软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007322] [CSTR: 32375.14.jos.007322]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于群体情绪稳态化的社交网络谣言检测方法*

殷 茗[1], 乔 胜[1], 陈 威[1], 姜继娇[2]

[1](西北工业大学 软件学院, 陕西 西安 710072)

[2](西北工业大学 管理学院, 陕西 西安 710072)

通信作者: 殷茗, E-mail: yming@nwpu.edu.cn

摘 要: 网络信息来源众多、鱼龙混杂, 及时、准确地判断其是否为谣言是社交媒体认知域研究的关键问题. 先前的研究大多侧重于谣言的文本内容、用户特征或局限于传播模式中的固有特征, 忽略了用户参与事件讨论而产生的群体情绪及其产生且隐藏于谣言传播的情绪稳态特征的关键线索. 提出一种以群体情绪稳态为导向, 融合时序和空间稳态特征的社交网络谣言检测方法, 该方法基于谣言传播中的文本特征和用户行为, 将群体情绪的时序与空间关系稳态化特征相结合, 能够实现较强的表达能力和检测精度. 具体地, 该方法以用户对某事件或话题态度的情绪关键词作为基础, 利用递归神经网络构建时序关系的情绪稳态特征, 使群体情绪具有表达能力较强的时间一致性特征, 可以反映群体情绪随时间的趋同效应; 利用异构图神经网络建立用户与关键词、文本与关键词之间联系, 使群体情绪具有空间关系的细粒度群体情绪稳态特征; 最后, 将两类局部稳态特征进行融合, 具备全局性且提高了特征表达, 进一步分类可获得谣言检测结果. 所提方法运行于两个国际公开且被广泛使用的推特数据集上, 其准确率较基线中性能最好方法分别提高了 3.4% 和 3.2%, T-*F*1 值较基线中性能最好方法分别提高了 3.0% 和 1.8%, N-*F*1 值较基线中性能最好方法分别提高了 2.7% 和 2.3%, U-*F*1 值较基线中性能最好方法分别提高了 2.3% 和 1.0%.

关键词: 谣言检测; 群体情绪稳态; 时序关系; 空间关系; 社交网络

中图法分类号: TP18

中文引用格式: 殷茗, 乔胜, 陈威, 姜继娇. 基于群体情绪稳态化的社交网络谣言检测方法. 软件学报. http://www.jos.org.cn/1000-9825/7322.htm

英文引用格式: Yin M, Qiao S, Chen W, Jiang JJ. Collective Emotional Stabilization Method for Social Network Rumor Detection. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7322.htm

Collective Emotional Stabilization Method for Social Network Rumor Detection

YIN Ming[1], QIAO Sheng[1], CHEN Wei[1], JIANG Ji-Jiao[2]

[1](School of Software, Northwestern Polytechnical University, Xi'an 710072, China)

[2](School of Management, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: There are numerous and miscellaneous sources of online information. Judging whether it is a rumor in a timely and accurate manner is a crucial issue in the research of the cognitive domain of social media. Most of the previous studies have mainly concentrated on the text content of rumors, user characteristics, or the inherent features confined to the propagation mode, ignoring the key clues of the collective emotions generated by users' participation in event discussions and the emotional steady-state characteristics hidden in the spread of rumors. In this study, a social network rumor detection method that is oriented by collective emotional stabilization and integrates temporal and spatial steady-state features is proposed. Based on the text features and user behaviors in rumor propagation, the temporal and spatial relationship steady-state features of collective emotions are combined for the first time, which can achieve strong expressiveness and detection accuracy. Specifically, this method takes the emotional keywords of users' attitude towards a certain event or

* 基金项目: 陕西省自然科学基础研究计划 (2023-JC-YB-615); 教育部人文社会科学基金 (24YJAZH202); 陕西省社会科学基金 (2023R102)
收稿时间: 2024-05-21; 修改时间: 2024-08-13, 2024-10-09; 采用时间: 2024-11-11; jos 在线出版时间: 2025-02-26

topic as the basis and uses recurrent neural networks to construct emotional steady-state features of the temporal relationship, enabling the collective emotions to have temporally consistent features with strong expressiveness, which can reflect the convergence effect of the collective emotions over time. The heterogeneous graph neural network is utilized to establish the connections between users and keywords, as well as between texts and keywords so that the collective emotions possess the fine-grained collective emotional steady-state features of the spatial relationship. Finally, the two types of local steady-state features are fused, possessing globality and improving the feature expression. Further classification can obtain the rumor detection results. The proposed method is run on two internationally publicly available and widely used Twitter datasets. Compared with the best-performing method in the baselines, the accuracy is improved by 3.4% and 3.2% respectively; the T-*F*1 value is improved by 3.0% and 1.8% respectively; the N-*F*1 value is improved by 2.7% and 2.3% respectively; the U-*F*1 value is improved by 2.3% and 1.0% respectively.

Key words: rumor detection; collective emotional stabilization; temporal relationship; spatial relationship; social network

社交网络为谣言的大量产生和迅猛传播提供了天然平台. 早期谣言检测通过建立谣言信息收集平台如新浪谣言信息中心, 进行人工处理[1], 但这种方式存在较大的滞后性已淡出范围. 利用特征工程的传统机器学习进行谣言检测的方法[2−8], 虽然在效率上较人工处理方式有了提高, 但模型一般较为简单, 且需要人为设计数据特征, 费时费力. 利用深度学习的文本谣言检测[9−16]已取得了出色效果被广泛应用. 然而, 社交网络文本是体现信息特征的诸多方面之一, 且是网络用户情绪反应与传播的载体, 用户情绪作为贯穿谣言传播的关键线索被忽视. 现有谣言检测虽然考虑了情绪特征[17,18], 但群体情绪的研究甚少, 缺少通过群体情绪进行谣言检测的专门研究. 因此, 通过群体情绪检测社交网络谣言具有理论价值和意义.

社交网络用户的发布内容不仅包含用户本身特征[19−22], 且有社交网络的情绪表现, 个体情绪在社交信息快速传播中会演变为群体情绪[23−26]. 如图 1 所示, 用户 A 发布了有明显个体情绪的内容, 其相邻节点 B 或 D 看到这条内容并受到影响, 他们反馈出积极、消极或中立的态度, 这些情绪会持续影响于相邻节点, 随着过程推移逐渐达到稳定状态, 群体情绪会趋于一致, 反映在图 1 中即从状态 1 到状态 2 的群体情绪形成过程.

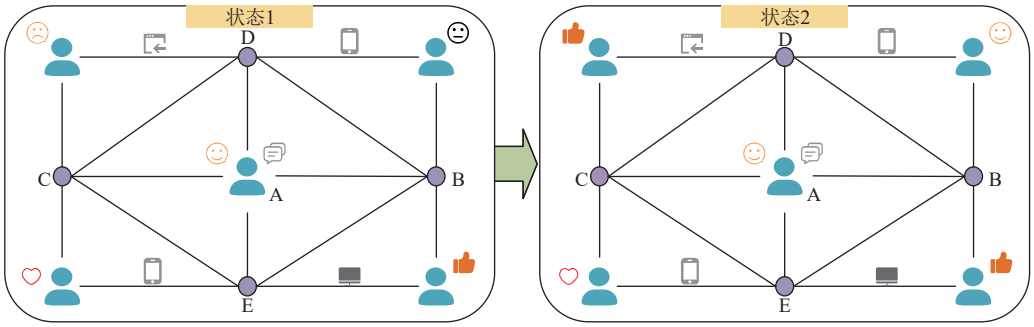


图 1　群体情绪形成过程

群体情绪通常被视为信息传播的加速器[27]. 在谣言传播的过程中, 群体情绪往往会从“中立”或“积极”迅速转变为“消极”, 因为谣言通常包含误导性或惊人的信息, 容易引发恐慌、愤怒或不安等负面情绪. 这些负面情绪可能在一段时间内反复波动, 因为谣言往往伴随不确定性和争议. 最终, 随着谣言被澄清或逐渐失去影响力, 情绪会趋于稳定, 可能回到“中立”或缓和的“消极”状态. 相较而言, 事实的传播通常会引发较少的情绪波动. 群体可能从“中立”迅速转向“积极”或短暂的“消极”, 但这种情绪变化通常较快趋于稳定, 并且情绪类型在短时间内不会频繁变化. 因此, 群体情绪稳态作为社交网络群体的重要特征, 是判断谣言的重要线索, 它对于社交网络谣言的及时发现和阻断具有重要意义.

本文引入群体情绪稳态特征, 对用户群体情绪演化的深入探讨, 综合考虑了情绪随时间变化和空间分布的双重影响, 从而构建出一种全面的谣言检测框架. 本文考虑到群体情绪稳态形成过程在时序关系和空间关系层面的形成特点, 将这两个形成方面作为群体情绪稳态特征的重要基础, 提出了一种融合群体情绪时序稳态特征和空间稳态特征的谣言检测方法. 该方法的时序稳态特征是以时间顺序为基础, 体现了群体情绪随时间变化达到稳态过

程的作用. 空间稳态特征是以用户、推特文本、情绪关键词所构成的异构图为基础, 反映了群体情绪在空间结构关系下的形成过程, 特征融合则是将两种局部特征作进一步融合, 使得新特征具备全局关系, 进而实现准确的谣言分类.

本文第 1 节介绍群体情绪谣言检测相关工作. 第 2 节进行问题定义. 第 3 节介绍基于群体稳态的谣言检测方法. 第 4 节展示不同数据集的对比实验结果, 并对结果进行充分讨论. 最后进行总结和展望.

1 相关工作

1.1 谣言检测方法

目前, 社交网络谣言检测可分为以特征工程[28−30]为基础的机器学习方法和以人工神经网络为基础的深度学习方法. 早期谣言检测大多采用机器学习方法, 相较早期人工处理方式已经有了长足进步. Guacho 等人[5]提出一种基于张量分解的半监督假新闻检测模型, 通过关联传播算法实现对新闻的分类. 但需要指出的是, 该方法所含的数据特征单一, 且因数据不平衡使得假新闻检测结果欠佳. Liang 等人[6]在文本内容基础上, 通过区别对比造谣用户与普通用户的行为, 提出了一种结合用户特征的谣言检测方法, 丰富了数据特征种类并使谣言检测结果更准确. Wu 等人[7]考虑到传播结构, 提出了一种基于图核的混合支持向量机分类模型, 可提取高阶传播模式下的主题、情感等语义特征. 然而, 由于机器学习方法突出特征处理, 使其只能在有限范围内使用. 同时, 由于特征维度低、种类多元性不足, 因而仍然存在过拟合现象, 这使得谣言检测的性能相对较低.

深度学习的异军突起使谣言检测方法全面发生转变, 其检测能力获得大幅提升. 与机器学习方法不同, 深度学习可以处理高维且复杂的数据. Ma 等人[31]首次利用递归神经网络分析推文在转发时随时间产生的语义变化, 并用于谣言检测, 有效增强了其检测可信度. Wan 等人[32]以谣言扩散为切入点, 通过研究谣言与真实信息之间存在的耦合关系, 提出一种谣言扩散干预模型, 借助约束算法实现了谣言扩散随时间的干预. Ma 等人[33]研究了谣言的传播过程, 发现谣言与真实信息之间的特殊关系, 提出了基于传播模式的谣言检测方法. Huang 等人[14]提出一种结合谣言传播过程中时间结构和空间结构的方法, 该方法将所构建的神经网络中的时间、空间结构当作整体特征来进行表示, 通过时间特征提取器、空间特征提取器和积分器这 3 个构件获取到每条信息的时间-空间特征. 然而, 影响谣言传播的因素不止一种, 利用上述深度学习方法虽然可以获得丰富且完整的文本特征, 但是他们侧重于网络结构的构建, 并没有考虑用户情绪线索.

1.2 利用群体情绪稳态的谣言检测

情绪有传播感染作用, 会影响用户的发帖、转发和点赞等行为[34]. Pröllochs 等人[35]通过谣言文本明确了情绪与谣言传播的关系. Horner 等人[36]通过分析谣言标题与用户情绪水平高低之间的关系, 提出了一个全过程模型以减少谣言传播. 然而, 这些研究没有对情绪传播过程及由单一用户情绪所形成的群体情绪进行研究.

目前已有的群体情绪研究主要聚焦在信息传播扩散和影响力方面研究[18,19,37,38]. 我们发现群体情绪是谣言判断的重要线索, 当某条消息在网络传播时, 会自然伴随着用户不断参与讨论的现象, 即消息受众的样本量逐渐增多, 有很强的统计意义. 随着群体情绪的形成, 群体内对消息真假性的态度会趋同, 这对判定谣言有莫大价值. 因此, 本研究聚焦于群体情绪从起初的莫衷一是到趋于稳态的过程, 并将其用于谣言检测, 进而捕捉到谣言传播中群体情绪的独特模式. 不同于现有方法通常依赖单一特征 (如文本特征等), 本文通过将群体情绪从初始波动到最终稳态的动态过程纳入分析, 借助情绪关键词与用户、谣言文本之间的关系, 提出了一种新的、更加高效的谣言检测方法.

2 问题定义

用户在社交网络中通过事件进行交互[39,40]. 本文将情绪划分为“积极”“消极”和“中立”这 3 种类别, 这种分类基于情感分析的常见框架, 在分析社交网络公众对某一事件真实性的态度时尤为关键. 在社交网络中, 用户对事件真实性的反应常常伴随着情绪表达, 这些情绪不仅是对事件本身的回应, 也反映了对事件真实性的认知和信任程度.

“积极”情绪可能表明用户对事件的真实性持肯定态度, 认为事件是真实的, 或对事件背后的信息源表示信任; “消极”情绪则可能表达怀疑、不信任或对事件真实性的否定, 反映了用户对信息的质疑或对事件背景的担忧; 而“中立”情绪则可能表明用户对事件的真实性持保留态度, 既不完全接受也不完全拒绝, 或对事件本身缺乏足够的信息支撑从而以做出判断. 通过对这 3 类情绪的细致分析, 能够更准确地描绘出社交网络用户对事件真实性的整体看法, 理解不同情绪倾向如何影响对信息的接受与传播. 具体如公式 (1) 所示:

$$sentiment = \begin{cases} 1, & \text{positive} \\ 0, & \text{neutral} \\ -1, & \text{negative} \end{cases} \tag{1}$$

基于上述分析, 本文提出了时序关系的群体情绪稳态的概念及分析方法, 定义如下.

定义 1. 时序关系的群体情绪稳态. 在社交网络信息传播过程中, 个体用户除受到事件内容的影响外, 更会受到群体情感倾向的影响[41], 这种影响逐步扩大最终会形成群体情绪, 稳态是群体情绪达到稳定或平衡的状态. 为了定量表示, 本文通过设定阈值来描述群体情绪稳态在时间点 t 内的可信度 R, 则 R 表示如下:

$$R = \frac{1}{N}\sum_{i=1}^{N} E_{t-i} \tag{2}$$

其中, N 表示时间窗口长度, E_{t-i} 表示在时间点 t 内的群体情绪值.

当 R 的值在区间 (0, 1] 时, 表示在考虑的时间段内, 群体情绪倾向于积极;

当 R 的值在区间 [−1, 0) 时, 表示在考虑的时间段内, 群体情绪倾向于消极;

当 R 的值为 0 时, 表示在考虑的时间段内, 群体情绪倾向于中立或情绪分布均衡.

E_{t-i} 的计算过程如公式 (3) 所示:

$$E_{t-i} = \frac{1}{M}\sum_{j=1}^{M} e_{j,t-i} \tag{3}$$

其中, $e_{j,t-i}$ 是个体 j 在时间点 t–i 的情绪值, 其由后文中的 VADER 工具计算得到, M 是个体总数. 通过对社交网络中用户的情绪关键词进行分析, 获取每个时间点上与事件相关的情绪表达. 这些情绪关键词的强度和倾向性 (积极、消极和中立) 通过情绪分析工具 (VADER) 量化为具体数值. 接着, E_{t-i} 是这些情绪强度在时间点 t–i 上的加权平均值, 反映了在该时间点上群体整体的情绪状态.

定义 2. 空间关系群体情绪稳态. 指社交网络中情绪状态在特定空间关系中达到一种平衡或稳定的情况. 在这里, 空间关系指的是社交网络中用户之间的互动模式、信息传播路径以及情绪关键词的分布和关联. 通过分析这些空间关系, 如用户与情绪关键词、推文与情绪关键词之间的关系, 可以揭示出情绪如何在社交网络中传播、变化, 并最终形成稳态. 为了定量表示, 通过设定阈值来描述空间关系的群体情绪稳态在积极、消极和中立这 3 类情绪下的可信度, 具体可表示如下:

$$E_P = \frac{1}{|K_P|}\sum\nolimits_{k\in K_P}\left(\frac{1}{|R_k|}\sum\nolimits_{r\in R_k} \text{情绪强度}(r,k,t)\right) \tag{4}$$

$$E_N = \frac{1}{|K_N|}\sum\nolimits_{k\in K_N}\left(\frac{1}{|R_k|}\sum\nolimits_{r\in R_k} \text{情绪强度}(r,k,t)\right) \tag{5}$$

$$E_n = \frac{1}{|K_n|}\sum\nolimits_{k\in K_n}\left(\frac{1}{|R_k|}\sum\nolimits_{r\in R_k} \text{情绪强度}(r,k,t)\right) \tag{6}$$

其中, E_P、E_N 和 E_n 分别代表积极、消极和中立这 3 类情绪的稳态可信度, K_P、K_N 和 K_n 分别表示积极情绪、消极情绪和中立情绪关键词的集合, R_k 表示与情绪关键词 k 相关联的社交网络关系集合, 情绪强度 (r, k, t) 表示在时间关系上的情绪强度, 其可以用定义 1 中的情绪分析计算得到, 具体如公式 (7) 所示:

$$(r,k,t) = S(k,t) \tag{7}$$

其中, $S(k, t)$ 表示时间点 t 上与情绪关键词 k 相关的情绪评分, 仍由 VADER 计算得到. 从社交网络中的文本或用户行为中提取出与事件或话题相关的情绪关键词 k, 对于每个 k, 使用情绪分析工具 (VADER) 来量化该关键词在

社交网络关系 r 下的情绪表达强度, 结合该情绪在时间点 t 的表达频率和强度, 得到对应的情绪强度值 $S(k, t)$.

结合定义 1 中对积极、消极和中立情绪设定的 [−1, 1] 的阈值, 将积极情绪稳态的阈值区间设定为 0–1, 即当 E_P 大于 0 时, 则判定为积极情绪稳态; 类似地, 当 E_N 小于 0 时, 则判定为消极情绪稳态; 当 E_n 为 0 时, 判定为中立情绪稳态. 按照此种方式可以使后续的时序和空间关系的稳态特征融合过程保持一致性和有效性, 确保两种方法的比较和整合更加准确和有意义.

3 基于群体情绪稳态的谣言检测方法

基于群体情绪稳态的谣言检测方法的框架如图 2 所示, 可分为 4 个阶段: (1) 时序关系的群体情绪稳态特征提取. 通过卷积神经网络 (convolutional neural network, CNN) 对预先提取到的情绪关键词处理, 使其具备时间结构关系, 由于社交网络群体情绪形成的时间跨度和间隔大、依赖关系强, 因此通过门控循环神经网络 (gated recurrent unit, GRU) 捕捉这些时间依赖[42], 得到群体情绪在时序层面的稳态特征. (2) 空间关系的群体情绪稳态特征提取. 以情绪关键词为主, 结合推特文本内容和用户特征构建异构图, 找出群体情绪稳态在图关系层面形成的细粒度关联性, 通过注意力机制得到重要程度接近的具备空间关系的群体情绪稳态特征. (3) 时序与空间稳态特征融合. 对时序关系和空间关系的群体情绪稳态特征进行融合, 以得到完备的、具有全局关系的稳态特征. (4) 谣言分类. 利用 *Softmax* 对融合后特征进行分类并得到谣言检测结果.

群体情绪稳态是一种综合性特征, 是信息传播过程中各方面的结果, 本文研究中体现在从推文中所提取到的情绪关键词, 通过构建时序关系得到的稳态特征, 还有此为基础构建的情绪关键词-用户和情绪关键词-推文异构图, 并进一步得到的空间关系稳态特征. 因此, 稳态特征本质上已不再只是单一的情绪反应.

3.1 阶段 I: 时序关系的群体情绪稳态特征提取

3.1.1 预处理谣言数据

关键词作为情绪传播的载体, 可反映群体情绪一定时间段内的变化. 预先将推特文本分割成单个的单词, 可以减少它们之间的耦合, 也可将其特征化后作为后续输入. NLTK 分词工具简单高效而被广泛使用[43,44]. 本文利用 NLTK 对推特文本进行分割, 并标注单词词性, 这样可以过滤掉常见的单词, 并对分割好的单词进行词干提取和词形还原, 得到所需的情绪关键词, 也可以避免无意义的单词输入到模型从而干扰最终结果.

对于分词后的数据集, 就可以对其进行情绪分析等处理. 本文利用 VADER 工具识别出文本的情绪倾向 (积极、消极、中立), 同时评估情绪的强度. 与一些基于深度学习的复杂情感分析工具相比, VADER 是一个轻量级工具, 不需要大量计算资源或复杂的模型训练. 它能够快速、准确地对短文本进行情感分析, 尤其是在处理社交媒体内容 (如推文、评论) 时表现出色. 它采用了基于词典和规则的混合方法, 能够很好地捕捉文本中的情感极性 (积极、消极、中立) 及情感强度. 这使得 VADER 在处理非正式、短小、口语化的文本时, 具有较高的准确性和鲁棒性, 特别适合在需要快速处理大量文本的应用场景中使用. 在谣言检测中, 情感往往是区分谣言和事实的重要方面. 谣言通常引发更强烈的情感反应, 而事实可能导致较为平稳的情感反应. 因此, 在数据预处理阶段提取这些情感特征, 对于后续的谣言检测至关重要. 由于 VADER 专门针对社交媒体文本进行了优化, 它能够很好地捕捉到谣言传播中所表现出的复杂情感特征. 通过在预处理阶段使用 VADER, 可以为谣言检测模型提供情感标签和情感强度信息, 这些信息能够帮助模型更好地理解谣言传播过程中情感的动态变化, 提高模型的整体检测性能. 具体地, 对于每条推文中的每个词汇以及词组, 利用 VADER 查找对应的情绪倾向并计算情绪强度. 需要特别指出的是, 本文使用 VADER 中默认的情绪强度计算指标, 即对于某一词汇, 若得分大于 0.05 时, 判定为积极情绪; 若得分小于 −0.05 时, 判定为消极情绪; 其他情况则视为中立情绪. 至此, 得到体现用户对相关事件感受态度的情绪关键词.

为了将得到的有上下文关系的情绪关键词作为后续神经网络模型的输入, 需要进行词向量化表示. 本文使用 Word2Vec 模型获取每一个情绪关键词的词嵌入向量.

3.1.2 CNN 构建时序关系

CNN 通过滑动窗口的方式对输入序列 (如单词或词向量) 进行卷积操作. 虽然 CNN 并不直接处理全局的时序

关系, 但它通过在局部窗口内进行卷积, 能够捕捉短范围内的依赖关系. 这些局部依赖关系实际上是时序关系的一部分. 因此, CNN 在文本处理中的卷积操作可以视为一种构建局部时序关系的方式.

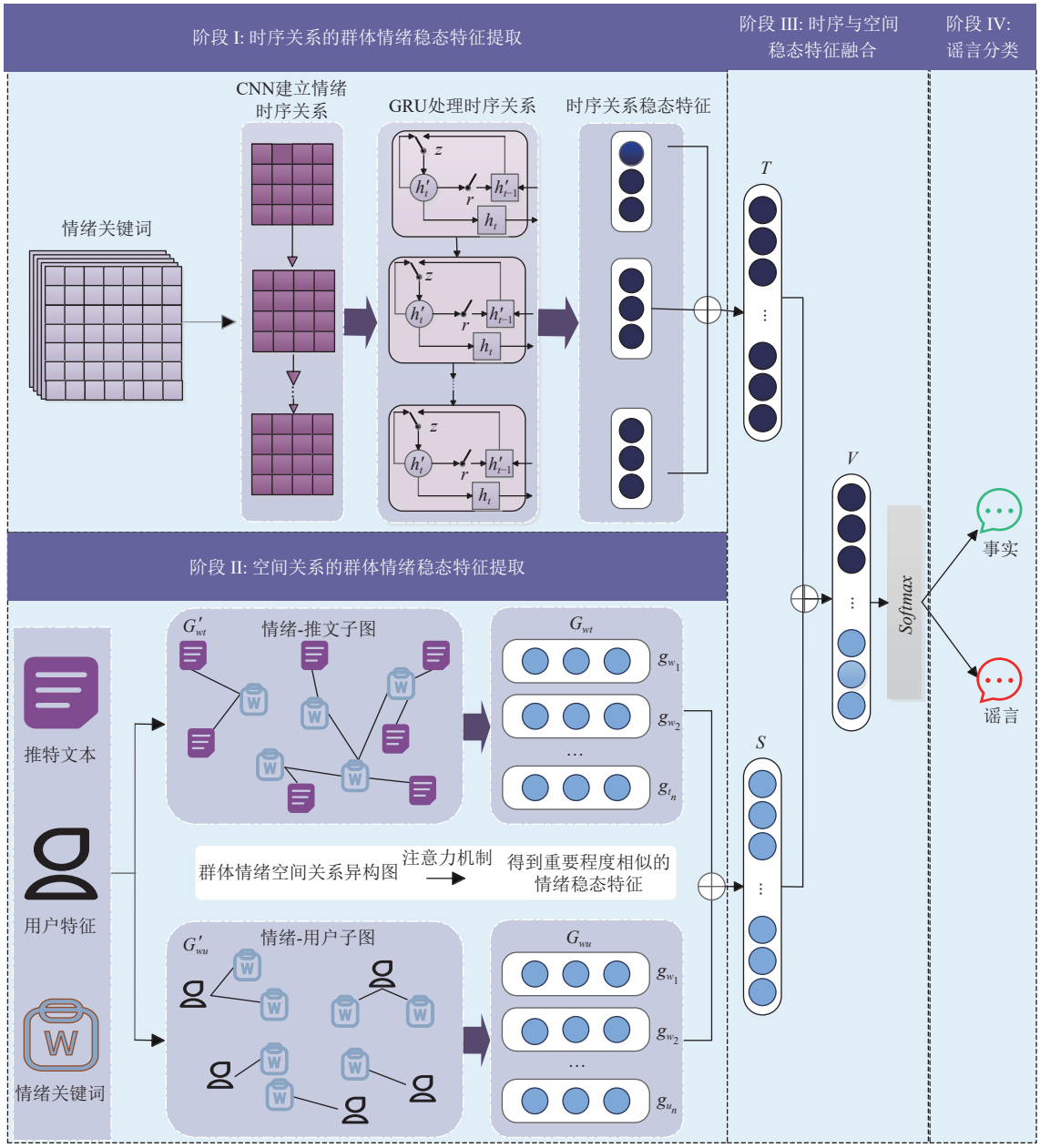


图 2 方法框架

在利用时序关系的群体情绪谣言检测时, 为了通过 CNN 对获取到的反映用户对事件真实性看法的情绪关键词提取局部特征, 特别是情绪特征, 需要对 CNN 改进, 即额外增加特征融合层作为 CNN 的输出层.

通过对推文内容进行词嵌入等一系列处理后, 得到每条推文中 (包括源推文和转发推文) 一系列维度为 d 的情绪关键词向量, 为构建出包含更多上下文信息的数据表示, 将这些词向量组合在一起形成的词拼接矩阵 M^v, 如公式 (8) 所示:

$$M^v = v_1 \oplus v_2 \oplus \ldots \oplus v_{n_v} \tag{8}$$

其中, $\oplus$ 为连接操作运算, $M^v \in R^{d \times n_v}$, n_v 表示词向量总数.

在 CNN 层, 由包括词嵌入层、卷积层以及池化层组成. 使用多个不同大小的滤波器在嵌入向量上进行卷积操作, 以捕捉情绪关键词的不同局部特征, 特别是情绪表达的关键部分利用卷积核 K_i 扫描词拼接矩阵 M^v , 其中 $K_i \in R^{d \times d_k}$, 这表示窗口大小为 d, 宽度为 k. 随后, 使用 *ReLU* 作为激活函数, 这种非线性函数有助于增强模型学习非线性关系的能力, 对于捕捉复杂的情绪特征尤为重要, 具体操作可用公式 (9) 描述.

$$c_i = ReLU((W_c, M^v_{i:i+d_k-1}) + b_i) \tag{9}$$

其中, W_c 表示对应的权重矩阵, b_i 为偏置项且 $b_i \in R$.

在池化层, 对卷积层的输出应用最大池化操作, 以减少特征的维度并提取最重要的情绪特征, 具体操作如公式 (10) 所示:

$$\tilde{c}_i = \max\{c_i, c_{i+1}, \ldots, c_{i+d_k-1}\} \tag{10}$$

在全连接层, 使用 *ReLU* 激活函数重新组合最终特征图 $\tilde{C}$, 以此表示谣言文本中情绪关键词的总体特征 F, 如公式 (11) 所示:

$$F = ReLU(W_f \tilde{C} + b) \tag{11}$$

将公式 (3) 得到的表示时间 t 的稳态值 e_t 通过全连接转换函数 f 扩展成与 F 同维度的向量, 随后将二者进行向量拼接得到 F_t, 具体如公式 (12) 所示:

$$F_t = F \oplus f(e_t) \tag{12}$$

此外, 为了防止过拟合, 在 CNN 层后会加入 *dropout*.

3.1.3 GRU 提取时序关系稳态特征

GRU 负责处理 CNN 提取出的局部特征, 并进一步捕捉全局的时序依赖关系. GRU 能够记住和处理较长时间范围内的信息, 因此这种组合方式可以兼顾局部模式识别和全局时序建模. 如果直接使用 LSTM 或 GRU 处理原始序列数据, 虽然能够捕捉全局时序信息, 但计算开销较大. 而通过 CNN 提取局部特征后再输入到 GRU 中, 可以减少模型的计算负担, 并且更高效地处理输入数据. CNN 提取的特征能够帮助 GRU 更聚焦于关键的时序关系, 提高模型的整体性能和效率.

本文在时序关系群体情绪稳态基础上, 通过改进 GRU 模型中的候选隐藏状态以提取时序稳态特征.

群体情绪是随时间不断变化的, 而 GRU 中的更新门 (z) 则可以决定由上一时间段传入下一时间段的信息, 因此情绪关键词特征随时间段变化的过程可以用更新门表示:

$$z = \sigma(W_z[h_{t-1}, F_t] + U_z \cdot e_t) \tag{13}$$

其中, W_z 表示权重矩阵, e_t 表示第 t 个时间段的稳态值, h_{t-1} 表示前一个 t–1 时间段保存的信息.

本研究需要考虑群体情绪随时间变化时, 以往时间段的情绪对后续时间段的影响, 因此需要利用重置门 (r) 处理以往时间段内的情绪信息, 其公式如下:

$$r = \sigma(W_r[h_{t-1}, F_t] + U_r \cdot e_t) \tag{14}$$

更新门和重置门的不断迭代, 使情绪特征随着时间变化的过程中能够有选择地保留至最后时间段中, 使所得到的序列特征逐渐达到某一特定状态. 将公式 (2) 得到的表示时间 t 的稳态值 e_t 用来调节候选隐藏状态, 可在每个时间段内考虑群体情绪稳态信息, 进而影响 GRU 更新其隐藏状态. tanh 对学习到的信息进行压缩, 起到稳定数值的作用. 据此, 可以得到一定时间段内多个用户对某一推文的情绪逐渐趋于一致的特征:

$$h'_{t-1} = h_{t-1} \otimes r \tag{15}$$

$$h'_t = \tanh(W_h[h_{t-1}, x_t] + W_{e_t} \cdot e_t) \tag{16}$$

$$h_t = (1 - z) \otimes h'_{t-1} + z \otimes h'_t \tag{17}$$

其中, $\otimes$ 是阿达玛乘积, h_t 是隐藏层输出, W_h 和 W_{e_t} 为权重矩阵.

将 GRU 输出的特征向量利用定义 1 中的群体情绪稳态分析方法得到最终时序关系的群体情绪稳态特征向量 T, 使用 *Softmax* 函数进行归一化处理, 如公式 (16) 所示, 代表预测结果为类别 i 的概率, 且 $0 \leqslant \hat{y}(s_i) \leqslant 1$.

$$\hat{y}(T_i) = Softmax(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}} \tag{18}$$

对于时序数据, 在本文所研究的群体情绪稳态方面表现在一定时间范围内群体情绪所达到的稳定模式或趋势, 卷积神经网络 (CNN) 和门控循环神经网络 (GRU) 能够有效地从时间序列数据中捕捉这些局部的稳定特征, 在处理时间序列数据时, 可以通过门控机制捕捉长期依赖关系. 对于情绪数据, GRU 能够记住时间序列中持续存在的情绪状态, 同时过滤掉短期的波动和噪声, 相较于 BERT 等 Transformer 类模型, CNN 和 GRU 更专注于局部的特征提取和长期的时间依赖. 这种专注性使得它们在处理稳态特征时更加直接有效, 不需要通过复杂的自注意力机制去处理全局依赖, 从而避免了模型过度复杂化和冗余计算.

3.2 阶段 II: 空间关系的群体情绪稳态特征提取

3.2.1 构建群体情绪异构图

在社交网络中, 一条推文一般包含发布该推文的用户、情绪关键词及推文内容等多种类型的实体, 而这些实体间存在不同形式的交互关系. 与同构图相比, 异构图能更精确地模拟和分析这些不同类型实体之间的复杂关系, 如用户与推文、推文与关键词之间的关系. 异构图还能够更好地捕捉社交网络数据的丰富语义信息, 提供更深入的洞察, 特别是在分析社交网络中的群体情绪动态时. 通过利用异构图模型, 可以更有效地分析和理解社交网络中用户的情绪表达、推文内容的情绪倾向以及这些情绪如何在网络中传播.

用户通过发帖、评论或转发等行为表达情绪. 这些行为中包含的情绪关键词可用来分析用户的情绪倾向. 例如, 通过分析用户发表或与之互动的推文中出现的情绪关键词, 可以揭示用户情绪状态, 倾向于积极、消极还是中立的情绪. 这种分析可以帮助理解社交网络中不同用户或用户群体的情绪特征和动态变化.

推文内容中的情绪关键词揭示了社交媒体上的群体情绪趋势. 通过分析哪些情绪关键词经常出现在推文中, 可以得知在特定事件或话题讨论中公众情绪的主导倾向. 这种分析可以帮助揭示社交网络上情绪的集体表达模式, 如在重大新闻事件或社会热点讨论中公众情绪的波动.

群体情绪图通常是用来表示不同群体之间的情绪关系或空间分布的, 这种图结构包含节点 (情绪状态) 和边 (关系). 在这种情况下, 数据的结构是图, 而不是像文本那样的序列. CNN 擅长处理具有规则结构的数据 (如文本), 而图数据通常具有不规则的连接结构, 节点的数量和连接方式不固定. 因此, 在这种情况下, 选择专门处理图结构的网络模型更为合理.

因此, 采用异构图模型以情绪关键词为核心进行用户与推文内容的情绪交互分析, 可以理解单个用户的情绪表达, 揭示社交网络中群体情绪的整体趋势和模式, 提供了对社交网络上的情绪动态和公众反应的深入理解. 图 3 为本研究构建的以情绪关键词为主的异构图, 包含推文内容、用户以及 3 类实体间的交互关系.

如图 3 所示, 对于异构图 $G = (V, E)$, V 表示多种类型的节点, 它由情绪关键词集合 W、用户集合 U 及推文集合 T 构成, 即 $V = W \cup U \cup T$. E 表示异构图的边, 它由情绪关键词与用户之间的交互关系 R_{wu}, 情绪关键词与推文内容之间的交互关系 R_{wt} 和情绪关键词之间的语义关系 R_{ww} 构成. 其中, 情绪关键词之间的关系用黑线连接, 情绪关键词-用户之间的关系用黄线连接, 情绪关键词-推文内容之间的关系用红线连接.

对于情绪关键词-用户关系 R_{wu} 的表示需要精确地反映用户通过发帖、评论或转发等行为表达自己的情绪. 为此, 可以利用情感强度加权这种方法, 根据情绪关键词在用户推文中的情感强度来加权这种关系. 不同于简单的正面、负面分类, 情感强度加权能够捕捉情感表达中的微妙变化, 特别适合于分析情感丰富且多样性强的社交文本数据, 这对于理解用户情感的复杂性和动态性非常重要. 此外, 情感强度加权还提供了一种分析社交媒体上情感趋势的额外维度, 有助于深入了解公众情绪的波动和驱动因素. 这个过程涉及对每条推文中的情感关键词强度进行量化, 然后计算其平均值或总和, 具体如公式 (19) 所示:

$$D(k) = \frac{1}{N}\sum_{i=1}^{N} I(k, t_i) \tag{19}$$

其中, $D(k)$ 是特定情绪关键词 k 的平均情感强度, N 是包含情绪关键词 k 的推文数量, t_i 是第 i 条包含情绪关键词 k 的推文, $I(k, t_i)$ 是情绪关键词 k 在推文 t_i 中的情感强度, 通常是一个介于−1 (消极) 到 1 (积极) 的值.

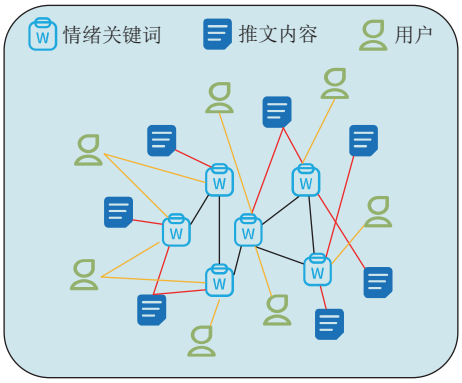


图 3　以情绪关键词为主的异构图

随后使用公式 (3)–公式 (5) 计算用户节点的稳态特征, 具体地, 将用户情绪表达映射后的特征向量记为 u, 如果 u 更频繁地表达积极情绪, 则 E_{P_u} 更高, 为该用户节点赋予更高 E_p 值, E_{N_u} 和 E_{n_u} 值同理. 情绪关键词节点 w 的稳态特征计算和用户节点一致. 最后得到 R_{wu} 的稳态特征 $Feature(wu)$. 如公式 (20)–公式 (22) 所示:

$$Feature(u) = [E_{P_u}, E_{N_u}, E_{n_u}] \tag{20}$$

$$Feature(w) = [E_{P_w}, E_{N_w}, E_{n_w}] \tag{21}$$

$$Feature(wu) = f(Feature(w), Feature(u)) \tag{22}$$

对于情绪关键词-推文关系 R_{wt} 的表示, 通常可使用频率关系方法, 本研究则使用上下文相关关系方法, 是因为社交网络文本通常较短且信息密度高, 上下文相关方法在处理这类短文本时能够提供更丰富和精确的信息解析, 而且能够更深入地捕捉推文中的复杂语义和情感层次, 这对于理解情绪关键词在不同上下文中的具体含义至关重要. 该过程需要使用余弦相似度计算推文内容和特定情绪关键词之间的上下文相关性.

随后根据公式 (3)–公式 (5) 计算推文节点 t 的稳态特征, 根据其中包含的情绪关键词为每个推文节点赋予特征. 由于情绪关键词节点的稳态特征已由公式 (19) 计算得出, 此处不再赘述. 最后得到 R_{wt} 的稳态特征 $Feature(wt)$. 具体如公式 (23) 和公式 (24) 所示:

$$Feature(t) = [E_{P_t}, E_{N_t}, E_{n_t}] \tag{23}$$

$$Feature(wt) = f(Feature(w), Feature(t)) \tag{24}$$

对于情绪关键词之间关系 R_{ww} 的表示, 使用 PMI (pointwise mutual information) 方法.

由于此前已由公式 (19) 计算得出情绪关键词节点的稳态特征, 故可以进一步得到 R_{ww} 的稳态特征 $Feature(ww)$. 具体如公式 (25) 所示:

$$Feature(ww) = f(Feature(w), Feature(w)) \tag{25}$$

为了更好地获取异构图中情绪关键词和推文内容与群体情绪形成的关系, 利用异构图的多元特征分析优势[45], 将异构图分解为情绪-用户子图和情绪-推文子图, 这种划分可以进行更具针对性的分析, 使得能够深入理解情绪如何在社交网络中传播. 情绪-用户子图侧重于分析用户的情绪倾向, 而情绪-推文子图则专注于评估具体推文内容的情绪表达. 通过分别构建两个子图, 可以更细粒度地分析情绪关键词与用户及推文之间的复杂关系, 同时能够提供更全面和多维度的情绪分析.

情绪-推文子图中的节点为异构图中的情绪关键词和推文内容节点, 边亦为异构图中的边, 如图 4 所示. 情绪-用户子图中的节点为异构图中的情绪关键词和用户节点, 边亦为异构图中的边, 如图 5 所示.

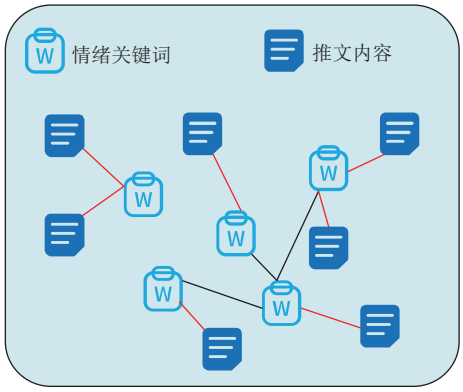


图 4 情绪-推文子图

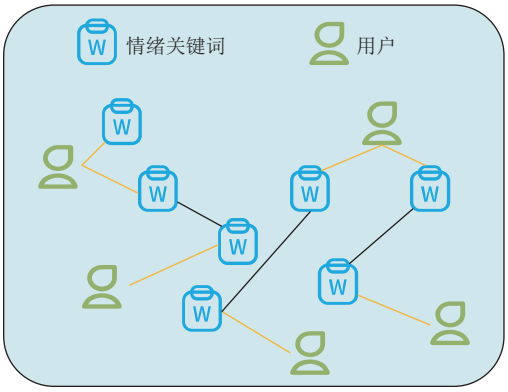


图 5 情绪-用户子图

3.2.2 训练群体情绪异构图

考虑到情绪-用户子图和情绪-推文子图中每个节点的邻居对于学习节点嵌入以进行谣言检测具有不同的重要性, 需要在图注意力网络的基础上, 对子图中的节点应用注意力机制. 子图节点注意力机制允许模型在学习节点表示时考虑到每个节点在特定情绪上的相对重要性, 从而更精确地捕捉与情绪稳态相关的动态.

在原异构图中, 情绪关键词集合 W 表示为 $H_W=\{h_{w_1},h_{w_2},\ldots,h_{w_n}\}$, h_{w_i} 是关键词 w_i 的词向量表示, 且 $h_{w_i}\in R^N$, N 是词向量的维度. 推文集合 T 表示为 $H_T=\{h_{t_1},h_{t_2},\ldots,h_{t_n}\}$, h_{t_i} 是每条推文 t_i 在异构图中的表示. 用户集合 U 表示为 $H_U=\{h_{u_1},h_{u_2},\ldots,h_{u_n}\}$, h_{u_i} 是从每个用户 u_i 中提取的用户行为或用户数据, 且 $h_{u_i}\in R^F$, F 是用户特征维度. 需要注意的是, 当 u_i 无法正常获取时, 利用正态分布对其进行初始化, 以便于后续处理.

对于分解后构建的情绪-推文子图, 无论是推文还是关键词, 它们的特征表示是统一的, 使得它们可以直接在同一特征空间内进行比较和分析, 这样做有助于捕捉推文内容和情绪关键词之间的关系. 而对于情绪-用户子图, 情绪关键词和用户在本质上有不同的特征集, 不适合用相同的方式表示. 例如, 用户可以根据其行为、个人资料等进行特征化, 而情绪关键词的特征则更侧重于其内容、上下文, 故其不同类型的节点具有不同的特征空间. 为解决上述问题, 针对情绪-用户子图中的情绪关键词和用户节点, 分别利用变换矩阵 M_{Φ_w} 和 M_{Φ_u} 将这两种节点映射到相同的向量空间中, 具体如公式 (26) 和公式 (27) 所示:

$$H'_W=M_{\Phi_w}\cdot H_W \tag{26}$$

$$H'_U=M_{\Phi_u}\cdot H_U \tag{27}$$

其中, H'_W 和 H'_U 分别为情绪关键词和用户节点在情绪-用户子图上的映射表示. 通过使用变换矩阵, 就可以利用注意力机制统一处理不同类型节点的不同特征空间子图.

经过上述处理, 情绪-推文子图中的节点可以表示为 $H_{wt}=\{h_{w_1},h_{w_2},\ldots,h_{w_n},h_{t_1},h_{t_2},\ldots,h_{t_n}\}$, 其中, $h_{w_i}\in H_W$, $h_{t_i}\in H_T$. 情绪-用户子图的节点可以表示为 $H_{wu}=\{h_{w_1},h_{w_2},\ldots,h_{w_n},h_{u_1},h_{u_2},\ldots,h_{u_n}\}$, 其中, $h_{w_i}\in H'_W$, $h_{u_i}\in H'_U$. 随后通过使用多头自注意力机制来学习子图中不同节点的权重, 多头自注意力机制在学习子图中节点的权重时具有更高的效率和灵活性, 这对于复杂的社交网络结构和最终利用群体情绪稳态进行精确的谣言检测来说是非常关键的. 对于子图中的节点对 (i,j), 通过自注意力机制 f 来学习系数 $e_{i,j}$, 它表示节点 j 对于节点 i 的重要程度, 具体计算过程如公式 (28) 所示.

$$e_{i,j}=f(Wh_i,Wh_j) \tag{28}$$

其中, W 表示线性变换矩阵.

3.2.3 获取空间关系稳态特征

为了更有效地捕捉子图中节点的局部结构特征, 使得模型能够集中分析每个节点 (用户或推文) 及其直接社交邻居的情绪表达, 这就需要在计算注意力系数时, 模型仅考虑子图中每个节点与其邻居节点之间的关系, 而不是与所有节点的关系, 这种局部聚焦有助于捕捉社交网络中的情绪交互和传播模式. 即对于 $e_{i,j}$, $j \in N_i$, N_i 表示子图中节点 i 的直接邻居节点. 在得到子图中节点的权重之后, 利用 *Softmax* 对其进行归一化处理, 以确保所有进入节点 i 的注意力系数之和为 1, 从而获得归一化的自注意力系数 $\alpha_{i,j}$.

随后聚合子图中节点 i 的邻居表示及其相应的系数, 以更新节点 i 的嵌入表示, 如公式 (29) 所示:

$$Z_i = \sigma\left(\sum_{j\in N_i} \alpha_{i,j} W h_j\right) \tag{29}$$

其中, Z_i 为节点更新后的嵌入表示, σ 为非线性函数, W 仍为权重矩阵, N_i 表示节点 i 的邻居节点集合. 经过上述子图节点注意力操作, 得到了具有关于情绪-推文子图的节点嵌入表示 $H'_{wt} = \left\{h'_{w_1}, h'_{w_2}, \ldots, h'_{w_n}, h'_{t_1}, h'_{t_2}, \ldots, h'_{t_n}\right\}$, 这有助于理解不同推文与特定情绪之间的关系. 情绪-用户子图的节点嵌入表示为 $H'_{wu} = \left\{h'_{w_1}, h'_{w_2}, \ldots, h'_{w_n}, h'_{u_1}, h'_{u_2}, \ldots, h'_{u_n}\right\}$, 可以揭示不同用户对特定情绪的反应.

为确保模型不仅关注单个节点的信息, 还关注这些节点如何在整个网络中相互作用, 以更准确地识别与谣言传播相关的情绪特征, 需要对两个子图进行整体的注意力应用[46]来整合这些子图节点特征. 对于节点嵌入 H'_{wt} 和 H'_{wu} , 情绪-推文子图和情绪-用户子图的权重计算如公式 (30) 所示:

$$(\beta_{wt}, \beta_{wu}) = Att(H'_{wt}, H'_{wu}) \tag{30}$$

其中, *Att* 为进行子图注意力的前馈神经网络.

为学习情绪-推文子图和情绪-用户子图的权重, 需要使用非线性变换, 将原始节点表示转换为更高级的形式, 从而有效地捕捉节点的情绪特征和关系. 计算经过变换的节点表示与经过整体注意力后的子图之间的相似性, 以此结果作为节点重要性的判断依据, 即节点的重要性是通过其表示与一个全局注意力向量的相似度来确定. 随后, 通过计算子图中所有节点重要性的平均值, 得到整个子图的重要性, 该过程实际上是将子图中所有节点信息综合起来, 以评估子图作为一个整体的重要性. 在情绪-推文子图中, 节点表示包括推文内容和与特定情绪相关的特征, 通过上述过程, 可以确定哪些推文中的情绪表达对整个社交网络中的群体情绪稳态最为关键. 在情绪-用户子图中, 节点表示涉及用户的行为、情绪反应, 通过分析节点的重要性, 可以揭示用户在群体情绪的形成和影响社交网络中的群体情绪稳态所起的作用. 此外, 利用上述方法, 模型能够识别和量化这两个子图中节点对于整体情绪稳态的贡献程度, 这对于检测与谣言传播相关的群体情绪非常重要. 具体过程如公式 (31) 和公式 (32) 所示:

$$I_{wt} = \frac{1}{|H'_{wt}|} \sum_{h_i \in H'_{wt}} \mathbf{a}^{\mathrm{T}} \cdot \tanh(W' h_i) \tag{31}$$

$$I_{wu} = \frac{1}{|H'_{wu}|} \sum_{h_i \in H'_{wu}} \mathbf{a}^{\mathrm{T}} \cdot \tanh(W' h_i) \tag{32}$$

其中, W' 为权重矩阵, $\mathbf{a}$ 为两个子图共有的注意力向量.

在分析子图重要性之后, 利用 *Softmax* 函数来对其进行归一化处理, 得到最终的情绪-推文子图的权重 β_{wt} 和情绪-用户权重 β_{wu} . 接着, 通过学习到的子图的权重系数, 融合子图中情绪关键词节点的表示, 以获得原始情绪关键词的表示 Z_W , 具体如公式 (33) 和公式 (34) 所示:

$$Z_W = \{z_1, z_2, \ldots, z_w\} \tag{33}$$

$$z_i = \sum_{\Phi \in \{wt, wu\}} \beta_\Phi \cdot z_{w_i} \tag{34}$$

其中, w 表示所有情绪关键词的数量, z_{w_i} 表示子图中的情绪关键词节点, Z'_Φ 表示子图中有整体关系的节点. 最后, 将 Z_W 作为输入, 利用定义 2 提出的群体情绪稳态分析方法得到空间关系的群体情绪稳态特征向量 S.

3.3 阶段 III: 时序与空间稳态特征融合

社交网络有非结构化的特点, 这就让群体情绪的稳态过程复杂多面, 前面得到的时序关系稳态特征 T 和空间关系稳态特征 S 均仅具备局部意义, 为了得到完整有效的全局关系特征, 需要对 T 和 S 进行特征融合, 得到新的全局稳态特征, 其融合公式:

$$V' = T \oplus S \tag{35}$$

新的稳态特征 V' 结合了原有的稳态特征, 可以在后续获得更佳的谣言分类表现. 为了防止过拟合, 我们使用 *dropout* 函数处理融合过程中出现的正则化项:

$$V = dropout(V') \tag{36}$$

3.4 阶段 IV: 谣言分类

谣言分类器由一个全连接层和一个 *Softmax* 层组成, 将融合后的群体情绪全局稳态特征 V 作为输入, 输出 P 是一个二维向量, 分别表示谣言和真实情况概率:

$$P = Softmax(W_\theta V + b) \tag{37}$$

其中, W_θ 是全连接层的参数, b 是偏置值.

综上, CES 方法的过程可描述如下, 分为 3 部分: 第 1 部分 (步骤 1, 2) 通过对推特数据集的一系列处理得到具有时序关系的群体情绪稳态特征; 第 2 部分 (步骤 3–5) 主要负责提取具有空间关系的群体情绪稳态特征; 第 3 部分 (步骤 6, 7) 负责将前面得到的两种局部稳态特征融合, 得到全局稳态特征, 并对其分类以得到谣言检测结果.

算法 1. CES.

输入: 推特数据集 $E = \{e_1, e_2, \ldots, e_n\}$;
输出: 谣言分类结果 Y.

BEGIN
1. 按第 3.1.1 节的方式对情绪关键词向量化得到初始特征向量;
2. 重复如下操作得到时序关系稳态特征 T: 根据公式 (8)–公式 (12) 利用改进的 CNN 模型构建时序关系, 根据公式 (13)–公式 (17) 利用改进的 GRU 模型处理以往时间段内的情绪信息得到 r, 不断迭代 z 和 r 并根据公式 (2) 得到逐渐趋于阈值的时序关系稳态特征;
3. 按第 3.2.1 节的方式构建群体情绪异构图 G, 根据公式 (17)–公式 (23) 并结合定义 2 分别得到情绪关键词-用户、情绪关键词-推文、情绪关键词-情绪关键词这 3 类关系的稳态特征;
4. 按第 3.2.2 节的方式将 G 分解为情绪-推文子图和情绪-用户子图, 根据公式 (28) 得到关于节点对的注意力学习系数;
5. 按第 3.2.3 节的方式, 根据公式 (29)–公式 (34) 得到空间关系的群体情绪稳态特征 S;
6. 根据公式 (33) 将步骤 2 得到的 T 和步骤 5 得到的 S 融合得到群体情绪全局稳态特征 V, 并利用公式 (34) 处理正则化项;
7. 根据公式 (37) 得到谣言分类结果 Y.
END

此外, CES 方法的训练过程可描述如算法 2.

算法 2. CES 方法的训练算法.

输入: 推特数据集 $E = \{e_1, e_2, \ldots, e_n\}$, 真实标签集合 Y^*;
输出: 训练后得到的参数集合 Θ.

BEGIN

1. 初始化模型参数集合 Θ, 最大迭代次数 EPOCH, 当前迭代次数 epoch;
2. FOR epoch = 1 TO EPOCH DO
3. 计算 E 中的每个事件 e 对应的预测标签 Y;
4. 计算损失值 *loss*;
5. 根据 *loss* 利用优化算法更新参数集合 Θ;
6. END FOR

END

4 实验过程与结果

4.1 数据集

本研究使用 Twitter15 和 Twitter16 国际公开真实谣言数据集, 它们由 Huang 等人[47]在 Ma 等人[13]提出数据集的基础上扩充获得, 已被当作标准数据集而广泛用于国际谣言检测领域. 数据集的描述信息如表 1 所示.

表 1　数据集的描述信息

描述项	Twitter15	Twitter16
源推文总数	1 490	818
用户总数	276 663	173 487
非谣言总数	744	410
谣言总数	372	207
未证实谣言总数	374	201

为了确保实验过程合理且防止过拟合, 本研究利用 holdout 方法[48], 按 80%、10% 和 10% 的比例将数据集随机划分为训练集、验证集和测试集. 其中, 训练集用于训练模型参数, 验证集用于模型性能的初步评估, 测试集用于模型的泛化能力评估.

4.2 评估指标

本研究使用准确率 (Accuracy) 和 $F1$-score 指标评估谣言检测方法的性能, 由于实验数据集中包含谣言 (T)、非谣言 (N) 和未证实的谣言 (U) 这 3 类标签, 因此使用 $F1$-score 进行结果评估时, 将上述 3 类标签对应的 $F1$-score 值分别记为 T-$F1$、N-$F1$ 和 U-$F1$.

4.3 基线方法

为了验证本文提出的 CES 方法的效果, 我们选择了 5 个谣言检测方法作为基准方法同 CES 进行比较.

• GRU-RNN[31]: 该模型将谣言检测的问题定义为一个事件级别的分类任务, 其中每个事件包含多个相关的微博帖子, 将每个事件中的帖子按时间顺序排列, 形成一个时间序列, 时间序列的长度根据帖子的数量确定. 利用递归神经网络能够捕捉序列数据中的时序依赖关系的特性, 有效地提高了谣言的检测准确性, 尤其是在谣言传播的早期阶段. 它能够与本文提出的时序关系的群体情绪稳态特征进行直接对比, 从而验证本文方法在时间维度上的有效性.

• TRNN[33]: 该模型结合了谣言传播过程中的树形结构和 Transformer 的优势, 在利用树形结构建模过程中, 将 Twitter 上的对话结构化为树形结构, 提出了两种树形 Transformer 的变体, 自底向上和自顶向下的树形转换器, 以及这两种方法的结合, 实验结果证明了利用树形 Transformer 甄别谣言的高效性. 在本文中, TRNN 被用来对比空间和传播路径的特征, 这与本文提出的空间关系群体情绪稳态特征形成了良好的对照, 可以有效展示本文方法在复杂网络结构下的优势.

• KAN[11]: 该模型提出一种知识感知注意力网络, 结合了新闻内容和外部知识图谱, 通过新闻-实体 (N-E) 和新闻-实体上下文 (N-EC) 两种注意力机制来评估知识实体及其上下文的重要性. 该模型与本文中空间关系的群体情

绪稳态模块中所使用的异构图注意力网络在技术层面有相似性, 特别在处理复杂的数据关系方面. 将本文方法与 KAN 进行对比, 有助于展示本文方法在结合情绪稳态特征的优势.

• RVNN[13]: 基于自下而上和自上而下树结构的 RNN 模型, 通过 BU 模型和 TD 模型. 按照推文的非顺序传播结构, 融合推文结构语义特征和内容语义特征, 进行谣言检测. 该模型结合谣言语义特征和传播结构特征, 与本文的特征融合方法相似, 而且在谣言的网络传播和时间动态方面与本文的时序和空间关系类似. 与本文提出的时序和空间关系稳态特征模型可以形成互补的对比, 进一步验证本文方法的全局特征提取能力.

• STS-NN[14]: 一种空间-时间结构神经网络模型, 该模型首先将消息传播视为按时间顺序排列的消息序列, 然后对序列中的每个消息应用 STS-NN 单元. 每个 STS-NN 单元包含空间捕获器、时间捕获器和整合器这 3 个组件, 用于捕获每条消息的空间-时间信息. 该模型所使用的技术与本文使用的 CNN-GRU 和异构图注意力网络相结合的方法在处理谣言信息的空间和时间结构方面具有一致性. 本文方法同样融合了时序和空间特征, 通过与 STS-NN 的对比, 可以展示本文所提方法在时空特征融合和情绪稳态捕捉方面的独特优势.

4.4 参数设置

本文在使用 Word2Vec 模型对经过情绪分析后得到的情绪关键词训练词向量时, 词嵌入维度、窗口大小、最小词频、负采样和 epoch 值对模型训练和最终生成的词向量质量有重要影响. 以上参数设置如表 2 所示.

考虑到情绪关键词的数量相对有限, 词嵌入在 150–350 维度范围内足以捕捉词汇的语义信息, 同时避免过度复杂化; 窗口大小决定了 Word2Vec 模型在预测当前词的上下文时考虑的邻近词汇的范围, 较小的窗口可以更好地捕捉情绪关键词间的紧密联系, 此外, 结合一般谣言检测的设置情况, 将窗口大小设置为 5; 最小词频是一个词在文本数据中出现的最低频次, 低于这个频次的词将被模型忽略, 考虑到情绪关键词的重要性和某些情绪关键词可能存在的稀有性, 低频词也可能承载丰富的情绪信息, 选择较低的最小词频阈值有助于包含这些重要但不常见的词汇, 故将最小词频大小设置为 3; 负采样有助于提高训练效率, 对于更多关注情绪关键词的数据集, 负采样为 8 即可很好地完成词向量的训练, 过多的负样本会降低模型效率; 对于 epoch 值, 由于情绪关键词的数量相对较少, 较高 epoch 值可确保模型有足够机会来准确地学习每个词的嵌入表示.

在使用 Word2Vec 模型训练词向量之后, 将这些词向量作为输入, 利用 CNN 和 GRU 进行谣言检测时, CNN 和 GRU 的超参数设置对模型的性能至关重要. 具体数值设置如表 3 所示.

表 2 Word2Vec 超参数设置

参数项	参数值
词嵌入维度	150、200、250、300、350
窗口大小	5
最小词频	3
负采样	8
epoch	20

表 3 CNN 和 GRU 参数设置

参数项	参数值
CNN的滤波器大小	2、3、4
CNN的滤波器数量	128
GRU的隐藏层单元数	128
GRU的层数	2
*dropout*值	0.5
学习率	0.005
批次大小	64

CNN 滤波器 (即卷积核) 的大小决定了卷积操作中考虑的输入数据的区域大小, 与处理图像不同, 在处理文本数据时, 选择 2、3、4 的大小为了能够捕捉双词、三词和四词组合, 这些通常在情绪表达中很常见; 由于每个滤波器都可以捕捉输入数据的不同特征, 因此更多的滤波器意味着能够捕捉更多种类的特征, 在 CNN 的以往应用中, 滤波器数量为 128 在多种任务中表现良好, 是一个平衡模型复杂度和特征提取能力的折衷选择, 同时避免了过于庞大的模型规模所带来的负面影响; GRU 隐藏层单元数决定了模型在处理序列数据时能够记住的信息量, 选择 128 个单元可以平衡捕捉复杂序列依赖的能力和计算效率, 提供足够的复杂度来捕捉序列数据中的情绪动态; GRU 层数设置为 2 层, 通常足以捕捉长短期依赖, 过多层数会导致过拟合和增加计算负担; 由于本实验所用的数据集规模较小, *dropout* 值设置为 0.5 是为了防止过拟合, 同时有助于提高模型的泛化能力; 学习率大小设置为 0.005 时, 平衡了学习速度和稳定性, 避免梯度消失或梯度爆炸; 因本实验以情绪关键词向量为主, 故批次大小不需

要与其他以整文本形式进行实验的情况相同, 将批次大小设置为 64, 虽然会在一定程度使训练速度变慢, 但可能提供更好的泛化能力.

利用异构图注意力网络提取空间关系的群体情绪稳态特征时的参数设置如表 4 所示. 由于多头注意力机制的头数设置应需要根据具体情况灵活调整, 没有一成不变的规则, 一般通过实验来确定最佳头数, 故选择 8、4、2、1 的方式, 提供了从多到少不同层次的视角, 可以帮助找到适合本研究的头数目; 注意力输出维度与词嵌入的最佳维度 200 保持一致, 这样确保了经过注意力机制处理的特征与原始的词嵌入在维度上是直接可比的, 从而简化模型的架构并减少需要额外变换的需求, 减轻不必要的计算负担; 本研究所使用的异构图注意力网络, 涉及权重、偏置以及线性变换的过程中会产生负值, 而使用 LeakyReLU 作为激活函数则可以处理这些负值, 同时可以提供更好的特征表示; 由于后续要进行关于时序和空间关系群体情绪稳态的特征融合, 故学习率和批次大小和第 3 节中的保持一致; 本研究使用 He 作为权重初始化工具, He 初始化是专为 *ReLU* 及其变体 LeakyReLU 设计的, 它考虑了这些激活函数在正输入上的线性特性和在负输入上的小梯度, 从而防止梯度消失或爆炸; 异构图注意力网络在处理图结构数据时经常遇到稀疏梯度的问题, 而 Adam 优化器由于其自适应性质, 特别适合处理这种稀疏梯度, 能够更有效地更新网络权重; 由于融合模型中包含重要的时间序列和网络结构的稳态特征, 故将 *dropout* 值设置为 0.3 这一较低数值以保持关键信息的完整性.

表 4　异构图注意力网络参数设置

参数项	参数值
注意力头数	8、4、2、1
注意力输出维度	200
激活函数	LeakyReLU
学习率	0.005
批次大小	64
权重初始化	He
优化器	Adam
特征融合*dropout*值	0.3

4.5　模型性能结果

在 Twitter15 数据集运行的模型性能结果如图 6 和表 5 所示, 在 Twitter16 数据集运行的模型性能结果如图 7 和表 6 所示. 可以发现, 我们提出的 CES 算法在 Twitter15 和 Twitter16 数据集的准确率分别为 0.791 和 0.773, 较 KAN 方法分别提高 3.4% 和 3.2%, 说明所提出方法优于其他基线, 证明了所提出方法的有效性.

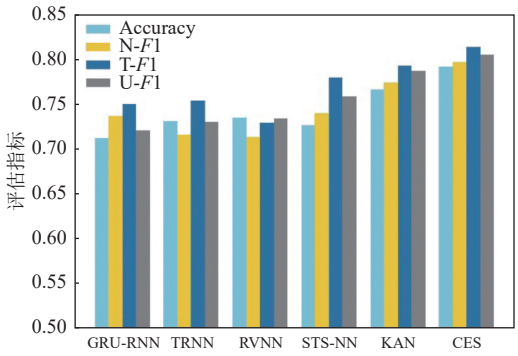


图 6　本方法与基准方法在 Twitter15 数据集的性能对比结果

表 5　Twitter15 数据集实验结果

方法	Accuracy	N-*F*1	T-*F*1	U-*F*1
GRU-RNN	0.711	0.736	0.749	0.720
TRNN	0.730	0.715	0.753	0.729
RVNN	0.734	0.712	0.728	0.733
STS-NN	0.726	0.739	0.779	0.758
KAN	0.765	0.773	0.792	0.786
CES	**0.791**	**0.796**	**0.813**	**0.804**

通过实验结果数据, 本研究发现:

(1) GRU-RNN 方法整体表现偏弱, 在 Twitter15 和 Twitter16 上的准确率均为最低 (分别为 0.711 和 0.693). 其

中, 在 Twitter15 数据集上, 本文所提出的 CES 方法在准确率、N-*F*1 值、T-*F*1 值和 U-*F*1 值方面分别较 GRU-RNN 方法提高 11.25%、8.15%、8.54% 和 11.67%. 性能得到大幅度提升的原因是, GRU-RNN 方法虽然充分考虑到谣言文本的非结构特点, 也获取到丰富的时间序列特征, 但仅使用 RNN 类模型带来的缺陷之一为泛化能力差, 因此在检测结果的准确性方面仍然具备较大的提升空间.

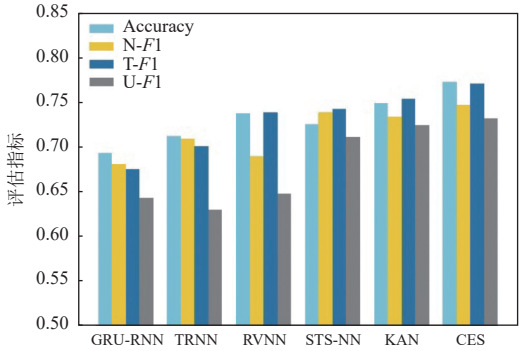


图 7 本方法与基准方法在 Twitter16 数据集的性能对比结果

表 6 Twitter16 数据集实验结果

方法	Accuracy	N-*F*1	T-*F*1	U-*F*1
GRU-RNN	0.693	0.681	0.675	0.643
TRNN	0.712	0.709	0.701	0.630
RVNN	0.738	0.690	0.739	0.648
STS-NN	0.726	0.739	0.743	0.711
KAN	0.749	0.734	0.754	0.725
CES	**0.773**	**0.747**	**0.771**	**0.732**

(2) 虽然 RVNN 在 Twitter15 和 Twitter16 数据集上的准确率 (分别为 0.734 和 0.738) 高于 STS-NN (准确率均为 0.726), 但其他评价指标却均低于 STS-NN. 这是因为 RVNN 模型通过递归神经网络处理传播树结构, 在处理结构上的复杂性方面不如 STS-NN 模型高效, 且 RVNN 可能在捕捉上下文和细节方面不如 STS-NN 模型强, 这会影响对特定类别的精确率和召回率, 从而影响到 T-*F*1、N-*F*1 和 U-*F*1 值. 此外, RVNN 的信息整合和传递方式可能导致某些关键信息的丢失, 尤其是在长距离的信息传递过程中. 在 Twitter15 数据集上, 本文所提的 CES 方法在准确率、N-*F*1 值、T-*F*1 值和 U-*F*1 值方面分别较 RVNN 提高 7.8%、11.8%、11.7% 和 9.7%, 说明 CES 方法较好地解决了 RVNN 中存在的问题, 取得了可观的效果.

(3) KAN 模型由于通过集成外部知识图谱来增强新闻内容的实体识别和对齐, 提供了丰富的上下文和背景信息, 这有助于更准确地识别和验证谣言内容, 故其在 Twitter15 数据集上的全部指标、Twitter16 数据集上除 N-*F*1 之外的指标均高于 STS-NN, 尤其在 Twitter15 数据集上的准确率较 STS-NN (0.726) 增长了 5.4%, 为两数据集各项指标提升最高. 此外, KAN 采用了两种注意力机制, 不仅考虑了新闻的语义表达, 还结合了知识层面的关系, 而 STS-NN 没有采用双重注意力机制, 这限制了它在谣言检测方面的更好表现.

(4) 本文提出的 CES 方法在两个数据集的各项评估指标均优于 KAN 模型, 其中在 Twitter15 数据集上的提升效果最突出, T-*F*1、N-*F*1 和 U-*F*1 较 KAN 分别提升了 3.0%、2.7% 和 2.3%, 在 Twitter16 数据集上分别为 1.8%、2.3% 和 1.0%. 这些结果说明, 将时序与空间关系融合以表达群体情绪的稳态特征, 在谣言检测任务中取得良好的表现, 这也进一步验证了本文提出的群体情绪稳态方法在揭示谣言特性方面的高效性.

实验中注意力网络头的数量对 Twitter15 和 Twitter16 数据集上准确率、N-*F*1 值和 T-*F*1 值的影响分别如图 8–图 10 所示. 可以看出, 随着注意力网络头数量的增加, 准确率、N-*F*1 和 T-*F*1 值相应的变化趋势均为先增高后降低, 且在注意力网络头数量为 4 时均达到最大. 其中, 在两个数据集上准确率平均值为 0.782, N-*F*1 平均值为 0.772, T-*F*1 平均值为 0.792. 出现先增高后降低的趋势是因为不同注意力网络对应不同层次的关联计算, 当注意力网络头数量较小时, 可以处理的信息量不充足, 通常难以得到有效特征. 注意力头数继续增加时, 使模型能够从不同角度同时捕捉信息, 增强了对特征的理解和表达能力, 从而提高准确率、N-*F*1 和 T-*F*1 值. 当注意力网络头数量较大时, 虽然可以处理的数据量也增多, 但会使模型变得过于复杂, 从而学习到训练数据中的噪声, 而非概括性特征. 这会进一步导致模型的泛化能力下降, 过拟合现象随之而来.

另外, 本研究给出了 VADER 在 Twitter15 数据集中推文在积极、中立和消极这 3 种情绪分类的性能结果, 如

表 7 所示. 在分析积极和消极情绪时的表现较分析中立情绪性能更佳, 这是由于积极和消极情感的表达通常较为直接和明确, 包含清晰的情感词汇, 中立情感往往缺乏明确的情感词汇. 同时, 在社交网络的评论数据集中, 情绪分布往往是非对称的, 积极和消极评论的数量相对较多, 而中立评论一般比例较小. 这种不对称的情绪分布使得 VADER 在训练时对极端情绪的识别能力更强, 而对中立情绪的处理能力则相对较弱.

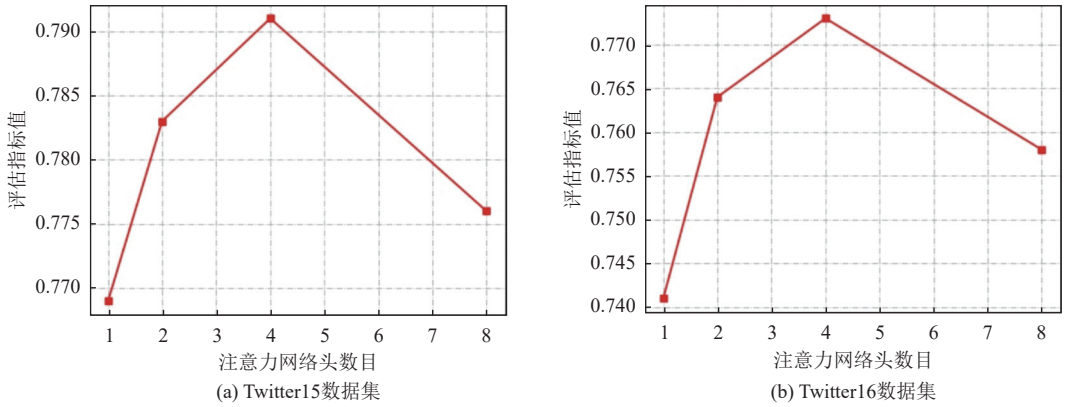


图 8 注意力网络头数量对准确率的影响结果

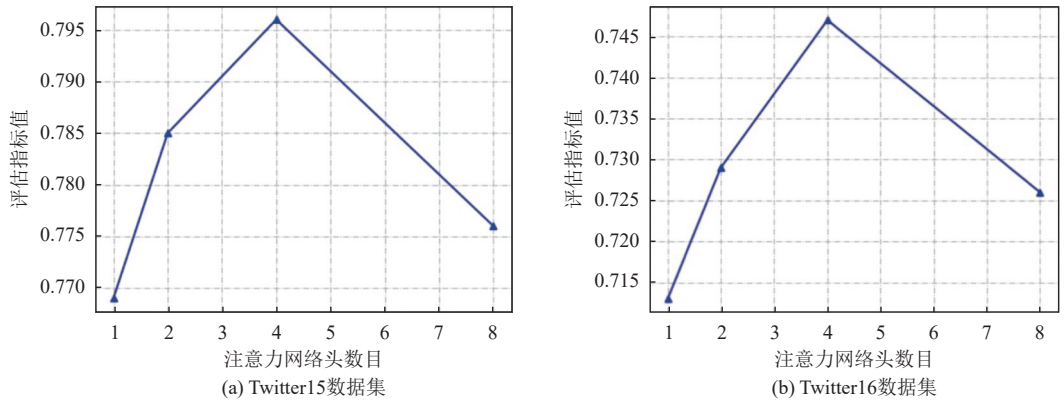


图 9 注意力网络头数量对 N-*F*1 值的影响结果

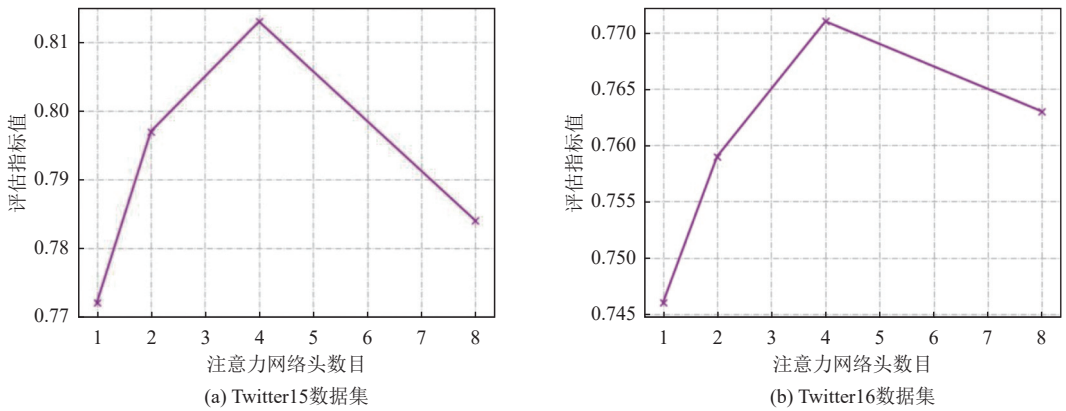


图 10 注意力网络头数量对 T-*F*1 值的影响结果

表 7 VADER 在 Twitter15 数据集推文在积极、中立和消极情绪分类的性能结果

情绪类别	Accuracy	Recall	Precision	*F*1-score
积极	0.84	0.79	0.82	0.80
消极	0.78	0.71	0.76	0.73
中立	0.67	0.63	0.69	0.66

4.6 消融实验结果

为了验证所提出的方法中各部分对整体的性能贡献情况, 本研究设计了 4 组消融实验.

(1) CES 表示在全部模块都保留的条件下进行的实验.

(2) CES\E 表示在只去除情绪关键词特征的条件下进行的实验.

(3) CES\T 表示在时序关系的群体情绪稳态条件下进行的实验.

(4) CES\S 表示在空间关系的群体情绪稳态条件下进行的实验.

消融实验在 Twitter15 数据集的实验结果如图 11 和表 8 所示, 在 Twitter16 数据集的实验结果如图 12 和表 9 所示. 可以发现, CES\E 的谣言检测效果最差, 在 Twitter15 数据集, 较 CES 的准确率、N-*F*1、T-*F*1 和 U-*F*1 分别下降了 21%、20%、16% 和 18%, 下降幅度较 CES\T 和 CES\S 最大. 同时, 在 Twitter16 数据集的情况也类似, 较 CES 的准确率、N-*F*1、T-*F*1 和 U-*F*1 分别下降了 23%、19%、21% 和 21%. 证明情绪关键词特征对实验结果的作用十分明显, 从而说明利用群体情绪稳态进行谣言检测起到了重要作用. CES\S 的各项指标下降最小, 在 Twitter15 数据集上的准确率、N-*F*1、T-*F*1 和 U-*F*1 分别下降了 3%、3%、2% 和 4%.

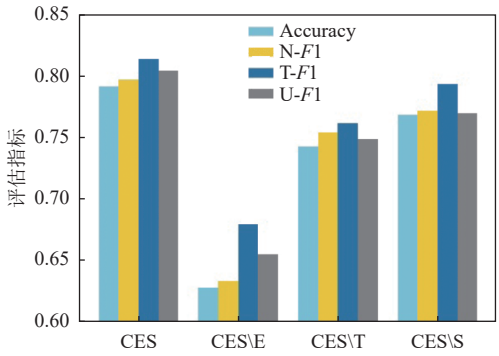


图 11 Twitter15 数据集消融实验可视化结果

表 8 Twitter15 数据集消融实验结果

模型	Accuracy	N-*F*1	T-*F*1	U-*F*1
CES	0.791	0.796	0.813	0.804
CES\E	0.628 **(−21%)**	0.633 **(−20%)**	0.679 **(−16%)**	0.655 **(−18%)**
CES\S	0.742 (−6%)	0.754 (−5%)	0.761 (−6%)	0.748 (−7%)
CES\T	0.768 (−3%)	0.771 (−3%)	0.793 (−2%)	0.769 (−4%)

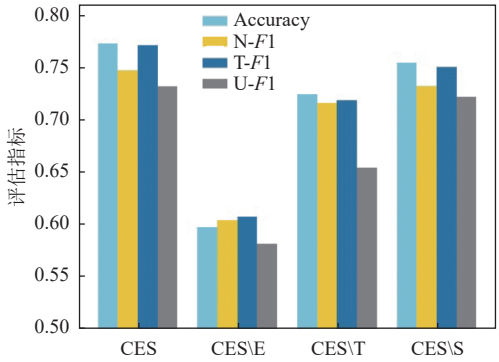


图 12 Twitter16 数据集消融实验可视化结果

表 9 Twitter16 数据集消融实验结果

模型	Accuracy	N-*F*1	T-*F*1	U-*F*1
CES	0.773	0.747	0.771	0.732
CES\E	0.597 **(−23%)**	0.603 **(−19%)**	0.607 **(−21%)**	0.581 **(−21%)**
CES\S	0.724 (−6%)	0.716 (−4%)	0.718 (−7%)	0.654 (−11%)
CES\T	0.754 (−2%)	0.732 (−2%)	0.750 (−3%)	0.722 (−1%)

对于 CES\E, 由于情绪关键词特征对于谣言检测至关重要, 这些关键词能够直接反映群体在面对谣言时的情绪反应强度和方向性, 进而指示信息的可疑性和潜在的谣言性质, 去除这一特征导致模型无法有效捕捉这些关键

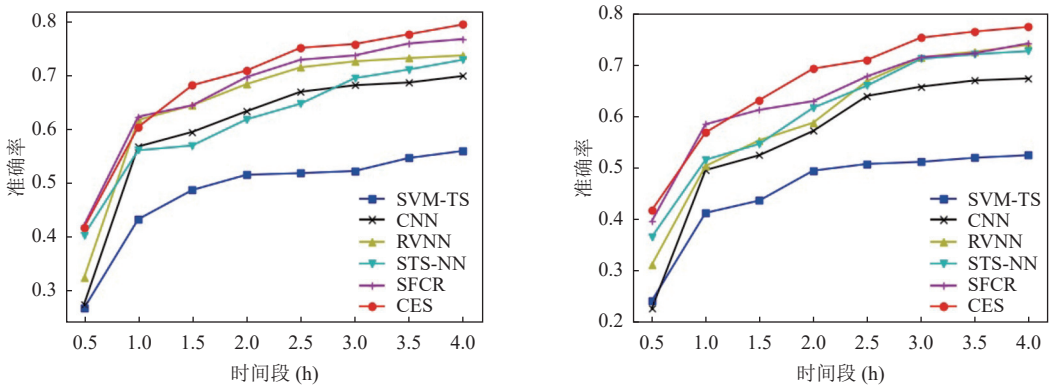
的群体情绪线索, 因而整体性能大幅下降.

对于 CES\T, 其表现次于 CES\S, 但依然较为出色, 表明时序关系的群体情绪稳态特征在谣言检测中的重要性, 时序特征捕捉了情绪随时间的变化和演变过程, 尤其是谣言引发的情绪波动和逐渐稳定的动态模式.

对于 CES\S, 在消融实验中表现出最小的性能下降, 这表明基于情绪关键词-用户和情绪关键词-推文两种空间关系的群体情绪稳态特征对谣言检测具有显著的贡献. 空间关系特征侧重于情绪在用户和推文之间的分布, 谣言通常通过影响特定群体或用户群而扩散, 这些群体可能具有相似的观点, 因此在情绪关键词的使用上表现出一定的共性. 模型通过这些空间关系特征, 从而更准确地检测出谣言.

4.7 案例分析

本研究通过比较在谣言传播开始的 4 h 内 CES 和基准方法的准确率, 在 Twitter15 和 Twitter16 数据集上运行的准确率随时间变化的趋势分别如图 13 所示. 可以看出, CES 在第 1 小时后各个时间段的准确率均高于基准方法, 这是因为处理数据时需要学习时间, 随着谣言传播, CES 性能也逐渐提高, 这对应于群体情绪的形成并逐渐达到一定程度的稳定状态. 本文提出的 CES 方法在谣言传播的开始阶段就可以更准确地甄别谣言, 有着其他方法不具备的优势, 说明群体情绪对谣言检测具有积极的指导意义.



(a) Twitter15数据集 (b) Twitter16数据集

图 13 方法运行准确率随时间变化图

为了直观展现案例中群体情绪稳态的形成过程, 本文利用词云分别展示了在第 1 小时、第 2 小时和第 4 小时的群体情绪状态, 该案例为 Twitter15 数据集中一个群体情绪表现为消极的谣言事件. 效果如图 14–图 16 所示. 其中每个单词代表着一个或多个用户对相应事件内容评论的情绪关键词, 单词字体的大小在表示其出现的频率, 字体越大, 意味着单词出现次数越多. 可以发现, 第 1 小时群体情绪的特点为多样且复杂, 人们对信息内容看法不一, 有怀疑、相信、漠视, 也有持中立态度; 第 2 小时群体情绪的特点表现为对消息内容持怀疑、否定的用户增多, 占比也增大, 说明随着用户参与度的扩大, 对事件内容真实性的判断也逐渐有所依据; 第 4 小时群体情绪的特点则是基本一致, 虽然仍有个别用户相信, 但从根本上无法影响绝大多数用户对案例中信息内容真实性的否定, 也从侧面甄别了谣言, 这种特点即为群体情绪稳态的鲜明反映.

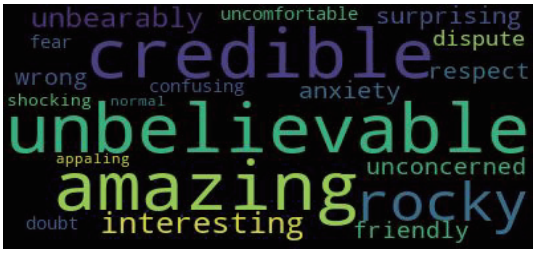


图 14 第 1 小时群体情绪的词云

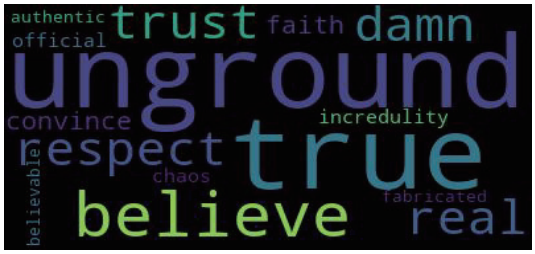


图 15 第 2 小时群体情绪的词云

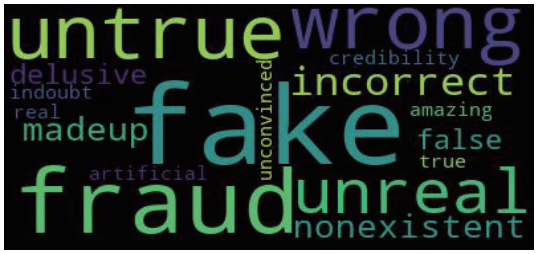
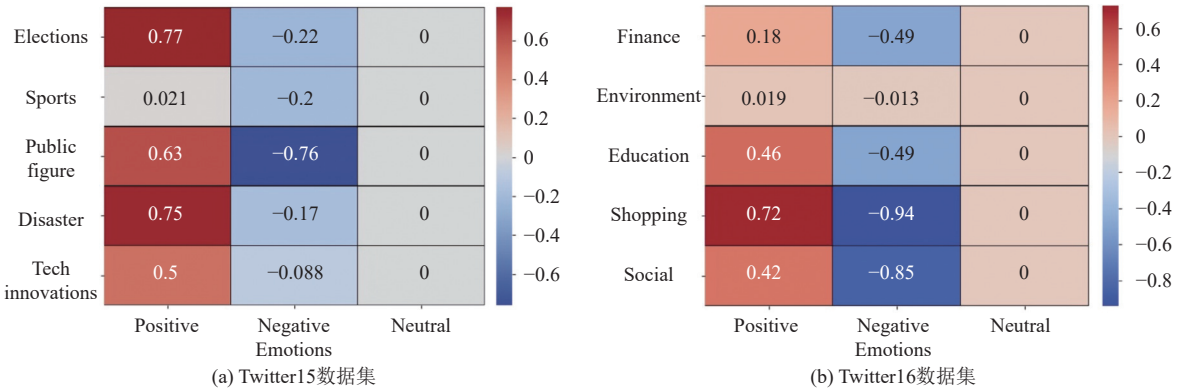


图 16　第 4 小时群体情绪的词云

图 17 表示 Twitter15 数据集上的“选举”“运动”“公众人物”“自然灾难”和“科技创新”这 5 类话题和 Twitter16 数据集上的“金融”“环境问题”“教育”“购物”和“社会问题”这 5 类话题下的群体情绪分布热图, 横轴表示积极、消极和中立这 3 种不同类型的群体情绪, 当分值大于 0 时表示积极情绪, 当分值小于 0 时表示消极情绪, 当分值等于 0 时表示中立情绪.



(a) Twitter15数据集　　(b) Twitter16数据集

图 17　5 类话题下的群体情绪分布热图

在图 17(a) 中, “选举”话题下积极情绪分值较高 (0.77), 接近 0.8, 而该话题下对应案例已被证实为真实事件, 这反映了公众对选举事件的积极反应以及这些反应在社交媒体上的表现形式, 支持者们可能会在社交媒体上积极宣传他们支持的候选人或政党, 分享积极的政治消息和观点. 集体认同感在选举期间往往会增强, 这种认同感可以通过社交媒体上的积极交流和支持表达得到加强. 此外, 对某一政党或候选人胜利的预期可能会在其支持者中产生积极的情绪反应. 这种观察有助于理解社交网络上关于选举的讨论动态, 及公众如何在数字平台上表达和分享他们对政治事件的感受和态度. 通过分析这种情绪分布, 可以深入地了解公众对重大政治事件的反应, 而且可以得知公众的积极情绪反应和观点基于事实, 而非虚假或误导性信息.

如图 17(b) 所示, “购物”话题下消极情绪分值很高 (−0.94), 接近−1, 而数据集中该话题几个对应案例已被证实为谣言, 这在一定程度说明购物相关的谣言, 如虚假促销活动或错误的产品信息, 可能迅速在社交媒体上扩散, 引起群体中广泛的不满或急促的购买决策, 最终形成消极的群体情绪, 而即使这些信息事后被证明是虚假的, 但这种情绪的变化可能破坏消费者的整体信心和购物体验的正面性.

表 10 是利用 VADER 工具计算 Twitter15 数据集中推文的情绪强度结果的示例. 其中, VADER 并不是直接对每个词或词组进行评分, 而是根据该词或词组的积极分值、消极分值和中立分值给出综合加权分值作为某一词或词组的情绪强度结果. 对于积极词或词组赋予正值, 对于消极词或词组赋予负值. 这种方法能够有效获取到社交网络中用户交流所产生的情绪信息.

表 10　利用 VADER 在 Twitter15 数据集的部分推文内容词汇及词组的情绪强度计算结果

单词/词组	积极分值	中立分值	消极分值	综合分值
okay	0.20	0.70	0.10	0.10
amazing	0.85	0.15	0.00	0.85
fake	0.00	0.30	0.70	−0.70
hype	0.60	0.40	0.00	0.60
fake news	0.00	0.70	0.30	−0.40
super excited	0.90	0.10	0.00	0.90
not happy	0.00	0.40	0.60	−0.60
really cool	0.80	0.20	0.00	0.80
totally wrong	0.00	0.30	0.70	−0.70
best day ever	0.90	0.10	0.00	0.90
not that great	0.00	0.60	0.40	−0.40
fake news	0.00	0.70	0.30	−0.40

5　总结与展望

针对以往研究中仅侧重于谣言的文本内容、用户特征或局限于传播模式中的固有特征, 而忽略用户产生的群体情绪的问题, 本文将群体情绪稳态引入谣言检测中, 群体情绪稳定过程对于谣言检测之所以有效, 是因为谣言和事实在最终情绪状态上的动态特征存在显著差异. 通过分析和捕捉这种差异, 能够更早、更准确地识别谣言传播. 这种动态过程的分析, 不仅补充了对最终情绪稳态的理解, 还提供了识别谣言传播中的关键线索. 相较于传统方法, 本文融合了群体情绪的时序与空间稳态特征, 突破了仅依赖文本或用户特征的局限, 使得谣言检测在准确性方面有所提升. 通过融合时序和空间两类关系的群体情绪稳态特征进行谣言检测, 在公开权威的 Twitter15 和 Twitter16 数据集的实验结果表明, 所提出方法的整体表现优于基准实验, 表现出良好的性能, 消融实验证明了所提出方法中时序和空间关系稳态特征对谣言检测效果的提升作用, 案例分析表明群体情绪稳态方法在谣言传播早期就能甄别谣言, 群体情绪对谣言检测具有积极作用.

虽然本文以群体情绪稳态为主并结合用户和推特文本特征实现了有效地谣言检测, 然而一些关键用户会对群体情绪及其稳态形成产生影响, 这些用户发言一定程度会主导其他参与用户对事件的态度. 对此, 后续我们将研究关键用户对群体情绪稳态形成的影响机制, 探究关键用户影响下的群体情绪稳态特征. 我们注意到情绪强度计算可能还有其他更有效的情绪分析方法. 我们也认识到, 细粒度的情感分析可能为群体情绪稳态的研究带来新的视角, 未来研究中, 我们计划探索更细致的情感分类方法, 研究其对谣言检测性能的潜在提升, 以确保在性能提升的同时, 维持模型的高效性. 同时, 大语言模型近年来的显著进展, 其在文本分析任务中表现出强大的能力. 未来研究中, 我们将考虑大语言模型与群体情绪稳态特征相结合, 进一步探索这种集成方法在谣言检测中的潜力.

References:

[1] Yan YQ, Wang YJ, Zheng P. A graph-based pivotal semantic mining framework for rumor detection. Engineering Applications of Artificial Intelligence, 2023, 118: 105613. [doi: 10.1016/j.engappai.2022.105613]

[2] Li ZM, Zhao Y, Duan T, Dai JQ. Configurational patterns for COVID-19 related social media rumor refutation effectiveness enhancement based on machine learning and fsQCA. Information Processing & Management, 2023, 60(3): 103303. [doi: 10.1016/j.ipm.2023.103303]

[3] Li ZM, Du XY, Zhao Y, Tu Y, Lev B, Gan L. Lifecycle research of social media rumor refutation effectiveness based on machine learning and visualization technology. Information Processing & Management, 2022, 59(6): 103077. [doi: 10.1016/j.ipm.2022.103077]

[4] Hsu IC, Chang CC. Integrating machine learning and open data into social chatbot for filtering information rumor. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(1): 1023–1037. [doi: 10.1007/s12652-020-02119-3]

[5] Guacho GB, Abdali S, Shah N, Papalexakis EE. Semi-supervised content-based detection of misinformation via tensor embeddings. In: Proc. of the 2018 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining. Barcelona: IEEE, 2018. 322–325. [doi: 10.1109/ASONAM.2018.8508241]

[6] Liang G, Yang J, Xu C. Automatic rumors identification on Sina Weibo. In: Proc. of the 12th Int'l Conf. on Natural Computation, Fuzzy Systems and Knowledge Discovery. Changsha: IEEE, 2016. 1523–1531. [doi: 10.1109/FSKD.2016.7603402]

[7] Wu K, Yang S, Zhu KQ. False rumors detection on Sina Weibo by propagation structures. In: Proc. of the 31st IEEE Int'l Conf. on Data Engineering. Seoul: IEEE, 2015. 651–662. [doi: 10.1109/ICDE.2015.7113322]

[8] Zhang YS, Peng YY, Duan YX, Zheng J, You JQ. The method of Sina Weibo rumor detecting based on comment abnormality. Acta Automatica Sinica, 2020, 46(8): 1689–1702 (in Chinese with English abstract). [doi: 10.16383/j.aas.c180444]

[9] Nasir JA, Khan OS, Varlamis I. Fake news detection: A hybrid CNN-RNN based deep learning approach. Int'l Journal of Information Management Data Insights, 2021, 1(1): 100007. [doi: 10.1016/j.jjimei.2020.100007]

[10] Jain V, Kaliyar RK, Goswami A, Narang P, Sharma Y. AENeT: An attention-enabled neural architecture for fake news detection using contextual features. Neural Computing and Applications, 2022, 34(1): 771–782. [doi: 10.1007/s00521-021-06450-4]

[11] Dun YQ, Tu KF, Chen C, Hou CY, Yuan XJ. KAN: Knowledge-aware attention network for fake news detection. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2021. 81–89. [doi: 10.1609/aaai.v35i1.16080]

[12] Davoudi M, Moosavi MR, Sadreddini MH. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. Expert Systems with Applications, 2022, 198: 116635. [doi: 10.1016/j.eswa.2022.116635]

[13] Ma J, Gao W, Wong KF. Rumor detection on Twitter with tree-structured recursive neural networks. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 1980–1989. [doi: 10.18653/v1/P18-1184]

[14] Huang Q, Zhou C, Wu J, Liu LC, Wang B. Deep spatial-temporal structure learning for rumor detection on Twitter. Neural Computing and Applications, 2023, 35(18): 12995–13005. [doi: 10.1007/s00521-020-05236-4]

[15] Suthanthira Devi P, Karthika S. Rumor identification and verification for text in social media content. The Computer Journal, 2022, 65(2): 436–455. [doi: 10.1093/comjnl/bxab118]

[16] Xu F, Li MH, Huang Q, Yan KY, Wang MW, Zhou GD. Knowledge graph-driven graph neural network-based model for rumor detection. Scientia Sinica Informationis, 2023, 53(4): 663–681 (in Chinese with English abstract). [doi: 10.1360/SSI-2022-0170]

[17] Wang YN, Wang J, Wang HY, Zhang RL, Li M. Users' mobility enhances information diffusion in online social networks. Information Sciences, 2021, 546: 329–348. [doi: 10.1016/j.ins.2020.07.061]

[18] Xiong X, Qiao SJ, Wu T, Wu Y, Han N, Zhang HQ. Spatio-temporal feature based emotional contagion analysis and prediction model for online social networks. Acta Automatica Sinica, 2018, 44(12): 2290–2299 (in Chinese with English abstract). [doi: 10.16383/j.aas.2018.c170480]

[19] Li C, Peng H, Li JX, Sun LC, Lyu LJ, Wang LH, Yu PS, He LF. Joint stance and rumor detection in hierarchical heterogeneous graph. IEEE Trans. on Neural Networks and Learning Systems, 2022, 33(6): 2530–2542. [doi: 10.1109/TNNLS.2021.3114027]

[20] Hu HB, Chen WH, Hu YX. Opinion dynamics in social networks under the influence of mass media. Applied Mathematics and Computation, 2024, 482: 128976. [doi: 10.1016/j.amc.2024.128976]

[21] Cai YC, Wang HZ, Ye HL, Jin YW, Gao W. Depression detection on online social network with multivariate time series feature of user depressive symptoms. Expert Systems with Applications, 2023, 217: 119538. [doi: 10.1016/j.eswa.2023.119538]

[22] Liu ZY, Yang T, Chen W, Chen JC, Li QR, Zhang J. Sentiment analysis of social media comments based on multimodal attention fusion network. Applied Soft Computing, 2024, 164: 112011. [doi: 10.1016/j.asoc.2024.112011]

[23] Yan GH, Zhang XL, Pei HY, Li YY. An emotion-information spreading model in social media on multiplex networks. Communications in Nonlinear Science and Numerical Simulation, 2024, 138: 108251. [doi: 10.1016/j.cnsns.2024.108251]

[24] Indu V, Thampi SM. Misinformation detection in social networks using emotion analysis and user behavior analysis. Pattern Recognition Letters, 2024, 182: 60–66. [doi: 10.1016/j.patrec.2024.04.007]

[25] Subbaiah B, Murugesan K, Saravanan P, Marudhamuthu K. An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network. Artificial Intelligence Review, 2024, 57(2): 34. [doi: 10.1007/s10462-023-10645-7]

[26] Tiwari D, Nagpal B, Bhati BS, Mishra A, Kumar M. A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques. Artificial Intelligence Review, 2023, 56(11): 13407–13461. [doi: 10.1007/s10462-023-10472-w]

[27] Li WM, Li YQ, Liu W, Wang C. An influence maximization method based on crowd emotion under an emotion-based attribute social network. Information Processing & Management, 2022, 59(2): 102818. [doi: 10.1016/j.ipm.2021.102818]

[28] Qazvinian V, Rosengren E, Radev DR, Mei QZ. Rumor has it: Identifying misinformation in microblogs. In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011. 1589–1599.

[29] Lopez-Herrejon RE, Linsbauer L, Galindo JA, Parejo JA, Benavides D, Segura S, Egyed A. An assessment of search-based techniques for reverse engineering feature models. Journal of Systems and Software, 2015, 103: 353–369. [doi: 10.1016/j.jss.2014.10.037]

[30] Correa Bahnsen A, Aouada D, Stojanovic A, Ottersten B. Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 2016, 51: 134–142. [doi: 10.1016/j.eswa.2015.12.030]

[31] Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong KF, Cha M. Detecting rumors from microblogs with recurrent neural networks. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. New York: AAAI, 2016. 3818–3824.

[32] Wan PF, Wang XM, Wang XY, Wang L, Lin YG, Zhao W. Intervening coupling diffusion of competitive information in online social networks. IEEE Trans. on Knowledge and Data Engineering, 2021, 33(6): 2548–2559. [doi: 10.1109/TKDE.2019.2954901]

[33] Ma J, Gao W. Debunking rumors on Twitter with tree Transformer. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 5455–5466. [doi: 10.18653/v1/2020.coling-main.476]

[34] Silva A, Han Y, Luo L, Karunasekera S, Leckie C. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. Information Processing & Management, 2021, 58(5): 102618. [doi: 10.1016/j.ipm.2021.102618]

[35] Pröllochs N, Bär D, Feuerriegel S. Emotions in online rumor diffusion. EPJ Data Science, 2021, 10(1): 51. [doi: 10.1140/epjds/s13688-021-00307-5]

[36] Horner CG, Galletta D, Crawford J, Shirsat A. Emotions: The unexplored fuel of fake news on social media. Journal of Management Information Systems, 2021, 38(4): 1039–1066. [doi: 10.1080/07421222.2021.1990610]

[37] Alsaif HF, Aldossari HD. Review of stance detection for rumor verification in social media. Engineering Applications of Artificial Intelligence, 2023, 119: 105801. [doi: 10.1016/j.engappai.2022.105801]

[38] Faralli S, Rittinghaus S, Samsami N, Distante D, Rocha E. Emotional intensity-based success prediction model for crowdfunded campaigns. Information Processing & Management, 2021, 58(1): 102394. [doi: 10.1016/j.ipm.2020.102394]

[39] Bordoloi M, Biswas SK. Sentiment analysis: A survey on design framework, applications and future scopes. Artificial Intelligence Review, 2023, 56(11): 12505–12560. [doi: 10.1007/s10462-023-10442-2]

[40] Yang Z, Pang YC, Li Q, Wei SH, Wang R, Xiao YP. A model for early rumor detection base on topic-derived domain compensation and multi-user association. Expert Systems with Applications, 2024, 250: 123951. [doi: 10.1016/j.eswa.2024.123951]

[41] Zhu AQ, Zhang SX. An analysis method of user group's emotional tendency in hot topic of microblog. In: Abawajy JH, Choo KKR, Xu Z, Atiquzzaman M, eds. Proc. of the 2020 Int'l Conference on Applications and Techniques in Cyber Intelligence. Cham: Springer, 2021. 845–851. [doi: 10.1007/978-3-030-53980-1_124]

[42] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555, 2014.

[43] Bird S. NLTK: The natural language toolkit. In: Proc. of the 2006 Joint Conf. of the Int'l Committee on Computational Linguistics and the Association for Computational Linguistics on Interactive Presentation Sessions. Sydney: Association for Computational Linguistics, 2006. 69–72. [doi: 10.3115/1225403.1225421]

[44] Schneider N, Wooters C. The NLTK FrameNet API: Designing for discoverability with a rich linguistic resource. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations. Copenhagen: Association for Computational Linguistics, 2017. 1–6. [doi: 10.18653/v1/D17-2001]

[45] Bing R, Yuan G, Zhu M, Meng FR, Ma HF, Qiao SJ. Heterogeneous graph neural networks analysis: A survey of techniques, evaluations and applications. Artificial Intelligence Review, 2023, 56(8): 8003–8042. [doi: 10.1007/s10462-022-10375-2]

[46] Wang YH, Zhou YH, Mei YD. A joint attention enhancement network for text classification applied to citizen complaint reporting. Applied Intelligence, 2023, 53(16): 19255–19265. [doi: 10.1007/s10489-023-04490-y]

[47] Huang Q, Yu JS, Wu J, Wang B. Heterogeneous graph attention networks for early detection of rumors on Twitter. In: Proc. of the 2020 Int'l Joint Conf. on Neural Networks. Glasgow: IEEE, 2020. 1–8. [doi: 10.1109/IJCNN48605.2020.9207582]

[48] Tansey W, Veitch V, Zhang HR, Rabadan R, Blei DM. The holdout randomization test for feature selection in black box models. Journal of Computational and Graphical Statistics, 2022, 31(1): 151–162. [doi: 10.1080/10618600.2021.1923520]

附中文参考文献:

[8] 张仰森, 彭媛媛, 段宇翔, 郑佳, 尤建清. 基于评论异常度的新浪微博谣言识别方法. 自动化学报, 2020, 46(8): 1689–1702. [doi: 10.16383/j.aas.c180444]

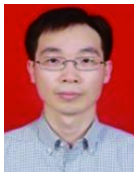
[16] 徐凡, 李明昊, 黄琪, 鄢克雨, 王明文, 周国栋. 知识图谱驱动的图卷积神经网络谣言检测模型. 中国科学: 信息科学, 2023, 53(4): 663–681. [doi: 10.1360/SSI-2022-0170]

[18] 熊熙, 乔少杰, 吴涛, 吴越, 韩楠, 张海清. 基于时空特征的社交网络情绪传播分析与预测模型. 自动化学报, 2018, 44(12): 2290–2299. [doi: 10.16383/j.aas.2018.c170480]

殷茗(1978－) 女, 博士, 副教授, 主要研究领域为智能数据分析, 社交网络, 文本计算.

陈威(2000－), 男, 硕士生, 主要研究领域为社交网络文本挖掘, 深度学习.

乔胜(1997－) 男, 硕士生, 主要研究领域为社交网络文本挖掘, 深度学习.



姜继娇(1979－), 男, 博士, 副教授, 主要研究领域为社交网络, 自然语言处理.