计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：基于知识图谱的多特征融合谣言检测方法
作者：刘小洋，李慧，张康旗，段迪，文癸凌
DOI：10.19734/j.issn.1001-3695.2023.10.0425
收稿日期：2023-10-09
网络首发日期：2023-12-09
引用格式：刘小洋，李慧，张康旗，段迪，文癸凌．基于知识图谱的多特征融合谣言检测方法[J/OL]．计算机应用研究.
https://doi.org/10.19734/j.issn.1001-3695.2023.10.0425

基于知识图谱的多特征融合谣言检测方法 *

刘小洋†，李　慧，张康旗，段　迪，文癸凌
(重庆理工大学 计算机科学与工程学院，重庆 400054)

摘　要：为了解决谣言检测中由于缺乏外部知识而导致模型难以感知内隐信息，进而限制了模型挖掘深层信息的能力这个问题，提出了基于知识图谱的多特征融合谣言检测方法(KGMRD)。首先，对于每个事件，将帖子和评论共同构建为一个文本序列，并利用分类器从中提取其中的情感特征，利用 ConceptNet 基于文本构造其知识图谱，将知识图谱中的实体表示利用注意力机制与文本的语义特征进行聚合，进而得到增强的语义特征表示；其次，在传播结构方面：对于每个事件，基于帖子的传播转发关系构建传播结构图，使用 DropEdge 对传播结构图进行剪枝，从而得到更有效的传播结构特征；最后，将得到的特征进行融合处理得到一个新的表示。在 Weibo、Twitter15 和 Twitter16 三个真实数据集上，使用 SVM-RBF 等 7 个模型作为基线进行了对比实验。实验结果表明：对比当前效果最好的基线，提出的 KGMRD 方法在 Weibo 数据集的 Acc。指标提升了 1.1%；在 Twitter15 和 Twitter16 数据集的 Acc。指标上提升了 2.2%，实验证明提出的 KGMRD 方法是合理的、有效的。

关键词：知识图谱；注意力机制；情感词典；谣言检测
中图分类号：TP399　　doi: 10.19734/j.issn.1001-3695.2023.10.0425

Knowledge graph based multi_feature fusion rumor detection

Liu Xiaoyang†, Li Hui, Zhang Kangqi, Duan Di, Wen Guiling
(*School of Computer Science & Engineering, Chongqing University of Technology, Chongqing* 400054, *China*)

Abstract: In order to solve the problem that it is difficult for the model to perceive implicit information due to the lack of external knowledge in rumor detection, which limits the ability of the model to mine deep information, this paper proposed Knowledge Graph Based Multi_feature Fusion Rumor Detection (KGMRD) method. First, for each event, construct posts and comments together into a text sequence and use a classifier to extract the emotional features. This paper constructed a knowledge graph based on text using ConceptNet and aggregated the entity representation in the knowledge graph with the semantic features of text using the attention mechanism, so as to obtain the enhanced semantic feature representation. Secondly, in terms of communication structure: for each event, this paper built its communication structure diagram based on the propagation and forwarding relationship of the post, and used DropEdge to prune the communication structure diagram, so as to obtain more effective communication structure characteristics. Finally, fused the obtained features to get a new representation and compared seven models including SVM-RBF on three real datasets of Weibo, Twitter15 and Twitter16. The experimental results show that: compared with the current baseline with the best effect, the proposed KGMRD method has the best Acc. on the Weibo dataset and improves the Acc. by 1.1%. And there is a 2.2% improvement on Twitter15 and Twitter16 dataset in Acc. The experiment proves that the proposed KGMRD method is reasonable and effective.

Key words: knowledge graph; attention mechanism; emotion dictionary; rumor detection

0　引言

虚假信息是故意传播以误导或欺骗为目的的虚假或者不准确的消息，它对无论是社会还是个人都有极大的影响[1]。Soroush 等人[2]在《Science》上发表的最新成果将虚假信息与真实信息的传播结构进行了对比，发现虚假信息的传播范围更远、更快、更深、更广。虚假信息由于其巨大的负面影响而成为一个重要问题，它引起研究人员的广泛关注[3]。谣言多集中爆发于突发事件下，在这种情况下由于人们对事实的认知有限，在恐慌心理的影响下民众更倾向于相信并传播谣言。因此，研究社交网络上虚假信息的传播特征，尽早识别出谣言，对社交网络的发展和治理有重大意义。

传统的谣言检测方法主要是利用深度学习或机器学习方法从发布的帖子本身出发，特征工程集中在文本内容方面的挖掘，得到单纯基于内容特征的谣言检测方法[4]，然而这些方法在谣言检测中不能取得较好的效果。Ma[5]等人从空间

收稿日期：2023-10-09；**修回日期：**2023-11-20　**基金项目：**重庆市教委人文社科重点项目(23SKGH247)；重庆理工大学研究生创新基金资助项目(gzlcx20232069)

作者简介：刘小洋(1980—)，男(通信作者)，安徽安庆人，教授，硕导，博士/博士后，主要研究方向为社交网络分析与复杂网络等(lxy3103@163.com)；李慧(1999—)，女，河南洛阳人，硕士研究生，主要研究方向为社交网路分析、谣言检测等；张康旗(1999—)，男，贵州普定人，硕士研究生，主要研究方向为社交网络分析、信息传播与舆情分析等；段迪(2000—)，男，重庆巴南人，硕士研究生，主要研究方向为社交网络分析与数据挖掘等；文癸凌(2001—)，男，四川南充人，硕士研究生，主要研究方向为社交网络分析、推荐系统等.

结构出发，考虑帖子在传播过程中的信息，提出基于传播结构特征的谣言检测方法，以获得模型更好的表现。现有研究基于传播结构特征和文本内容特征[6]提高了谣言检测模型的效果，但是仍存在局限性。包括：a)帖子自身的局限性，例如文本篇幅较短，现有的方法从中提取到的语义信息有限。b)网络用语存在缩写、别名等现象，例如：“特朗普”、“川普”、“特朗普先生”均表示同一个人，是对“唐纳德·特朗普”的提及，这些知识层面的提及和联系有助于提升判断帖子内容的可信度。然而，这些信息不能直接与文本中的实体相关联，因此需要引入外部知识来增强实体间的联系，将知识信息整合到知识图谱中进而增强谣言检测的效果。

针对上述问题，本文提出了基于知识图谱注意力机制的多特征融合谣言检测方法。该方法充分挖掘文本中的语义信息，结合外部知识提取实体背景知识并利用注意力机制将其进行聚合得到外部知识增强的语义信息；利用情感词典和情感分类器抽取文本中的情感特征；构建传播结构图并提取传播结构特征；最终进行融合并分类。

本文的主要贡献有：

a)结合英文的 Twitter 15 和 Twitter 16 数据集和中文的 Weibo 数据集，结合社交网络中发帖与评论的信息；以帖子和其评论转发为节点构建了谣言传播结构图，使用 GCN 提取了传播图的结构特征。

b)利用外部知识，构建知识图谱，将其嵌入表达作为背景知识通过注意力机制与文本语义特征进行聚合以获得语义特征的高阶表达，接着与情感特征以及结合传播结构特征进行特征融合得到谣言检测的更有效表示，基于此提出了一种基于知识图谱的多特征融合谣言检测方法。

c)将提出的 KGMRD 方法在 Weibo、Twitter 15 和 Twitter 16 三个真实数据集上进行大量实验，并与 SVM-RBF 等 7 种机器学习和深度学习模型进行了对比分析以验证所提 KGMRD 方法的合理性与有效性。

1　相关工作

近年来，社交媒体的兴起加剧了谣言的产生与传播，谣言对社会稳定性的影响使得谣言检测吸引了大量研究者的注意力。早期的谣言检测主要依赖于从文本内容、用户信息、传播结构等方面提取谣言的特征以对带有标签的帖子进行分类。这些特征主要是通过人工提取的，属于劳动密集型。如 Yang 等人[4]提出了基于文本特征的时间序列并融合了各种社会语境信息的谣言检测方法。Ma 等人[5]用传播树模拟了微博帖子的传播方式，使用基于内核传播树 Kernel 通过区分传播树结构之间的相似性以达到区分不同类型谣言的高阶模式。然而这些方法太依赖特征工程，需要大量的人力的投入，是费时费力的工作。

随着数据量的攀升以及数据种类的多样性，人工提取特征的难度也逐渐加大，为了解决这个局限性并学习谣言的高级特征，更多的深度学习的方法被用于挖掘谣言的各种隐藏特征以用于自动谣言检测。谣言的传播结构和时间特征也被考虑以提高谣言检测的准确性。Bi 等人[7]从微博信息传播网络的语义信息出发，构建其异构图，使用节点级注意力结合微博节点的邻居节点以生成具有特定语义的节点嵌入，再使用语义级注意力融合提取到的不同语义，进而得到更高级的语义表示。GCN 能够更好地从图中或者树中捕获全局结构特征，注意力机制能更好的聚合文本内容以从中获得更加关键的隐藏特征。随着对谣言检测这一领域的不断深入研究，也有一些研究者将注意力放在外部知识上，希望借助外部知识来增强文本的语义表达，进而获得更高效的表达。如 Castillo 等人[8]依据情感词典提取了 Twitter 谣言文本和非谣言文本中的情感词，进而达到谣言检测的目的。还有学者引入知识图谱以补充帖子内容，以产生更好的表示用于谣言检测。Sun 等人[9]使用双动态 GCN 对传播中的消息动态和背景知识进行融合建模。

然而这些方法忽略了实体之间的知识级相关性，无法根据知识图谱中特定的背景语义来捕捉实体间的高阶语义信息。基于此提出了基于知识图谱的多特征融合谣言检测方法，充分挖掘文本中的语义信息，结合外部知识提取实体背景知识并利用注意力机制将其进行聚合得到外部知识增强的语义信息；使用情感词典和情感分类器抽取文本中的情感特征；构建传播结构图并提取传播结构特征；最终进行融合并分类。图 1 是针对谣言案例结合本文提出的引用外部知识对谣言进行分析的结果。

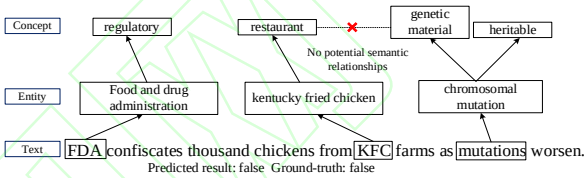


图 1　KGMRD 模型在谣言样本上的应用

Fig. 1　Application of KGMRD model to rumor samples

2　提出的 KGMRD 模型

本文提出的基于知识图谱的多特征融合谣言检测模型的总体框架如图 2 所示。

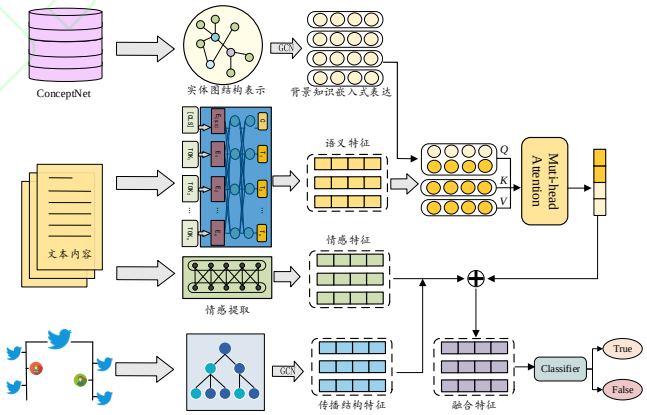


图 2　基于知识图谱的多特征融合谣言检测模型

Fig. 2　A multi_feature fusion rumor detection model (KGMRD) that based on knowledge graph

图 2 中，首先对于帖子中的文本，利用知识蒸馏获得 ConceptNet 中关于文本中实体的背景知识并将其利用 GCN 提取其嵌入式表达。接着，利用预训练模型 Bert 获得文本的语义特征，为了获得增强的语义表达，使用了多头注意力机制将实体与语义信息进行聚合；利用外部知识情感词典获得文本的情感特征表示，将其与增强的语义信息表达进行融合；传播结构的特征提取依赖于基于节点之间存在评论-转发关系，将源帖或者源微博(推文)、转发微博、评论的用户，作为传播结构图中的节点，用以构造传播图。使用 GCN 提取传播结构图中的结构信息，由于原始的传播结构中包含大量无用或者冗余的节点或边，进而会干扰提取出的结构特征。因此在这里使用 DropEdge 方法随机去除冗余的边和节点，以减少干扰，提取更有效的结构特征进而提高谣言检测的准确度。

对于每一个事件 E_i 有相应的标签 Y_i 与之对应，来表示事件的性质，$y_i \in \{\text{TR,FR}\}$ (TR 代表是谣言，FR 代表不是谣言)，在一些数据集中 y_i 还有其他取值，(TR: true rumor，FR: false rumor，UF: unverified rumor，NR: non-rumor)。谣言检测的目的就是通过学习谣言数据中的特征，并构造分类器，根据学习到的特征使用谣言分类器来区分真实性未知的微博或推文。

$$f: E_i \to Y_i \tag{1}$$

其中：E_i 是将要确定的事件，Y_i 是分类器给出的事件真实性的标签。

2.1　知识蒸馏

知识图谱是结构化的数据模型，具有描述真实世界实体的数百万个条目，例如，人，物、地点。知识图谱中的实体表示为图节点，实体之间的关系表示为边。知识图谱已经被广泛用于推荐系统[10]，以及对话生成中。一些方法尝试将知识图谱引入到谣言检测中，利用从外部知识中提取背景知识信息，以补充语义相关性来达到更好的谣言检测效果[11~14]。

识别文本中的实体并利用知识相关性策略来学习知识增强的实体表示。外部知识库中的先验知识能够为谣言检测提供更丰富的信息，因此从外部知识图谱中抽取出和文本相关的知识，即文本中包含的实体的概念解释，这些概念可以用作背景知识来增强文本的语义表示。本文使用 ConceptNet 知识图谱[12]作为外部知识。ConceptNet 是一个用自然语言描述人类一般知识的大规模知识图，它包含 590 万元组，310 万个概念和 38 个关系。将每个元组(头部概念、关系、尾部概念、置信度分数)表示为τ=(x，r，c，s)。为了到一个与给定的帖子文本相关的概念集 CE，使用了知识蒸馏来达到这一目的，其具体实现如下：首先，对于帖子 P_i,通过实体链接方法，把文本中的模糊实体提及链接到知识图谱中的正确实体 e，接着对于每个被识别的实体 $e \in EP$，从 ConceptNet 中通过概念化获得其概念信息，从而得到其概念集 CE。将文本中所有实体的概念集 CE 进行合并得到帖子 P_i 的背景知识 $BG_{P_i} = \{C_{E_1}, C_{E_2}, ..., C_{E_n}\}$，考虑实体-概念 $KG = <V, E>$，其中 V 表示节点集合，E 表示边集合。引入图 KG 的邻接矩阵 $\boldsymbol{A}$ 及其度矩阵 $\boldsymbol{D}$，其中 $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{A}_{ij}$，$\boldsymbol{A}$ 的对角元素设为 1 且有自环，每个节点与 d 维特征向量相关联，使用图 KG 的特征矩阵 $\boldsymbol{X} \in \mathbb{R}^{d \times v}$ 来表示所有顶点的初始特征，其中第 i 行对应于第 i 个节点的特征向量。基于邻接矩阵 $\boldsymbol{A}$ 和度矩阵 $\boldsymbol{D}$，每个 GCN 层输入特征矩阵 $\boldsymbol{X}$，可以得到背景知识的表示：

$$\boldsymbol{Z}_1 = \sigma(\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_0) \tag{2}$$

$$\boldsymbol{Z}_l = \sigma(\hat{\boldsymbol{A}}\boldsymbol{Z}_{l-1}\boldsymbol{W}_l) \tag{3}$$

其中：$\boldsymbol{Z}_l$ 代表第 l 层图卷积层的隐层特征，$\hat{\boldsymbol{A}} = \tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}$ 是归一化后的邻接矩阵，$\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$，$\boldsymbol{I}$ 是一个单位矩阵，表示邻接矩阵与其单位矩阵的自连接，$\tilde{\boldsymbol{D}}_{ii} = \sum_j \tilde{\boldsymbol{A}}_{ij}$ 表示第 i 个节点的度，$\boldsymbol{W}$ 表示可训练参数，$\sigma(\cdot)$ 是激活函数。

2.2　文本语义特征提取

Bert 是一种基于 Transformer 构架的高级预训练词嵌入模型[13]，本文使用 Bert 中作为句子编码器以获得句子的上下文表示，将其作为文本的语义特征。

$$P = Bert - CLS(w_1, ..., w_n) \tag{4}$$

2.3　外部知识增强注意力

在获得文本语义特征以及实体表达后，为了表征外部知识的相对重要性，将文本语义特征序列 P_t 投影到注意力机制中的 Q，K，V 向量中，即 $Q = PW_Q$，$K = PW_K$，$V = PW_V$。其中 $W(\cdot)$ 是可训练参数矩阵，通过这种方法可以得到语义特征和背景知识更加有效的聚合，其表示如下：

$$P' = Att([Z;P], P, P) = \text{softmax}(\frac{[Z;Q]K^T}{\sqrt{d}})\text{V} \tag{5}$$

其中：[;]表示拼接，P'是通过注意力机制融合的具有更有效表达的结果，多头注意力机制被用于获得多头注意力层的输入结果：

$$Multihead(P') = Concat(P'_1, P'_2, ..., P'_n)W \tag{6}$$

其中：P'是每个注意力层的输出，n 是注意力层的层数，W 是可训练参数。

2.4　情感特征提取

对于帖子和评论的向量表示，将其长度控制为 L，对于文本长度大于 L 的，将其裁剪为 L，长度小于 L 的将其用 0 向量进行填充，使其长度为 L。接着对于长度为 L 的文本序列 $C = [c_1, c_2, c_3, ..., c_i, ..., c_L]$，其中 c_i 是文本的第 i 个单词。将这些向量表示输入到提出的模型中，提取其中的情感特征。

为了使获得的情感特征更具有解释性，使用了情感分类器和情感词典从文本内容中提取特征。给定的文本输入序列为 L，其中 c_i 是文本中的第 i 个单词 ，目标是从文本 C 中提取情感特征。

1)情感分类

对于情感分类，使用公开的情感分类器去获得帖子文本的情感分类特征。给定情感分类器 F_{emo}，和帖子文本 C，假设输出的维度是 d_f，因此对文本 C 的预测是 $F_{emo}(C)$。从而能够获得文本的情感分类特征 $emo_T^{category} = F_{emo}(C)$，其中 $emo_T^{category} \in \mathbb{R}^{d_f}$。

2)情感词典

为了更好的获得句子的情感表示以及充分利用情感词典信息，本文将情感词典加入到情感特征提取任务中，为模型提供额外的情感特征信息。将情感词典记为 $D = \{d_1, d_2, ..., d_m\}$，其中情感词典 D 包含 m 种情感，对于情感 $d \in D$，情感字典提供了一个包含 L 个情感单词的单词表 $F = \{f_1, f_2, ..., f_L\}$.

在给定文本 C 的情况下，逐渐将每个单词和完整文本在左右情绪中的得分进行汇总以丰富表示。

对于情绪 d，首先计算单词级别的得分 $score(C_i, d)$,其中 c_i 是文本序列 C 中的第 i 个单词，如果单词 c_i 在词典 F 中，不仅考虑它的出现频率，还考虑其上下文中的程度词和否定词。

接着对文本分词，找出文档中的情感词、否定词以及程度副词，查看每个情感词之前有无否定词及程度副词，将它之前的否定词和程度副词划分为一个组，若有否定词，则将情感词的情感权值乘以否定词的值，若有程度副词就乘以其程度值，然后将所有组的得分加起来，大于 0 的归于正向情感，小于 0 的归于负向，得分的绝对值大小反映了文本的消极或积极的程度，通过这种方式获得每个单词的情感得分，其计算方式如下：

$$score(c_i, d) = \frac{G * neg(c_i, s) * deg(c_i, s)}{L} \tag{7}$$

$$G(c_i) = \begin{cases} 1, & if \quad c_i \in F \\ 0, & otherwise \end{cases} \tag{8}$$

其中：s 是左侧上下文的窗口大小，$neg(c_j)$ 和 $deg(c_j)$ 分别是单词 c_j 的负值和程度值，这些可以通过情感词典查找到。

$$neg(c_i, s) = \prod_{j=i-s}^{i-1} neg(c_j) \tag{9}$$

$$deg(c_i,s)=\prod_{j=i-s}^{i-1}deg(c_j) \tag{10}$$

将所有获得的单词的得分 $score(c_i,d)$ 进行相加，得到基于文本的情感得分 $score(C,d)$，文本情感得分的计算如下：

$$score(C,d)=\sum_{i=1}^{L}score(c_i,d),\forall d\in D \tag{11}$$

将获得的文本级情感得分拼接得到基于情感词典的情感特征：

$$emo_C^{lexcion}=score(C,d_1)\oplus score(C,d_2)\oplus...\oplus score(C,d_m) \tag{12}$$

在得到这两种特征后，将所有获得的不同种类的情感特征进行拼接进而得到文本的情感特征 emo_C，如式所示。

$$emo_C=emo_C^{category}\oplus emo_C^{lexcion} \tag{13}$$

其中：$emo_C\in\mathbb{R}^d$。

2.5 传播结构特征提取

基于帖子及其转发和评论关系为其构造了帖子的传播结构图 $G<V,E>$，其中 V 做为传播结构图的节点集包含了帖子发布者和用户节点，E 是传播结构图的边集表示节点之间有评论或者转发行为。由于近年来，卷积模型在图域中的应用越来越广泛，在现有的卷积模型中 GCN 的表现是最有效的模型之一，因此，在模型中采用 GCN，其计算“消息传递”架构的方式如下：

$$\boldsymbol{H}_k=M(\boldsymbol{A},\boldsymbol{H}_{k-1};\boldsymbol{W}_{k-1}) \tag{14}$$

其中：$\boldsymbol{H}_k$ 和 $\boldsymbol{H}_{k-1}$ 分别是由第 k 层和第 $k-1$ 层的卷积层计算得来的隐藏向量矩阵，M 是信息传播函数，$\boldsymbol{A}$ 是传播结构图的邻接矩阵，$\boldsymbol{W}_{k-1}$ 表示可训练的参数。由 ChebNet et al.[15]的对消息传播函数的定义可知上式可写成如下形式：

$$\boldsymbol{H}_k=M(\boldsymbol{A},\boldsymbol{H}_{k-1};\boldsymbol{W}_{k-1})=\sigma(\hat{\boldsymbol{A}}\boldsymbol{H}_{k-1}\boldsymbol{W}_{k-1}) \tag{15}$$

其中：$\hat{\boldsymbol{A}}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}$ 是归一化后的邻接矩阵，$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$，$\boldsymbol{I}$ 是一个单位矩阵，邻接矩阵与其单位矩阵相加表示添加自连接，$\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}$ 表示第 i 个节点的度，$\sigma(\cdot)$ 是激活函数。

由于传播图 G 的节点多，较为复杂，为了防止在 GCN 训练过程中出现过拟合现象同时为了减少由于过平滑引起的信息缺失，采用了 DropEdge 机制在模型训练时随机删减掉原始图中的边。假设传播结构图的总共边数为 N_e，弃边率为 p,那么 DropEdge 后的邻接矩阵 $\boldsymbol{A}'$ 由以下方式计算得到：

$$\boldsymbol{A}'=\boldsymbol{A}-\boldsymbol{A}_{drop} \tag{16}$$

其中：$\boldsymbol{A}_{drop}$ 是对 G 中的边集 E 进行随机采样后形成的邻接矩阵，邻接矩阵中的边的数目为 $N_e\times p$。

GCN 被用于提取谣言的传播结构特征，传播图 G 的隐藏特征矩阵 H_1 可以由以下公式获得：

$$\boldsymbol{H}_1=\sigma(\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_0) \tag{17}$$

$$\boldsymbol{H}_k=\sigma(\hat{\boldsymbol{A}}\boldsymbol{H}_{k-1}\boldsymbol{W}_{k-1}) \tag{18}$$

其中：$\boldsymbol{H}_k$ 表示是 GCN 中第 k 层特征，$\boldsymbol{W}_k$ 表示参数矩阵 $\boldsymbol{X}$ 是基于传播树构建的特征矩阵，模型中用 ReLU 函数作为激活函数。

2.6 分类预测

在获得了具有情感信息的语义特征和根节点增强的传播结构特征之后，将这些特征进行拼接从而获得融合特征 F：

$$F=\text{concat}(P',emo_c,\boldsymbol{H}) \tag{19}$$

最后的预测值由拼接后的向量输入一个多层感知机 MLP 和 Softmax 层从而获得对帖子准确性的概率 $\hat{y}$ 的预测，如下式：

$$\hat{y}=\text{Softmax}(\text{MLP}(F)) \tag{20}$$

其中：$\hat{y}$ 是判断谣言类别的概率值，映射该值能得到最终的标签。

3 实验设置

3.1 数据集

为了验证模型的有效性并使实验结果具有普遍性，在中文的 Weibo 数据集和两个英文的数据集 Twitter 15 和 Twitter 16 上进行实验。传播结构图中的节点表示用户的源帖子，边表示转发或者评论关系。在 Weibo 数据集中有两种标签，分别是 True Rumor(TR)和 False Rumor(FR), 在 Twitter 数据集中有四种标签，分别为 True Rumor(TR)、False Rumor(FR)、Unverified Rumor(UF)和 None Rumor(NR)。数据集详细信息如表 1 所示。

表 1　数据集的统计

Tab. 1　Statistic of Datasets

Statistic	Weibo	Twitter15	Twitter16
# posts	3,805,656	331,612	204,820
# Users	2,746,818	276,663	173,487
# Events	4664	1490	818
# True rumors	2351	374	205
# False rumors	2313	370	205
# Non-rumors	0	372	205
# Unverfied rumors	0	374	203

3.2 实验设置

在实验环节，实验配置为 Windows10、CPU Xeon Gold 6226R * 2、128GB 内存、NVIDIA Quadro RTX A6000 * 2。在实验中用下面的模型作为谣言检测模型的基准，与本文提出的 KGMRD 方法进行了比较分析。

a)SVM-RBF[16]，一种基于 SVM 并结合了 RBF 内核的检测模型，它是使用了新浪微博的具体特征构建的分类器。

b)RvNN[17]是一种基于树状结构 RNN 的模型，该模型考虑了谣言传播结构的自上而下和自底向上两个方向的特征。

c)VAE-GCN[18]提出了基于 GCN 的图卷积编码解码的谣言检测模型，学习谣言的文本和传播结构特征以进行谣言检测。

d)Bi-GCN[19]是一种基于 GCN 的谣言检测模型，考虑谣言的传播和扩散结构，并通过根节点特征增强来增强节点表示。

e)PPC[20]结合循环神经网络和卷积神经网络的谣言检测模型，该模型考虑了用户特征在传播路径上的全局和局部变化。

f)HAGNN[21]提出基于图神经网络的谣言检测模型，捕获不同粒度文本内容的高级表示，融合传播结构进行谣言检测。

g)GCNFEM [22]使用图卷积网络表示谣言传播树，以源和响应帖子为图，并根据随着时间的推移发现的对谣言的响应来更新节点表示，进而达到谣言检测的目的。

采用 *Accuracy*(*Acc.*)，*Precision*(*Prec.*)，*Recall*(*Rec.*)和 *F*1-score(*F*1)对提出的KGMRD方法进行性能评估。在Weibo数据集上采用 *Acc.*、*Prec.*、*Rec.*和 *F*1；在 Twitter 15 和 Twitter 16 数据集采用 *Acc.*和 *F*1 进行评价。

3.3 结果分析

在 Weibo 数据集上，将提出的 KGMRD 方法与经典的 SVM-RBF 等7种基线模型进行分析，其实验结果如表2所示。

表 2 中，提出的 KGMRD 方法以 94.6%的准确率(*Acc.*)成为对比的 7 种模型中表现最好的模型，与最佳基准相比有 1.1%的提升，其中 *F*1 值达到了 94.5%，与最佳基准相比有 0.5% 的提升，精确率(*Prec.*)更是达到了 95.9%。因此 KGMRD 方法整体来说优于其他模型。

在 Twitter 15 和 Twitter 16 数据集上将提出的 KGMRD

方法与传统的 SVM-RBF 等 7 种基线模型进行了对比分析，实验结果如表 3 和 4 所示。

表 2　在 Weibo 数据集上的实验结果

Tab. 2　Experimental Results of Weibo Dataset

Method	Acc.	Class	Prec.	Rec.	*F*1
SVM-RBF	0.879	F	0.777	0.656	0.708
		T	0.579	0.708	0.615
RvNN	0.908	F	0.912	0.897	0.905
		T	0.904	0.918	0.911
VAE-GCN	0.935	F	0.958	0.921	0.940
		T	0.917	0.934	0.936
Bi-GCN	0.926	F	0.931	0.899	0.917
		T	0.921	0.947	0.935
PPC	0.921	F	0.896	0.962	0.923
		T	0.949	0.889	0.918
HAGNN	0.928	F	0.844	0.931	0.905
		T	0.916	0.898	0.919
GCNFEM	0.911	F	0.903	0.925	0.909
		T	0.919	0.913	0.927
KGMRD	0.946	F	0.959	0.932	0.945
		T	0.923	0.951	0.944

表 3　在 Twitter 15 数据集上的实验结果

Tab. 3　Experimental Results of Twitter 15 Dataset

Method	Acc.	*F*1			
		NR	FR	TR	UR
SVM-RBF	0.318	0.455	0.037	0.218	0.225
RvNN	0.723	0.682	0.758	0.821	0.654
VAE-GCN	0.856	0.749	0.795	0.905	0.809
Bi-GCN	0.849	0.752	0.779	0.823	0.815
PPC	0.842	0.811	0.875	0.818	0.790
HAGNN	0.865	0.813	0.870	0.905	0.896
GCNFEM	0.863	0.769	0.852	0.911	0.773
KGMRD	0.887	0.859	0.869	0.893	0.901

表 4　在 Twitter 16 数据集上的实验结果

Tab. 4　Experimental Results of Twitter 16 Dataset

Method	Acc.	*F*1			
		NR	FR	TR	UR
SVM-RBF	0.553	0.670	0.085	0.117	0.361
RvNN	0.737	0.662	0.743	0.835	0.708
VAE-GCN	0.868	0.795	0.809	0.947	0.885
Bi-GCN	0.833	0.879	0.815	0.853	0.792
PPC	0.863	0.843	0.868	0.820	0.837
HAGNN	0.874	0.815	0.809	0.880	0.865
GCNFEM	0.860	0.753	0.859	0.909	0.772
KGMRD	0.895	0.871	0.887	0.922	0.908

图 3 是提出的 KGMRD 模型在 Twitter16 和 Weibo 数据集上的三条样本案例得到的结果，模型输出预测概率，经过分类器映射得到 false 或 true 的结果，将输出结果与真实标签对比，表明提出的模型对 Text(1)~Text(3)的预测均准确。

表 3、4 中，KGMRD 方法在 Twitter 15 和 Twitter 16 两个数据集上以 88.7%和 89.5%的准确率(*Acc.*)成为表现最好的模型，与基线中表现最好的 HAGNN 模型的准确率(*Acc.*)相比分别有 2.2%和 2.1%的提升。此外，从表 3 和 4 可以看到 KGMRD 能在两个数据集上的 TR 指标分别达到了 89.3%和 92.2%，在 NR、FR 和 UR 上的精确率(*Prec.*)也都能达到 85%以上。

Text(1): FDA confiscates several thousand chickens from KFC farms as mutations worsen.
Predicted result: false
Ground-truth: false

Text(2): Florida couple arrested for selling fake real estate on the moon.
Predicted result: false
Ground-truth: fasle

Text(3):据新华网：澳大利亚一位教授研究了1983-2000年间105次坠机事件2000多名幸存者的受访记录，总结出6条自救方法：①别与家人分开；②学会解安全带；③距离逃生口近；④背朝飞行方向；⑤带上防烟口罩；⑥听乘务员讲解。紧急情况下它们能救命！扩散！
Predicted result: true
Ground-truth: true

图 3　KGMRD 模型在 Twitter16、Weibo 数据集上的预测结果样例

Fig. 3　Example of prediction results on Twitter16 and Weibo of KGMRD

通过表 2~4 可以看出，与 SVM-RBF 等模型相比，本文提出的 KGMRD 及 GRU、PPC 等模型在一系列评价指标上均有较大的提升，且都达到了 88%以上的准确率(*Acc.*)，表明了基于神经网络的深度学习检测方法在原理上大幅优于基于传统机器学习的检测方法，证明了神经网络模型在不依赖于特征工程的同时，有着更好的谣言特征提取能力。在五个深度学习检测模型中， KGMRD、VAE-GCN 和 Bi-GCN 等结合了 GCN 来提取谣言的传播结构特征，在检测精度上优于其他三个模型，表明了以图结构来对传播过程进行建模并以图卷积神经网络来提取谣言在传播过程中的结构特征是有效的。本文提出的 KGMRD 利用外部知识增强文本的语义特征表达的谣言检测模型，在各项指标上优于其他模型，表明了通过外部知识增强文本语义对于提升谣言检测的精度是合理有效的。总体上，提出的 KGMRD 方法在不同程度上均优于其他的传统的机器学习及深度学习的 7 种模型。

3.4　消融实验

为了验证 KGMRD 方法中各个模块的有效性，设计了相应的消融实验。消融实验的模型如下：

a)KGMRD/KGA:去掉模型中的知识图谱和注意力机制模块，即将语义特征、情感特征和传播结构特征相结合进行谣言检测。

b)KGMRD/GCN：去掉模型中的知识图谱和注意力机制模块，即不考虑帖子的传播结构，将使用注意力机制聚合了外部知识而获得的增强的语义特征与情感特征融合进行谣言检测。

c)KGMRD/E：去掉模型中的情感特征提取模块，即增强的语义特征与传播结构特征相结合进行谣言检测。

在 Weibo、Twitter 15、Twitter 16 数据集上对以上三种模型进行验证，以衡量不同模块的性能和合理性，并与本文提出的 KGMRD 模型进行对比，实验结果如图 4 所示。图 4 是四种模型在 Twitter 15、Twitter16 数据集上的结果。图 5 是针对两条不同的谣言样本案例 Text(1)和 Text(2)，以上四个模型分别对这两个样本的预测概率值。从图 4 和 5 中可以看出与其他三种模型相比，本文提出的模型 KGMRD 有更好的表现，进而证实了模型各模块的有效性。

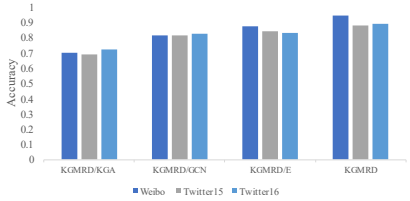


图 4　KGMRD 模型上的消融实验

Fig. 4　Ablation study of KGMRD

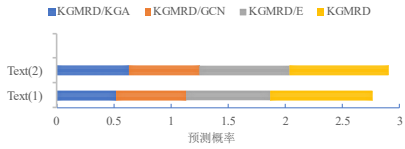


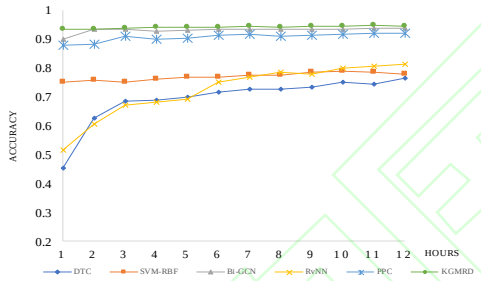
图 5　KGMRD 模型在 Text(1)和 Text(2)上的消融实验的预测结果

Fig. 5　Ablation study of KGMRD on Text(1) and Text(2)

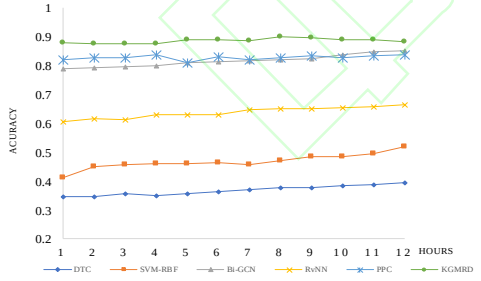
3.5　早期检测

由于随着时间的增加，谣言扩散的范围会越来越广，产生的负面影响也会越来越大，所以尽早地检测出谣言的存在并抑制其传播非常重要，对谣言的早期发现能力也成为一个衡量谣言检测效果的重要指标。为了验证该模型对谣言早期检测的有效性，在三个数据集上的实验过程中设置了一系列的检测截止日期，并对从释放时间到截止日期时间的数据进行了实验。早期检测结果如图 3 所示。

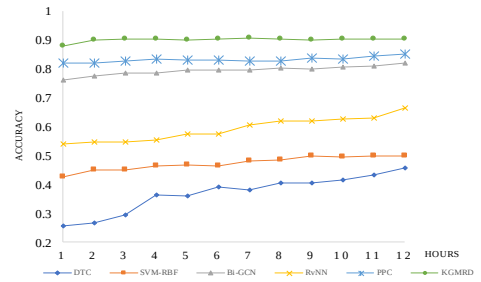
图 6 中，(a)~(c)分别显示了 KGMRD 方法与传统的机器学习方法 DTC、SVM-RBF 等以及深度学习方法 PPC 等模型在 Weibo、Twitter 15 和 Twitter 16 数据集上当设置不同截至时间的情况下性能对比。图 6 显示，本文提出 KGMRD 方法在源帖早期就达到了较高的准确率(*Acc.*)，此外，在每个截止时间本文提出的模型都明显优于其他模型。表明提出的 KGMRD 方法不仅有利于长期的谣言检测，而且有助于谣言的早期检测。



(a)在 Weibo 数据集的早期检测结果



(b) 在 Twitter 15 数据集的结果



(c) 在 Twitter 16 数据集的结果

图 6 早期检测结果

Fig. 6　Early detection results

4　结束语

本文利用外部知识中的信息提出了一种自动谣言检测方法 KGMRD。考虑了帖子中的实体信息与外部知识的链接，结合注意力机制将二者更好的聚合以得到增强的语义特征，提取帖子中的情感特征，考虑帖子传播结构特征；将增强的语义特征与情感特征以及结构特征融合进而得到融合特征并进行谣言检测。为了评估 KGMRD 模型的合理性、有效性，在 Weibo、Twitter 15 和 Twitter 16 数据集上进行实验，对比 SVM-RBF 等 7 种不同的模型；为了验证各个模块的有效性在三个数据集上进行了消融实验，实验结果表明，KGMRD 方法综合来说优于传统的 SVM-RBF 等 7 种基线模型；全面论证了提出的 KGMRD 方法的合理性与有效性。

下一步将考虑从源帖的图片、音频、视频等不同的模态信息中提取谣言特征，实现多模态谣言检测。

参考文献：

[1] Miró-Llinares F, Aguerri J C. Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a‘threat’[J]. European Journal of Criminology, 2023, 20 (1): 356-374.

[2] Vosoughi S, Roy D, Aral S. The spread of true and false news online [J]. Science, 2018, 359 (6380): 1146-1151.

[3] 庞源焜, 张宇山. 句子级状态下 LSTM 对谣言鉴别的研究 [J]. 计算机应用研究, 2022, 39 (4): 2038-2041 (Pang Yuanhun, Zhang Yushan. Rumor identification research based on sentence-state LSTM. [J]. Journal of Application Research of Computers, 2022, 39 (4): 2038-2041.)

[4] Zhang Xueyao, Cao Juan, Li Xirong, *et al.* Mining dual emotion for fake news detection [C]// Proceedings of the web conference 2021. 3465-3476.

[5] Castillo C, Mendoza M, Poblete B. Information credibility on twitter [C]// Proceedings of the 20th international conference on World wide web. 2011: 675-684.

[6] Ma Jing, Gao Wei, Wong K F. Detect rumors in microblog posts using propagation structure via kernel learning [C]. Association for Computational Linguistics, 2017, 9870 (242): 643-654.

[7] Bi Bei, Wang Yaojun, Zhang Haicang, *et al.* Microblog-HAN: A micro-blog rumor detection model based on heterogeneous graph attention network [J]. PloS one, 2022, 17 (4): 12-20.

[8] Castillo C, Mendoza M, Poblete B. Information credibility on twitter [C]// Proceedings of the 20th international conference on World wide web. 2011, 845 (3): 675-684

[9] Sun Mengzhu, Zhang Xi, Zheng Jiaqi, *et al.* Ddgcn: Dual dynamic graph convolutional networks for rumor detection on social media [C]// Proceedings of the AAAI conference on artificial intelligence. 2022, 36 (4): 4611-4619.

[10] Zhang Chengyang, Huang Xianying, An Jiahao. MACR: Multi-information Augmented Conversational Recommender [J]. Expert Systems with Applications, 2023, 213: 118981.

[11] 郭秋实, 李晨曦, 刘金硕. 引入知识表示的图卷积网络谣言检测方法 [J]. 计算机应用研究, 2022, 39 (7): 2032-2036. (Guo qiushi, Li Chenxi, Liu Jinshuo. Rumor detection with knowledge representation and graph convolutional network [J]. Journal of Application Research of Computers, 2022, 39 (7): 2032-2036.)

[12] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5. 5: An Open Multilingual Graph of General Knowledge. In AAAI2017.

4444–4451.

[13] Dun Yaqian, Tu Kefei, Chen Chen, *et al.* Kan: Knowledge-aware attention network for fake news detection [C]// Proceedings of the AAAI conference on artificial intelligence. 2021, 35 (1): 81-89.

[14] Tseng Yuwen, Yang Huikuo, Wang Weiyao, *et al.* KAHAN: knowledge-aware hierarchical attention network for fake news detection on social media [C]// Companion Proceedings of the Web Conference 2022. 2022: 868-875.

[15] Welling M, Kipf T N. Semi-supervised classification with graph convolutional networks [C]// International Conference on Learning Representations. 2016, 475 (35): 499-508. \

[16] Yang Fan, Liu Yang, Yu Xiaohui, *et al.* Automatic detection of rumor on sina weibo [C]// Proceedings of the ACM SIGKDD workshop on mining data semantics. 2012: 1-7.

[17] Ma Jing, Gao Wei, Wong K F. Rumor detection on twitter with tree-structured recursive neural networks [C]. Association for Computational Linguistics, 2018, Beijing, China, 8-13.

[18] Lin Hongbin, Zhang Xi, Fu Xianghua. A Graph Convolutional Encoder and Decoder Model for Rumor Detection [C]// 2020 IEEE the 7th International Conference on Data Science and Advanced Analytics. IEEE, 2020, 67 (3): 300-306.

[19] Bian Tian, Xiao Xi, Xu Tingyang, *et al.* Rumor detection on social media with bi-directional graph convolutional networks [C]// Proceedings of the AAAI conference on artificial intelligence. 2020, 549-556.

[20] Liu Yang, Wu Yifang. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks [C]// Proceedings of the AAAI conference on artificial intelligence. 2018, 32 (1): 1-8.

[21] Xu Shouzhi, Liu Xiaodi, Ma Kai. *et al.* Rumor detection on social media using hierarchically aggregated feature via graph neural networks. Appl Intell 53, 3136–3149 (2023) . https://doi. org/10. 1007/s10489-022-03592-3

[22] Thota N. R. , Sun Xiaoyan, Dai Jun. (2023) . Early Rumor Detection in Social Media Based on Graph Convolutional Networks. 2023 International Conference on Computing, Networking and Communications (ICNC) , 516-522.