.6791 | 0.8853 | 0.7687 |
| TextGCN | 0.7893 | 0.8692 | 0.6437 | 0.7397 | 0.7136 | 0.8942 | 0.7903 |
| BERT | 0.7962 | 0.8931 | 0.6334 | 0.7412 | 0.6804 | 0.9028 | 0.7760 |
| KMGCN-NoKDVisual | 0.8006 | 0.8697 | 0.6785 | 0.7621 | 0.7573 | 0.8953 | 0.8205 |
| CompareNet | 0.8274 | 0.8974 | 0.7204 | 0.7992 | 0.7764 | 0.9123 | 0.8389 |
| KDRD_word2vec (ours) | **0.8389*** | 0.9126 | **0.7355** | 0.8145** | **0.7943** | **0.9211** | **0.8530*** |
| KDRD_BERT (ours) | 0.8376* | **0.9168** | 0.7334 | **0.8149**** | 0.7796 | 0.9208 | 0.8443* |
| Weibo | | | | | | | |
| BOW_SVM | 0.6351 | 0.7561 | 0.5736 | 0.6523 | 0.6361 | 0.7980 | 0.7079 |
| KCNN | 0.7328 | 0.7354 | 0.7529 | 0.7440 | 0.7220 | 0.7235 | 0.7227 |
| TextGCN | 0.7966 | 0.9651 | 0.5684 | 0.7154 | 0.7225 | **0.9851** | 0.8336 |
| KMGCN-NoKDVisual | 0.8297 | 0.9029 | 0.6987 | 0.7879 | 0.7670 | 0.8934 | 0.8254 |
| BERT | 0.8135 | **0.9770** | 0.6753 | 0.7893 | 0.7432 | 0.9514 | 0.8345 |
| CompareNet | 0.8692 | 0.9376 | **0.8477** | **0.8904** | 0.8374 | 0.8921 | 0.8639 |
| KDRD_word2vec (ours) | **0.8901**** | 0.9255 | 0.8100 | 0.8639 | **0.8621** | 0.9696 | **0.9127**** |
| KDRD_BERT (ours) | 0.8798 | 0.9364 | 0.8461 | 0.8890 | 0.8314 | 0.8921 | 0.8607 |

a) KDRD_word2vec indicates the vector initialization of word using word2vec; KDRD_BERT stands for the vector initialization of word using BERT; * refers to the p-value corresponding to the comparison significance test of the reference system <0.05; ** denotes the p-value corresponding to the comparison significance test of the reference system <0.01.

性能一般. 由于 CompareNet 集成了图卷积神经网络和世界知识, 其性能优于 BOW_SVM, KCNN 和 TextGCN.

(2) 在不同语义知识类型层面, 义原和同义词均起到至关重要的作用. 单个情况下, 义原性能优于同义词, 原因在于义原包含了更详细的层次化语义信息. 此外, 义原和同义词两种信息可以与实体知识有效集成, 发挥两者的互补优势.

(3) 在标签类别层面, 总体而言, Real (不是谣言) 性能也要总体优于 Fake (谣言) 类别. 相比于其他基准模型, 由于我们模型融入了更丰富的语义知识和实体知识, 在谣言类别检测上面也体现了一定的优势.

(4) 在数据集层面, WeChat 采集微信公众号中发布的新闻信息, 而 Weibo 是微博社区中被识别的谣言和官方正确信息. 总体而言, Weibo 语料库包含官方正确信息, 而 WeChat 中的微信公众号信息质量参差不齐. 从而 Weibo 数据集上的性能要优于 WeChat 数据集.

(5) 在词向量初始化层面, 采用 word2vec 作为初始化词嵌入取得的谣言检测性能优于 BERT 作为初始化向量, 原因与英文数据集上类似.

表 7 显示了平均池化下图卷积神经网络的层数 (1 层、2 层和 3 层) 对中文谣言检测性能的影响. 与在英文数据集上取得的性能类似, 采用 2 层的图卷积神经网络性能最优; 同时, 平均池化性能优于

表 6 中文数据集上消融实验结果

**Table 6** Results of the ablation experiment on the Chinese datasets

| | Accuracy | Fake | | | Real | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 |
| WeChat | | | | | | | |
| entity knowledge | 0.8137 | 0.8833 | 0.7099 | 0.7872 | 0.7727 | 0.9051 | 0.8337 |
| semantic knowledge | 0.8285 | 0.9013 | 0.7216 | 0.8015 | 0.7728 | 0.9134 | 0.8372 |
| entity knowledge +sememe | 0.8359 | 0.9034 | 0.7321 | 0.8088 | 0.7906 | 0.9208 | 0.8508 |
| entity knowledge +synset | 0.8315 | 0.9052 | 0.7198 | 0.8019 | 0.7827 | 0.9178 | 0.8449 |
| entity knowledge +semantic knowledge | **0.8389** | **0.9126** | **0.7355** | **0.8145** | **0.7943** | **0.9211** | **0.8530** |
| Weibo | | | | | | | |
| entity knowledge | 0.8461 | 0.9247 | 0.7624 | 0.8357 | 0.7941 | 0.9386 | 0.8603 |
| semantic knowledge | 0.8607 | 0.9167 | 0.7856 | 0.8461 | 0.8334 | 0.9437 | 0.8851 |
| entity knowledge +sememe | 0.8862 | 0.9254 | 0.8045 | 0.8607 | 0.8472 | 0.9681 | 0.9036 |
| entity knowledge +synset | 0.8876 | 0.9138 | **0.8112** | 0.8594 | 0.8568 | 0.9632 | 0.9069 |
| entity knowledge +semantic knowledge | **0.8901** | **0.9255** | 0.8100 | **0.8639** | **0.8621** | **0.9696** | **0.9127** |

表 7 向量平均池化下中文数据集上图卷积层数实验结果

**Table 7** Experimental results with different convolution layers under the average pooling mode on Chinese datasets

| | Accuracy | Fake | | | Real | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 |
| WeChat | | | | | | | |
| KDRD (1 layer) | 0.7665 | 0.8281 | 0.6471 | 0.7265 | 0.7224 | 0.8692 | 0.7890 |
| KDRD (2 layers) | **0.8389** | **0.9126** | **0.7355** | **0.8145** | **0.7943** | **0.9211** | **0.8530** |
| KDRD (3 layers) | 0.8245 | 0.8993 | 0.7117 | 0.7946 | 0.7621 | 0.9204 | 0.8338 |
| Weibo | | | | | | | |
| KDRD (1 layer) | 0.8085 | 0.8934 | 0.6921 | 0.7800 | 0.7321 | 0.9567 | 0.8295 |
| KDRD (2 layers) | **0.8901** | **0.9255** | **0.8100** | **0.8639** | **0.8621** | **0.9696** | **0.9127** |
| KDRD (3 layers) | 0.8834 | 0.9167 | 0.8092 | 0.8596 | 0.8512 | 0.9621 | 0.9033 |

最大池化.

总体而言, 在英文 Pheme 和中文 Weibo 上谣言检测性能较好, 而在英文 Twitter 和中文 WeChat 上谣言检测性能略低. 影响性能的潜在原因在于两个方面. (1) 谣言检测性能受语料库质量的影响. 其中, 英文 Pheme 进行了过滤, 仅包含 5 种突发事件对应的新闻; 中文 Weibo 语料库包含了大量官方正确新闻. 相比较而言, 中文 WeChat 和英文 Twitter 均来自于互联网中的用户贴文, 质量相对略低.

表 8　不同类型节点所占的比例

**Table 8**　Proportion of different types of nodes

| | Text original word node (%) | World knowledge node (%) | Language knowledge node (%) |
|---|---|---|---|
| Pheme | 68.40 | 17.00 | 14.60 |
| Twitter | 60.20 | 19.30 | 20.50 |
| WeChat | 59.50 | 21.60 | 18.90 |
| Weibo | 65.70 | 19.20 | 15.10 |

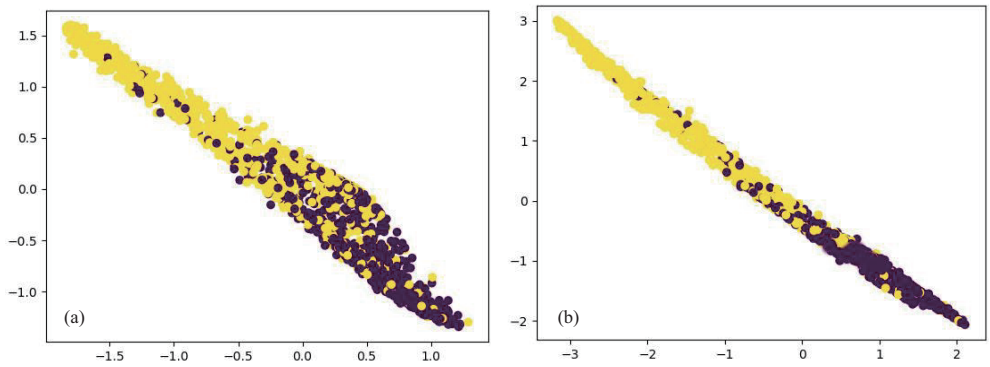


**图 6　英文 Twitter 数据集下可视化效果**

**Figure 6**　Visualization under the English Twitter dataset. (a) 1-layer GCN; (b) 2-layer GCN

(2) 谣言检测性能受语料库包含的谣言事件种类影响. 其中, Pheme 只包含 5 种突发事件对应的新闻, 中文 Weibo 包含娱乐、军事、政治、生活等多种事件. 相比较而言, 英文 Twitter 却包含 52 种事件, 而中文 WeChat 也包含多种事件 (娱乐、军事、政治、生活等) 的微信公众号. 随着事件种类的增加, 谣言检测的挑战性也加大.

### 4.4.3　模型可解释性分析

我们从两个方面较为深入地分析了本文所构建的图神经网络帮助检测谣言的原因.

其一: 我们分别统计了 4 个数据集中文本原单词节点、世界知识节点、语言知识节点 3 种类型节点所占的比例. 从表 8 的统计数据可以看出, 世界知识节点和语言知识节点大约占据了 30%∼40% 的比例, 通过采用 HowNet (义原和同义词) 以及 WordNet (上义词、下义词和同义词) 分别对中英文社交媒体文本的主题词进行扩充, 以及三元组形式的世界知识融入, 有效地增强了两种节点的共现性, 从而丰富了原社交媒体文本的表示, 从一定程度上缓解了数据稀疏共现问题, 提升了谣言检测性能.

其二: 我们利用 t-SNE 工具[14] 对本文模型进行了可视化. 为了较为深入地分析 GCN 的层数影响, 我们对第一层 GCN 下的输出和第二层 GCN 下的输出分别做了可视化. 图 6 和 7 分别显示了英文 Twitter 和中文 Weibo 数据集下的可视化效果. 其中, 黄色节点代表 Fake 类型, 褐色节点代表 Real 类型. 可以明确针对本文模型的输出向量进行聚类后 (图中两种不同颜色节点), 可以对真假谣言进行有效区分, 并且 2 层 GCN 的效果优于 1 层 GCN.

14) https://scikit-learn.org/stable/.

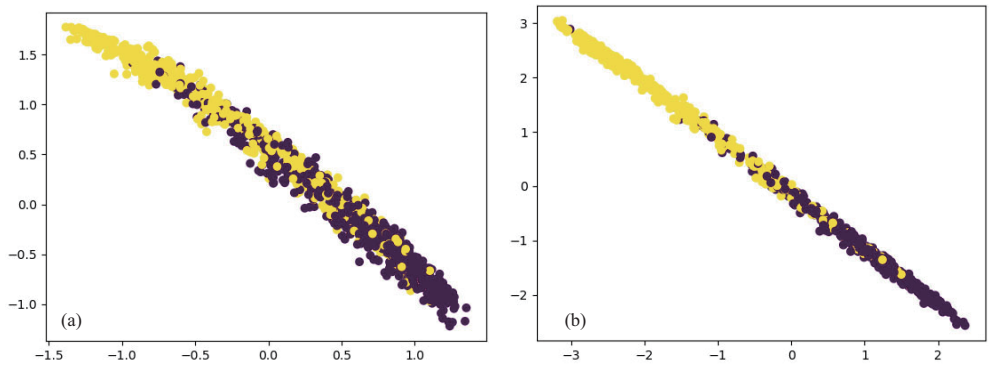


**图 7 中文 Weibo 数据集下可视化效果**

**Figure 7** Visualization under Chinese Weibo dataset. (a) 1-layer GCN; (b) 2-layer GCN

## 5 结论

本文构建了一个功能强大的语义 – 实体无向图, 该图由 3 种类型的节点和 6 种类型的边组成, 任何两种类型的节点之间的共现性可以在语义 – 实体图中被有效地增强. 在图卷积神经网络框架下, 我们成功地将语言知识 (例如, 上义词、下义词、同义词、义原) 和实体知识与外部知识库有机地结合起来, 用于谣言检测. 4 个国际基准谣言检测语料库上的实验结果表明了层次化语言知识和世界知识在谣言检测中起着互补作用, 详细的实验结果和可视化分析验证了本文提出的模型显著优于代表性的基准谣言检测模型.

**致谢** 感谢匿名审稿人提出的宝贵意见. 基于他们的意见, 本文的质量得以极大的提升.

## 参考文献

1 Newman N, Fletcher R, Kalogeropoulos A, et al. Reuters Institute Digital News Report. 2019. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_1.pdf
2 Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science, 2018, 359: 1146–1151
3 Zannettou S, Sirivianos M, Blackburn J, et al. The web of false information. J Data Inf Qual, 2019, 11: 1–37
4 Xu F, Sheng V S, Wang M W. A unified perspective for disinformation detection and truth discovery in social sensing: a survey. ACM Comput Surv, 2021, 55: 1–33
5 Rubin V, Conroy N, Chen Y M. Towards news verification: deception detection methods for news discourse. In: Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, Hawaii, 2015. 1–11
6 Zhao Z, Resnick P, Mei Q Z. Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of the International Conference Companion on World Wide Web, Chicago, 2015. 1395–1405
7 Chen X Y, Zhu D D, Lin D Z, et al. Rumor knowledge embedding based data augmentation for imbalanced rumor detection. Inform Sciences, 2021, 580: 352–370
8 Potthast M, Kiesel J, Reinartz K, et al. A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18), Melbourne, 2018. 231–240
9 Yang S, Shu K, Wang S H, et al. Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'19), Hawaii, 2019. 5644–5651
10 Hu L M, Yang T C, Zhang L H, et al. Compare to the knowledge: graph neural fake news detection with external knowledge. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP'21), 2021. 754–763

11 Miller G A. WordNet: a lexical database for English. Commun ACM, 1995, 38: 39–41
12 Dong Z D, Dong Q. HowNet—a hybrid language and knowledge resource. In: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, 2003
13 Wu K, Yang S, Zhu K Q. False rumors detection on Sina Weibo by propagation structures. In: Proceedings of the IEEE International Conference on Data Engineering (ICDE'15), Seoul, 2015. 651–662
14 John S T, Nello C. Kernel Methods for Pattern Analysis. Cambridge: Cambridge University Press, 2004
15 Ma J, Gao W, Wong K F. Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'17), Vancouver, 2017. 708–717
16 Kumar S, Carley K. Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19), Florence, 2019. 1173–1179
17 Li J W, Sujana Y, Kao H Y. Exploiting microblog conversation structures to detect rumors. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING'20), 2020. 5420–5429
18 Ma J, Gao W. Debunking rumors on Twitter with tree transformer. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING'20), 2020. 5455–5466
19 Yang X Y, Lyu Y F, Tian T, et al. Rumor detection on social media with graph structured adversarial learning. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence(IJCAI-PRICAI'20), 2020. 1417–1423
20 Horne B D, Adali S. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Proceedings of the 2nd International Workshop on News and Public Opinion (NECO'17), Montreal, 2017. 1–9
21 Li J W, Ott M, Cardie C, et al. Towards a general rule for identifying deceptive opinion spam. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'14), Baltimore, 2014. 1566–1576
22 Wang W Y. "liar, liar pants on fire": a new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017. 422–426
23 Castillo C, Mendoza M, Poblete B. Information credibility on Twitter. In: Proceedings of the International Conference Companion on World Wide Web (WWW'11), Hyderabad, 2011. 675–684
24 Yuan C Y, Ma Q W, Zhou W, et al. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING'20), 2019. 5444–5454
25 Mukherjee S, Weikum G. Leveraging joint interactions for credibility analysis in news communities. In: Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'15), Melbourne, 2019. 353–362
26 Li Q Z, Zhang Q, Si L. Rumor detection by exploiting user credibility information, attention and multi-task learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19), Florence, 2019. 5047–5058
27 Ciampaglia G L, Shiralkar P, Rocha L M, et al. Computational fact checking from knowledge networks. Plos One, 2015, 10: e0128193
28 Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks. Mach Learn, 2010, 81: 53–67
29 Shi B X, Weninger T. Discriminative predicate path mining for fact checking in knowledge graphs. Knowledge-Based Syst, 2016, 104: 123–133
30 Shiralkar P, Flammini A, Menczer F, et al. Finding streams in knowledge graphs to support fact checking. In: Proceedings of the IEEE International Conference on Data Mining (ICDM'17), Orleans, 2017. 859–864
31 Pan J Z, Pavlova S, Li C X, et al. Content based fake news detection using knowledge graphs. In: Proceedings of the International Semantic Web Conference (ISWC'18), 2018. 669–683
32 Xu B-B, Cen K-T, Huang J-J, et al. A survey on graph convolutional neural network. Chin J Comput, 2020, 43: 755–780 [徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述. 计算机学报, 2020, 43: 755–780]
33 Hu J W, Liu Y C, Zhao J M, et al. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. In: Proceedings of the 59th Annual Meeting of the Association for Computational

Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP'21), 2021. 5666–5675
34 Yao L, Mao C S, Luo Y. Graph convolutional networks for text classification. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19), Hawaii, 2019. 7370–7377
35 Wang Y Z, Qian S S, Hu J, et al. Fake news detection via knowledge-driven multimodal graph convolutional networks. In: Proceedings of the International Conference on Multimedia Retrieval (ICMR'20), Dublin, 2019. 540–547
36 Xu F, Sheng V S, Wang M W. Near real-time topic-driven rumor detection in source microblogs. Knowledge-Based Syst, 2020, 207: 106391
37 Zhang X Y, Zhang T, Zhao W T, et al. Dual-attention graph convolutional network. 2019. ArXiv:1911.12486
38 Kingma D P, Ba J L. Adam: a method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR'15), 2015
39 Li Q M, Han Z C, Wu X M. Deeper insights into graph convolutional networks for semi-supervised learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18), New Orleans, 2018
40 Jin Z W, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In Proceedings of the 2017 ACM on Multimedia Conference, New York, 2017. 795–816
41 Wang H W, Zhang F Z, Xie X, et al. DKN: deep knowledge-aware network for news recommendation. In: Proceedings of the 27th International Conference on World Wide Web (WWW'18), Lyon, 2018
42 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2019. ArXiv:1810.04805v2

# Knowledge graph-driven graph neural network-based model for rumor detection

Fan XU[1], Minghao LI[1], Qi HUANG[1], Keyu YAN[1], Mingwen WANG[1] & Guodong ZHOU[2*]

1. *School of Computer Information Engineering, Jiangxi Normal University, Nanchang* 330022*, China*;
2. *School of Computer Science & Technology, Soochow University, Suzhou* 215006*, China*
* Corresponding author. E-mail: gdzhou@suda.edu.cn

**Abstract** Rumors can be propagated quickly across online social media at very low costs, resulting in a significant negative impact on society. The conventional rumor-detection models mainly consider information such as propagation patterns, writing style, user credit, and world knowledge. However, the propagation pattern of rumors is often difficult to capture, the writing style is easy to imitate, and the user information composed of metadata (e.g., occupation, hometown, education, and age) is easy to forge. This paper presents a novel knowledge-driven graph convolutional network rumor-detection model. The model represents social media text as a semantic-entity undirected graph structure comprising edges containing six combinations of three types of nodes. The nodes, in turn, contain words from an original text, entity words extended by the world knowledge base, and semantic words extended by the language knowledge base. The semantic-entity graph can effectively enhance the co-occurrence between any two nodes to enrich the representation of the original social media text, alleviating the problem of sparse data co-occurrence. WordNet (hypernym, hyponym, and synonym) and HowNet (sememe and synonym) are adopted to extend the topic words of social media texts, and language and entity knowledge are successfully integrated through the framework of the graph convolution network. The experimental results based on four international benchmarks, Chinese and English databases, and visualization analyses show the effectiveness of our proposed model.

**Keywords** language knowledge, world knowledge, topic model, graph convolutional neural networks, rumor detection