文章编号：1003-0530(2022)06-1155-15

基于传播意图特征的虚假新闻检测方法综述

毛震东[1] 赵博文[1] 白嘉萌[2] 胡 博[3]

（1. 中国科学技术大学网络空间安全学院，安徽合肥 230027；2. 合肥综合性国家科学中心人工智能研究院，安徽合肥 230088；3. 中国科学技术大学信息科学技术学院，安徽合肥 230027）

摘 要： 虚假新闻的传播会对个人、社会和国家产生巨大的负面影响，因此虚假新闻的检测始终都是研究的热点问题。虚假新闻检测实质上是一种信息分类问题，旨在验证由文本，图像和视频等多媒体信息构成的新闻的真实性。本文对虚假新闻检测问题和当前的主流方法展开了比较系统的研究，并揭示了虚假新闻的一个本质，即与报道真实事件的真实新闻不同，假新闻通常是有意为之，有特定的传播意图如误导公众等。基于这一特性，本文首先将虚假新闻的传播意图大致分为三类，并根据对应的相关特征对当前的研究方法作了分析，旨在能让读者从一个全新的角度理解虚假新闻检测领域。本文还介绍了虚假新闻检测的问题定义、基本范式、常用数据集和指标，并给出了该领域的未来的一些发展方向。

关键词： 虚假新闻检测；舆情分析；深度学习；特征提取

中图分类号： TP181 **文献标识码：** A **DOI：** 10. 16798/j. issn. 1003-0530. 2022. 06. 003

引用格式： 毛震东，赵博文，白嘉萌，等. 基于传播意图特征的虚假新闻检测方法综述[J]. 信号处理，2022，38(6)：1155-1169. DOI：10. 16798/j. issn. 1003-0530. 2022. 06. 003.

Reference format： MAO Zhendong，ZHAO Bowen，BAI Jiameng，et al. Review of fake news detection methods based on the features of propagation intention[J]. Journal of Signal Processing，2022，38(6)：1155-1169. DOI: 10. 16798/j. issn. 1003-0530. 2022. 06. 003.

Review of Fake News Detection Methods Based on the Features of Propagation Intention

MAO Zhendong[1] ZHAO Bowen[1] BAI Jiameng[2] HU Bo[3]

（1. School of Cyber Science and Technology，University of Science and Technology of China，Hefei，Anhui 230027，China；2. Institute of Artificial Intelligence，Hefei Comprehensive National Center，Hefei，Anhui 230088，China；3. School of Information Science and Technology，University of Science and Technology of China，Hefei，Anhui 230027，China）

Abstract: The spread of fake news has a great negative impact on personal development，social stability and national security. Therefore，the detection of fake news has always been a hot issue. Fake news detection is typically a classification problem aiming at verifying the authenticity of news which is composed of multimedia contents such as texts，images and videos. This paper conducted a systematic study on fake news detection and its current mainstream methods. We pointed out an intrinsic characteristic of fake news，that is，fake news usually has specific propagation intentions，such as misleading the public，which is different from truth. Based on this characteristic，we first generally divided the propagation intentions of fake news into three categories，and analyzed the current research methods according to the corresponding relevant characteristics，aiming to enable readers to better understand this field from a new perspective. This paper also introduced the problem definition，basic paradigm，common datasets，evaluation metrics and state-of-the-

收稿日期：2021-10-12；修回日期：2022-04-05

art performances of fake news detection, and outlined some potential directions for future research.

Key words: fake news detection; public opinion analysis; deep learning; feature extraction

1 引言

随着互联网通讯技术的蓬勃发展,社交网络日渐占据了人们的主要日常生活。微博,抖音,Facebook,Twitter等社交网络平台的兴起虽然能够给广大普通用户提供发表新闻资讯和撰写评论文章的机会,但也同时在极大程度上为虚假新闻的发布和传播创造了全新的渠道。虚假新闻的定义是,为了达到某一目的而发布不实信息以欺骗他人的一类报道。虚假新闻对个人,社会和国家会产生不可估量的负面影响。对个人而言,一条关于奥巴马在爆炸中受伤的假新闻就曾引发了股市的崩盘,这给不少人的财产带来了损失;对社会而言,自然灾害发生时往往会有假新闻出现造成群众的恐慌,例如2011年的日本地震[1],2012年的飓风桑迪[2];对于国家而言,2016年的美国总统大选期间出现了大量假新闻[3-4],这对选举结果造成了严重的影响。所以,如何在当下的社交网络媒体信息中快速准确地检测出虚假新闻,确保新闻传播的真实性,是当前社交媒体分析一个亟需解决的问题。

在微博、推特等社交平台中,用户可以向平台主动举报可能是假新闻的信息,平台通过人工审核的方式来判定被举报的信息是否为假新闻。这样的方法虽然可以一定程度减少虚假新闻的进一步扩散,但是这种机制依赖于人工审查和专家知识,在人工审查阶段虚假新闻可能已被广泛传播。针对这些问题,研究者们提出了自动智能检测虚假新闻的方法。早期的研究关注于手工设计特征,例如统计特征[2,5-10],主题特征[11-12],单词特征[9,12-15]和句法特征[16-18]等,之后使用提取的特征训练有监督[5-6,19-22]或者无监督[23-24]的分类器来对新闻进行分类。由于虚假新闻的传播模式与真实新闻有很大差异,研究者们也基于新闻传播树和传播图[3,11,13]开展了大量研究。随着深度学习技术的发展,大量的研究开始使用深度学习技术来进行特征提取和传播模式的建模[18-19,25-38]。

真实新闻是将真相传递给群众,而虚假新闻是有目的性,诱导性的信息,通常有特定的传播意图,这是虚假新闻和真实新闻本质的不同。检测不同传播意图的虚假新闻对应有不同的针对性特征。本文将不同意图的虚假新闻特征分为以下4类,各类特征将在之后的章节中详细展开介绍。

(1)通用特征:虚假新闻虽然有不同的传播意图,但仍有一些通用的特征适用于所有意图的检测。

(2)意图误导公众的特征:这类假新闻通常用于商业目的,此时虚假新闻和真实新闻内容非常相似,用户不易区分,从而诱导用户来点击虚假新闻获得盈利。

(3)意图操纵舆论的虚假新闻特征:这类假新闻通常用于政治用途,此时假新闻中带有强烈的煽动性和情绪性的词语,从而达到操纵舆论的作用。

(4)意图吸引注意的特征:包含这种特征的新闻用于经济或娱乐目的,例如新闻中会附加与内容不相关的图片、视频来吸引用户注意。

现有对虚假新闻检测的综述文献大多根据虚假新闻检测使用的技术进行分类。例如Zhou等人[39]将现有的方法分为基于知识图谱的,基于新闻风格,基于传播模式和基于可信度网络的研究。Shu等人[40]以数据挖掘的角度将虚假新闻检测分为基于特征提取和模型构造的研究。此外Zubiaga等人[41]还根据虚假新闻检测任务的目标,概述了现有研究,包括事件真伪检测、事件跟踪、立场分类和准确性分类等。与这些分类方法不同,本文揭示了虚假新闻及其传播过程中存在的内在特性,即虚假新闻具有一定传播意图且具有相应的特征。本文根据这一新视角对现有的方法按传播意图特征做了分类和比较,以便能够更好地指导该领域的未来发展。

本文后续内容安排如下:第2节给出了虚假新闻检测的定义和基本范式,并介绍了虚假新闻检测领域中常用的数据集和评价指标;第3节将现有方法根据传播意图特征进行分类和比较;第4节介绍了基于特征的虚假新闻检测方法;第5节对当前方法的性能作了简要介绍;第6节对虚假新闻检测领域将来的工作进行了展望;第7节对基于传播意图

特征的虚假新闻检测技术作了总结。

2　虚假新闻检测

2.1　虚假新闻检测的问题定义

虚假新闻是指以不实信息误导大众，以带来政治效果或经济利益的新闻，是一种具有明显传播意图的不实信息。虚假新闻检测本质上是分类问题，其形式化定义如下：

给定一个事件 X，与 X 相关的一系列新闻 $M=\{m_1, m_2, \cdots, m_n\}$，这些新闻由一组用户 $U=\{u_1, u_2, \cdots, u_n\}$ 发布，新闻的形式可以是微博、推特等互联网社交媒体内容及其后续的转发、评论，也可以是发布者发表的文章、视频等。虚假新闻检测的目标是学习一个函数f(x)来判别该事件或某条新闻是否为不实内容。每条新闻都包含一组多媒体内容，包括文本描述，视觉内容和用户信息。其中文本描述包括原始新闻的文本内容、评论内容以及转发内容，视觉内容包括该新闻中的表情、图片、视频等信息，用户信息包括用户画像和社交关系。

2.2　虚假新闻检测的基本范式

虚假新闻检测的基本范式可以分为两步：相关特征的提取和对特征的编码分类，如图 1 所示。在第一步特征提取阶段，除了提取通用的特征如语义特征、主题特征、情感特征、视觉特征、时序特征和用户特征外，还可以根据不同的传播意图提取特定的特征。基于提取的特征，第二步将特征进一步编码并使用分类器来判别新闻的真假性。现有的分类方法可以分为传统机器学习分类和基于神经网络的方法。传统的机器学习方法首先选择合适的特征，再使用有监督或无监督的方法进行分类。基于神经网络的方法能够自动地进行新闻特征选择和融合，再进行分类，并结合卷积神经网络和循环神经网络处理新闻内容中的图像特征，文本特征和时序特征等多维度多模态特征。对于这个基本范式中的两个步骤我们将在后续章节详细介绍。

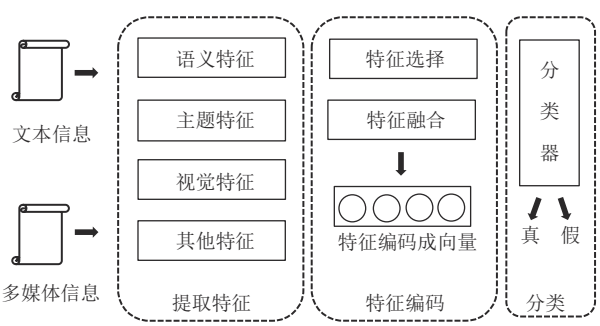


图 1　虚假新闻检测的基本范式

Fig. 1　The basic paradigm of false news detection

2.3　数据集

随着社会对虚假新闻检测研究的增多，研究者们制作了大量的基准数据集，这些数据集大多是从现实生活中的社交媒体上收集的，例如Twitter，新浪微博等。本文整理了一组虚假新闻检测领域经典的数据集，统计结果如表 1 所示。

表 1　虚假新闻检测领域经典数据集

Tab. 1　Common datasets for fake news detection

数据集名称	新闻总数	事件总数	虚假新闻条数	类型
微博[28]	3805656	4664	2313	文本、图像
Twitter15[6]	1490	1490	370	文本
Twitter16[28]	818	818	205	文本
BuzzFeedNews[42]	1627	1627	901	文本
LIAR[43]	12836	—	—	文本
FNC-1[30]	75385	2587	—	文本
FakeNewsNet[44]	23196	23196	5755	文本、图像
MediaEval[45]	15000	17	9000	文本、图像
CCMR[32]	15629	17	9404	文本

（1）微博[28]：该数据集是从新浪微博社区收集的信息，每一条微博都视为一条新的信息，并且与一个二元标签相关联，以表明该信息是否为真。数据集包含 4664 个事件，3805656 条信息，2746818 名用户，其中 2313 个事件是假新闻。

（2）Twitter15[6]和Twitter16[28]：该数据集是从美国Twitter公司收集的信息，也是虚假新闻检测领域最常用的数据集。前者包含 1490 条新闻和 276663 名用户，后者包含 818 条新闻和 173487 名用户。

（3）BuzzFeedNews[42]数据集包含新闻故事的标题和文本，这些新闻来自于 Facebook，内容为 2016 年美国大选相关的新闻文章，总共有 1627 条新闻和 901 条虚假新闻。其中包含 826 篇主流文章、356 篇左翼文章和 545 篇右翼文章。

（4）LIAR[43]是一个数量高达 12836 条的虚假新闻数据集，是从事实审核平台 PolitiFact 收集的 2007-2016 年间发布在该网站上的简短言论，包含一系列民主党人和共和党人发布的言论，有六种关

于真实度的标签。

(5)FNC-1[30]:该数据集包含了虚假新闻检测挑战赛的300个主题,每个主题与5~20篇新闻文章相关,总共有约2587篇文章和75385个事件。每个文档-标题对都用四种立场标签之一标注。训练集和测试集分别包含200个主题和100个主题的文档-标题对。

(6)FakeNewsNet[44]数据集收集自两个事实审核平台:GossipCop和PolitiFact。它包含23196篇新闻文章,其中5755篇是虚假新闻。此外,该数据集包含三种信息:带有标签的新闻内容,社会背景信息和时序信息。此外部分新闻还对应有相关图像。

(7)MediaEval[45]:该数据集用于检测社交媒体上的虚假多模态内容。其训练集包含与17个谣言相关事件的9000条虚假信息和6000条真实信息。其测试集包含与35个谣言相关的2000条新闻。每条推特信息中都包含文字内容、相关的图片视频内容和社交内容。

(8)CCMR[32]:该数据集是一个跨语言和跨平台的多媒体谣言验证数据集。它通过收集不同的搜索引擎返回的外部网页,扩展了MediaEval[45]数据集。CCMR共有15629条推特新闻,4625条谷歌网页,2505条百度网页,这些数据都与17个事件相关。

2.4 评价指标

虚假新闻检测是一项分类任务,评价指标包括精确率(Precision)、召回率(Recall)、F1值(F1 Score)和准确率(Accuracy)。首先要构造混淆矩阵,并计算真值。混淆矩阵中各项的定义如下:

(1)真阳性(True Positive,TP):待预测新闻是假新闻,预测结果为假新闻。

(2)真阴性(True Negative,TN):待预测新闻是真新闻,预测结果为真新闻。

(3)假阳性(False Positive,FP):待预测新闻是真新闻,预测结果为假新闻。

(4)假阴性(False Negative,FN):待预测新闻是假新闻,预测结果为真新闻。

根据上述定义,评价指标可以被定义以下公式(1)、(2)、(3)、(4):

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \tag{1}$$

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|} \tag{2}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \tag{4}$$

3 基于不同传播意图提取的虚假新闻特征

虚假新闻和真实新闻的区别在于虚假新闻是带有一定的传播意图,针对虚假新闻的传播意图可以提取相应特征用于虚假新闻检测。我们将这些特征分为通用特征、意图误导公众的特征、意图操纵舆论的特征和意图吸引注意的特征,如表2所示。我们将在后续小节详细介绍。

表2 不同传播意图的虚假新闻特征

Tab. 2 False news features based on different propagation intentions

传播意图	特征	代表文献
新闻通用特征	结构特征	[5]、[14]
	时间特征	[9]、[13]、[14]
	用户特征	[6]、[7]、[9]、[10]、[46]
	其他特征	[15]、[17]、[18]、[47]、[48]、[49]
意图误导公众	特殊符号统计特征	[2]、[5]、[6]、[7]、[8]
意图操纵舆论	情感特征	[8]、[9]、[12]、[15]、[50]、[51]、[52]
	风格特征	[53]、[54]
意图吸引注意	主题特征	[11]、[12]、[55]
	视觉特征	[9]、[16]、[45]、[49]、[56]
	点击诱饵特征	[8]、[57]

3.1 通用特征

通用特征指的是虚假新闻检测中具有普适性的特征。我们将通用特征具体分为结构特征、时间特征、用户特征和其他特征。

(1)结构特征:虚假新闻是被用户故意散播的,所以假新闻的传播过程与真实新闻有很大不同。研究者们提出可以建模新闻的传播路径,来获取传播过程中的结构特征。Castillo[5]提出可以获取新闻传播树中的结构特征,新闻传播树是指以某条新闻为根节点,该条新闻的转发新闻为子节点,并将子节点作为下一层的根节点按照上述过程迭代构建的树,此树结构可以代表新闻的传播路径。新闻传

播树中的最大深度和平均深度，根节点的度数，传播树中最大度数和平均度数都能反映出新闻的传播特征，文章提到，传播图中单层节点数目较大的新闻往往可信度较高。Kwon 等人[14]提出可以从三种类型的网络中提取特征，例如好友网络，好友网络最大连接图，以及新闻扩散图（类似上述新闻传播树定义，描述新闻主题在用户之间的扩散路径）。通过分析这些网络结构，他们得出结论，对于给定的新闻，如果新闻的传播方向是从度数低的用户到度数高的用户，或者在扩散图中孤点的比例很大（例如很多僵尸用户会发表虚假新闻获利，而这些用户没有社交网络），那么该条新闻很有可能是假的。

（2）时间特征：虚假新闻的始作俑者一旦发布虚假新闻后，会尽可能地使其流传并变得热门起来，因此虚假新闻随时间的传播特征与真实新闻不同。Kwon 等人[14]提取了新闻传播过程中的时间特征，他们观察到虚假新闻通常有多个周期性的转发和评论数量峰值，而真实新闻通常只有一个峰值。类似地，Wu 等人[13]提出了计算新闻的转发时间特征，也就是原始新闻和其被转发的平均时间差。因为恶意用户会刻意地反复转发虚假新闻，并评论相似的内容来提高虚假新闻的热度。Sun 等人[9]提出计算新闻的重复数，并作为一项特征。他们通过计算关键词的 Jaccard 系数来衡量两条新闻间的相似性，只要相似性超过预定义的阈值，则新闻将被视为重复。

（3）用户特征：用户故意散播虚假新闻的目的和行为不同，但有时会表现出类似的用户信息特征，通常可以从社会声誉和个人信息两个方面提取。一般来说，社会声誉高的用户不太可能发布假新闻，所以有时恶意用户会使用和声誉高的用户相似的昵称，以达到混淆视听的作用。为了克服这个问题，许多研究提出可以将用户的社会声誉作为用户特征。Gupta 等人[7]提出可以考虑用户的好友、关注者、粉丝的数量和社会地位，并检查用户是否被社交媒体验证为可信任用户（例如微博上的大 V），这些信息反映了一个用户是否可信。Sun 等人[9]提出如果一个用户很少被人关注但是其关注了很多其他用户，那么该用户很可能是虚假新闻的传播者。除此之外，他们还提到可以统计用户发布新闻中包含的强烈否定词和事件相关动词（即常用于事件描述而非日常生活的动词）的比例，这些比例越高，该用户是假新闻发布者的可能性就越大。

虚假新闻传播者倾向于隐藏个人信息，也就是说他们拥有的个人信息是不完整的，对此，许多的研究者展开了研究。Gupta 等人[7]提出虚假新闻发布者很可能是最近注册的新用户，因此，可以将用户的注册时间、个人描述、图片资料和位置信息作为用户特征。文中提出，虚假新闻的发布者和传播者通常是最近注册的用户，个人描述和图片资料信息较少，不存在定位信息。他们还检查了不同社交媒体上的个人资料是否会链接在一起，因为普通用户总是为了方便而链接它们，而虚假新闻发布者不会。此外，推文的位置、个人资料位置和事件位置的一致性也具有指示性[6]。Yang 等人[10]发现通过检测用户登录的客户端平台能够很好地检测新浪微博中的虚假新闻。客户端程序包括非移动端程序和移动客户端程序，其中非移动端包含网页版新浪微博，定时发帖工具和第三方应用程序；移动客户端程序是指安装在用户手机或平板中的应用。该文献指出如果一条新闻涉及到国外事件，并且从非移动客户端程序发出，那么该新闻有很大概率是虚假新闻。

（4）其他特征：除上述介绍的特征外，还有一些特征能够用于假新闻检测。例如 Hassan 等人[17]提出计算 TF-IDF（term frequency-inverse document frequency）值，这是衡量句子中每个单词的重要性的统计性数值。这可以帮助我们关注一些在假新闻经常使用但在真实新闻中很少使用的词，例如“震惊的”，“难以想象”等夸张的词语。Chen 等人[18]提出提取 TF-IDF 特征。他们首先构建一个包含所有新闻中 K 个关键词汇的字典，计算这些词汇的 TF-IDF 值。将每条新闻使用 TF-IDF 编码为一个 K 维向量，如果新闻中未出现某关键词汇，那么该维度为 0，反之该维度为预先计算好的 TF-IDF 值。文献[15]中的工作使用 Stanford 解析器基于上下文无关语法（CFG）树导出了一组包括所有词汇生成的规则，这些规则与父节点和祖父节点结合后编码为 TF-IDF 特征。同样地，词袋模型[46]，词性标注[17,47]，命名实体识别[48]等技术也被用来分析新闻中的关键字。上述特征的向量化表示，可以作为深度神经网络的

输入进行虚假新闻分类检测。

3.2 意图误导公众的相关特征

此类虚假新闻的传播意图在于误导公众，通常有一定的商业倾向，例如，让公众无法明确分辨消息的真实性，从而让用户更有可能购买特定产品。在这种情况下，虚假新闻的文本内容与真实新闻非常相似，除了在使用的符号或形式的统计特征上存在细微的差异，因此针对这一类虚假新闻的检测主要集中于提取特殊符号的特征。

提取特殊符号的特征主要关注的是捕捉一些通常用来误导公众的特殊单词或字符。例如，Castillo等人[5]提出关注新闻文本的长度，以及文本中是否包含问号或感叹号。他们注意到此类虚假新闻不仅会使用特殊符号以误导读者，一般还具有相似的长度。Gupta等人[2]和Castillo等人[5]提出对第一、第二、第三人称代词计数，Liu等人[6]提出验证信息中是否包含感官短语，如“I see”，“I hear”等等。这是因为如果包含此类单词虚假新闻看起来会更具有可信度。此外，Gupta等人[7]认为网络推文中的外部统一资源定位符（Uniform Resource Locator，URL）可以作为一个值得关注的证据，Biyani等人[8]从URL中提取了一些特征，如破折号、大写字母、逗号的频率等。还有一些工作考虑到了新闻带来的影响并据此计算了“@”标签、评论和转发的数量。

3.3 意图操纵舆论的相关特征

此类虚假新闻经常被用来操纵人们的观点，特别是为了政治目的。为了影响人们的态度或观点，此类新闻会使用许多情绪化的词语和特定的写作风格，因此此类虚假新闻的检测主要关注情感特征和风格特征。

（1）情感特征：情感特征可用于识别那些包含情绪化单词和句子、意图操纵舆论的煽动性言论[8-9,12-15,17,49-50]。为了提取此类特征，人们使用了许多情感分析工具。例如，Kwon等人[14]和Pérez-Rosas等人[15]都利用了一种被称为语言探究和单词计数（Linguistic Inquiry and Word Count，LIWC）的情感工具，以统计有特殊心理学含义的单词的数量。在此基础上，更多工作进一步研究提取了大量与情感相关的统计性特征。对于新浪微博的新闻，Sun等人[9]考虑了新闻是否包含强烈且负面的情绪词和意见词，Wu等人[13]利用单条消息中积极或消极的情绪词的数量计算该新闻的平均情绪得分。对于推特的新闻，Ma等人[12]提出使用多视角问答（Multiple-Perspective QA 3，MPQA3）情感词典和一些手动收集的常用表情符号来识别积极或消极的单词。

虚假新闻更喜欢使用一些情感极端的副词或形容词，因此命名实体识别（Named Entity Recognition，NER）技术[51]和词性（Parts of Speech，POS）相关的技术[17,51]得到了广泛应用。Hassan等人[17]提出利用自然语言处理工具包（Natural Language Toolkit，NLTK）标记提取POS特征。他们在语料库中收集了43个POS标签，并计算每个句子中属于这些标签的单词数量。此外，考虑到很多短语被反复提及以加强印象，Biyani等人[8]还提取了unigram和bigram特征用于虚假新闻检测。

（2）风格特征：风格特征用于建模新闻的书写样式和风格。因为为了操纵舆论，虚假新闻通常以独特的文体风格撰写，特别是对某一政党极度拥护的新闻更可能是虚假新闻，它们尤其倾向于操纵用户意见，其写作风格也十分不同于主流新闻。Potthast等人[52]分析了这类新闻的写作风格，并发现了其特征规律，他们基于文献[50]中提出的方案，鉴定两篇文章是否由极端政党分子编写。具体地说，给定两篇文章，每篇文章首先被划分为一组具有固定字数的小文本块。之后对于这两组文本块，通过迭代地移除这两组块中最具辨别力的特征来计算重构误差。最后从重构误差曲线可以看出，如果两篇文章属于同一作者，重构误差曲线将急剧下降。实验结果表明，左翼和右翼极端政党分子的新闻写作风格是相似的，但与主流真实新闻不同。因此，这种特征可用于检测政治性或类似意图操纵群众观点的新闻。

3.4 意图吸引注意的相关特征

这类虚假新闻主要用于商业或娱乐目的，如增加流量、点击率或制造轰动等，因此热点话题、图片和点击诱饵往往会出现在假新闻中。基于此，提取主题特征、视觉特征和点击诱饵特征来区分此类假新闻和真新闻是十分有效的。

（1）主题特征：一些虚假新闻倾向于利用耸人听闻的话题来吸引用户的兴趣，例如名人离婚或怀孕以及空难事故（例如“MH370航班失联”）。根据这一观察，一种直观的方法是将新闻按不同的主题

聚类，然后关注热点话题。例如，Jin 等人[11]提出使用聚类算法将新闻聚类成子事件，以子事件-中心事件的层级形式对主题信息进行分析。Ma 等人[12]提出了一种基于动态序列时间结构（Dynamic Series-Time Structure，DSTS）的分类器来检测虚假新闻。新闻主题的特征分布会随时间变化而变化（即消息传播过程），为了捕获这种特性以提高检测性能，该分类器使用隐含 Dirichlet 分布（Latent Dirichlet Allocation，LDA）模型[53]计算每种新闻的主题特征分布，之后通过捕捉随时间变化不同社交语境下的主题特征的变化实现虚假新闻的检测。。

（2）视觉特征：某些虚假新闻还倾向于关联图片或视频作为额外的视觉描述，这种图文并茂的方式比单纯的文本内容更能吸引眼球，因此可以使用视觉特征来验证消息的真实性。本文将视觉特征大抵分为视觉统计特征和视觉语义特征，前者侧重于统计分布，后者侧重于视觉内容的语义。

视觉统计特征用来检测带有过时或篡改图像的假新闻。过时图片是指以前曾在互联网上发布过的图片。为了确定图像是否过时，Sun 等人[9]计算了图像的时间跨度，即新闻发布时间（带有该图像的消息）与该图像原始发布时间的时间跨度。在这项工作中，作者使用一个图像搜索引擎从互联网上检索该图像的所有记录，并按时间顺序对搜索结果进行排序，时间最早的条目确定了该图像的原始发布时间。如果时间跨度大于预定义的阈值，则图像被视为过时，相应的新闻有较大概率可能是条假新闻。

篡改图像特征的研究要更为复杂。篡改图像的操作可分为三种类型：拼接、复制粘贴和修饰。拼接是指将另一幅图像中的对象添加到目标图像；复制粘贴是指将同一图像中的对象添加到不同位置；修饰指增强对比度、锐化边缘或使用滤色器。为了判断一张图像是否被篡改，一些工作[49]提出设计取证特征来评估图像的真实性，如对齐双 JPEG 压缩的概率图。然而，这些类型的特征在对来自社交网络的图像进行篡改检测时是不起作用的，因为这些图像通常经历多次重新保存过程，破坏了图像的取证痕迹。

与此不同，Jin 等人[54]提出了以下五种视觉特征来衡量图像分布：视觉清晰度得分、视觉连贯性得分、视觉相似性分布直方图、视觉多样性得分和视觉聚类得分。在这项工作中，关于同一事件的相关新闻片段被集中在一起，用于事件级别虚假新闻检测。这些特征的详细信息如下所示：

a）视觉清晰度得分：主要描述两组图像集之间的 Kullback-Leibler 散度。一组是针对目标事件的，包括有关该事件的所有新闻图像，另一组包括所有事件新闻中的所有图像，视觉清晰度评分衡量了这两个集合的分布差异。如果目标事件是真实事件，通常它包含的图像有不同的来源，其图像分布趋于一般化，相应的视觉清晰度分数则较低，而虚假事件通常具有有限的图像来源，其图像分布往往不同于平均值，因此具有较高的视觉清晰度分数。

b）视觉连贯性得分：视觉连贯性得分定义为事件中任意两幅图像之间的平均余弦相似度，主要衡量事件中图像的连贯性。如果事件相关的图像在视觉上非常相似，事件就可能是假的，因此通常假新闻事件的视觉连贯性得分较高。在这项工作中，利用到 GIST[58]模型为每幅图像提取了 512 维全局特征向量，便于计算一对图像的余弦相似度。

c）视觉相似度分布直方图：以细粒度级别衡量事件中图像的一致性。具体来说，首先根据受欢迎程度对图像进行排名（受欢迎程度与回复和评论数量呈正相关），然后使用在视觉连贯性评分中相同的计算方法，得到成对的相似度矩阵，最后将该矩阵元素的值映射到 H-bin 直方图中。

d）视觉多样性得分：主要衡量事件中图像的多样性。与视觉连贯性得分不同，它计算的是所有图像对的相异性加权平均值，其中代表性图像（受欢迎程度较高的图像）起着更重要的作用。通常，假新闻事件的视觉多样性得分较低，因为它们的图像多样性较低。

e）视觉聚类得分：主要衡量事件图像分布的聚类簇指标。该方法使用自底向上的聚类方法迭代合并最近的原子簇，合并标准是 GIST 特征的最近欧氏距离。虚假新闻事件的图像会比真实新闻事件的图像形成更少的簇，所以此特征可用于检测虚假新闻事件。

视觉语义特征旨在通过检测视觉内容、文本内容和事件在语义层面是否一致来检测虚假新闻。通常来讲，虚假新闻倾向于附加图片以增加其可信

度，然而这些图片实际上一般与新闻事件无关，文献[45]也说明了这一现象。为了检测带有图像的虚假新闻，Sun 等人[9]首先使用附加的图片作为查询，从搜索引擎中检索出类似图片并返回一组基于可信度排名的网站，然后从排名靠前的网站上爬取文本信息，最后计算新闻文本和上面爬取的文本之间的 Jaccard 系数。如果 Jaccard 系数的值较低，则该新闻被视为文本图像不匹配的虚假新闻。

此外，视觉内容也有助于将新闻分组以实现组级虚假新闻检测。具体而言，Jin 等人[16]提出将具有相同图像或视频的新闻划分到一个组，然后将同一组中新闻的特征聚合起来，用于组级虚假新闻检测。除此之外，一些工作[25-27,59]提出使用深度神经网络来提取视觉语义特征。例如，Jin 等人[25]设计了一种多模态融合网络可以利用图像特征作为检测虚假新闻的辅助线索。首先利用一个以 VGG-19 模型[60]作为主干网络并添加两个全连接层的子模型来提取 512 维的视觉表征，然后使用注意力机制把提取的视觉表征与文本表征进行聚合并连接，以推断虚假新闻的概率。此外，Qi 等人[59]还提出了结合新闻图像的频率域和像素域来提取视觉特征，用于虚假新闻检测。

(3)点击诱饵特征：某些虚假新闻倾向于使用耸人听闻的标题诱使用户点击特定网页，例如“震惊！美国人不再喝啤酒”。这类文章没有专业文章那么正式，可读性也更高。为了检测点击诱饵，Biyani 等人[8]提取非正式性和可读性的统计特征以区分点击诱饵，例如是否包含网络俚语或脏话，是否使用重复字符(如“ooh”、“aah”等)，以及标题和首句之间的相似性。此外，他们还进一步设计了衡量非正式程度和可读性程度的指标，计算如下：

a)Coleman-Liau 分数(CLScore)：根据人为经验计算阅读难度，公式为：

$$\mathrm{CLScore} = 0.0588L - 0.296S - 15.8 \tag{5}$$

其中 L 表示每 100 个单词所含字母的平均数量，S 表示每 100 个单词所含句子的平均数量。

b)RIX 和 LIX 指标：衡量可读性，公式为：

$$\mathrm{RIX} = \frac{LW}{S} \tag{6}$$

$$\mathrm{LIX} = \frac{W}{S} + \frac{100LW}{W} \tag{7}$$

其中 W 是单词计数，LW 是长单词(即超过 6 个字符)计数，S 是句子计数。

c)正式性度量(F-measure)通过计数文章中不同的词性标签，如名词、动词和形容词，来衡量正式程度。

除了上述的指标，新闻标题的句法结构风格也可以用于点击诱饵检测。有一种风格叫做前向指代(forward-reference)[57]，这类新闻标题通常挑逗性很强或者标题与文章之间有明显的信息差距。例如，给定一个标题：“这是最可怕的骗局”，用户可能很想知道“这”是指什么，因此点击网页。文献[8]的工作表明，前向指代通常以指示代词、人称代词、副词和定冠词为特征，可用于标题党检测。

4 基于特征的虚假新闻检测方法

基于上述提取的特征，分类算法可以用于进一步检测虚假新闻。相关的研究工作从传统的机器学习方法到最近的基于神经网络的方法层出不穷，如图 2 所示。传统的机器学习方法首先进行特征选择，然后进行分类，而基于神经网络的方法学习自动地进行特征选择、融合并分类。

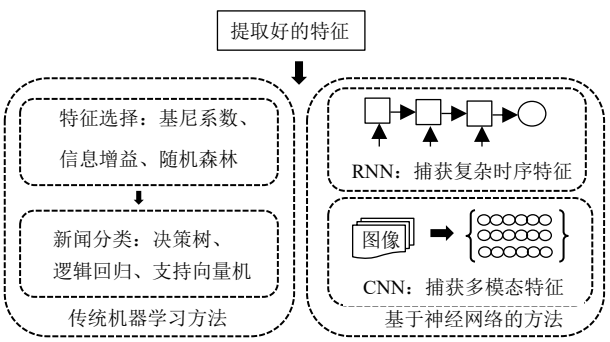


图 2 不同的虚假新闻检测方法

Fig. 2 Different fake news detection methods

4.1 传统机器学习方法

基于前文介绍的特征提取方法，传统的机器学习方法研究如何选择最合适的特征和分类器实现虚假新闻检测。特征的选择方法旨在降低特征维数并保留信息性特征，包括基尼指数(Gini index)、信息增益(information gain)和随机森林(random forest)等。例如，文献[5]、[6]、[17]使用基尼指数研究特征在构建决策树中的重要性。Kwon 等人[14]使用随机森林和逻辑模型寻找最有信息量的特征。具体地，他们重复进行 2 倍交叉验证并从特征集中

依次减少特征，以找到最重要的特征。Biyani 等人[8]利用信息增益对特征进行排序，并丢弃信息增益为零的特征。

Castillo 等人[5]则重点研究了 Twitter 上热点新闻爆发式传播时的时效性特征对新闻可信度评估的作用。他们首先在推特上爬取了两个月的数据，并人工区分为有新闻价值主题的信息和个人观点，随后又按范围不同提取四类不同的特征：基于新闻消息的特征、基于用户的特征、基于内容主题的特征和基于传播路径的特征。基于新闻消息的特征代表消息本身的一些性质，如推文长度、是否包含感叹号和问号、正面/负面情感词的数量、是否包含主题标签、是否为转发等；基于用户的特征代表发信息的用户自身的特征，如注册时长、粉丝数、关注数、原创推文数等；基于内容主题的特征是前两个特征经过计算的聚合，例如，带有主题标签的推文比例、包含 URL 的推文比例以及一组信息中正面和负面情绪词的比例；基于传播路径的特征则是包括树的深度或主题的初始推文数量等与消息转发传播路径树相关的特征。为了研究不同特征对新闻可信度评估的作用，他们使用最佳优先选择策略和决策树对上述四大类特征中的 15 种进行了分析验证，最终有三点发现：1）基于消息主题的特征（情绪信息、URL 等）与此任务非常相关，例如假设一条推文不包含任何 URL，那它有很大概率是不可信的新闻。2）基于用户的特征也有很高的相关性。例如通过可信用户（具有大量社交连接的活跃用户）传播、且有大量转发的消息通常可以被认为是高度可信的，因为这些用户为了自身的声誉倾向于传播可信的消息。3）在基于传播的特征中，转发数也是很重要的评判依据，有很多次转发的推文更有可能是真实新闻。这篇文章重点研究了大量不同的特征对新闻可信度的作用，提供了大量的数据和比较，有非常重要的研究意义。

对于上述特征，多种机器学习方法都可用于实现虚假新闻分类，如决策树（Decision Tree）、随机森林、梯度增强决策树（Gradient Boosted Decision Trees，GBDT）[20]、逻辑回归（Logistic Regression）、最大熵分类器（Max-Entropy classifier）[19]和不同核的支持向量机等。Wang 等人[20]构建了一个端到端的系统，用于自动判断一篇文章所含内容的真伪并对立场分类。首先对于给定的一篇事实检查文章生成综合候选集，然后基于频率分析构建了一个相对较小的表示矛盾的词汇库，根据矛盾词汇表，对文章关键成分计算 n-gram 权重向量，最后构建了一个梯度增强的决策树模型来预测相关文档是支持文档还是矛盾文档。此外还有一些工作为虚假新闻检测制定了特定的规则。Ciampaglia 等人[61]提出了一种语义接近度度量，该度量通过在知识图谱上查找概念节点（由提取的文本特征表示）之间的最短路径来执行事实检查。Wang 等人[62]通过屏蔽一部分用户来最小化虚假新闻的影响。他们通过综合考虑虚假新闻的全球流行度和对个体的吸引力，提出了一种动态伊辛传播模型（dynamic Ising propagation model）以同时减少虚假新闻的影响并维持用户体验。文献[21]、[22]则基于外部知识库研究基于事实的检查，从现有事实推断新闻的准确性。例如，Shi 等人[21]将事实检查视为从 Wikipedia 和 SemMedDB 知识库中提取的知识图谱的链接预测任务（link prediction），他们采用类似深度优先搜索（Depth-First-Search，DFS）的图遍历算法来检索元路径，并提取前 k 个判别路径作为特征来训练逻辑回归模型。这种方法为所分析事实的具体语义提供了一种可解释的、直观的解释，并且可以通过调查回归变量来描述所述事实是真是假。

4.2　基于神经网络的方法

受最近关于深度学习的研究进展的启发，大量的研究也逐渐聚焦于利用深度神经网络来检测虚假新闻。这类方法旨在使用一个网络结构学习自动地特征选择、特征编码、分类，实现端到端的虚假新闻检测。根据网络结构，这种方法大体可以分为两类，基于循环神经网络（Recurrent Neural Networks，RNN）的方法和基于卷积神经网络（Convolutional Neural Networks，CNN）的方法。

首先介绍基于 RNN 的研究工作。一些研究者研究了不同的手工提取特征和不同的神经网络结构对虚假新闻检测的影响。Volkova 等人[29]进行了大量的实验，证明了语言特征是信息量最大的，并且可以使用后期融合技术将其融合进神经网络以提高检测性能。Hanselowski 等人[30]介绍了他们在文章立场分类任务方面的工作，该任务被视为虚假新闻挑战赛（Fake News Challenge，FNC-1）中虚假新

闻检测的第一步。他们首先对一组手工提取的特征进行消融实验以选择最重要的特征，然后提出了一种多特征的层叠长短期记忆网络（stack Long Short-Term Memory，stackLSTM），该网络能够融合上述选定的特征以获得良好的结果。

一些工作还利用深层神经网络实现提取特征。在早期阶段，大量基于RNN的方法[18-19,28-34]关注于捕获虚假新闻随时间的变化。鉴于RNN可以识别证据的远距离相关性，Ma等人[28]提出使用RNN识别虚假新闻，首先将相关的新闻按不同时间间隔划分成组，并计算组中词汇术语的前k个TF-IDF值作为每个RNN单元的输入。通过最小化预测概率分布与真值之间的平方误差，该模型能够很好地区分虚假新闻和真实新闻。Rashkin等人[19]提出了一个LSTM模型，该模型以单词序列为输入，并将新闻的可靠性分为不同类别，即可信的、讽刺的、恶作剧和宣传性质的。Ruchansky等人[31]提出了一种混合深度模型，该模型结合了文本内容、用户反应和源用户信息以实现更准确的虚假新闻检测。混合模型由三个关键模块组成：捕获、评分和集成。捕获模块利用LSTM捕获用户响应的文本和时间特征，评分模块学习用户信息的表示并给每个用户打分。这两个模块进一步集成在第三个模块中以完成分类。

一些研究者还研究了基于某些特定类型的特征实现更细粒度的分类。Wen等人[32]提出利用门控循环单元（Gated Recurrent Unit，GRU）提取额外的多语言跨平台特征，该特征能捕捉到虚假新闻和来自不同社交媒体平台和不同语言的相应评论之间的一致性。文献[18,33-34]提出的方法聚焦于尤其突出的特征，例如情感词和挑逗性的句子，这些方法都采用了注意力机制。例如，Chen等人[18]将软注意力机制应用于RNN，使其可以同时关注特定的独特特征，并捕捉信息随时间的上下文变化，结果表明，该方法能够快速准确地检测出虚假新闻。此外有些工作并没有关注关键特征，而是关注了关键句子。De Sarkar等人[34]提出了一种用于讽刺性新闻检测的层级注意模型，该模型选择性地捕捉文档中的关键句子，没有使用手工制作的特征，仅将词语嵌入作为输入，便取得了很好的效果，这表明词语级别语义信息足以检测讽刺性新闻。

此外，近年来还有一些工作通过建模特定特征拓展了新的研究方向。例如，Shu等人[63]首次在社交媒体虚假新闻检测领域提出了具有解释性的模型dEFEND，该模型建模了新闻中的文本内容和用户评论间的关联，由新闻内容编码模块、用户评论编码模块和联合注意力模块构成。新闻内容编码模块通过从单词（Word Encoder）到句子（Sentence Encoder）的层级注意力神经网络在不同尺度获取新闻句子中的语法信息和句法信息，以得到新闻句子的向量表示。用户评论编码模块（Comment Encoder）通过多个注意力子网络在单词级别获取用户评论的隐层向量表示。联合注意力模块（Sentence-Comment Co-attention）通过学习捕捉新闻内容向量和评论文本向量之间的相关性，以筛选出有解释性的新闻句子和评论。本文所研究的可解释性的依据在于用户的评论通常含有一定的解释性证据，但有时错误的观点也会以一些真实的新闻内容中作为依据以迷惑他人，因此新闻中的句子也有一定的重要性。最后分类器将新闻内容特征和用户评论特征的拼接结果作为输入，最终输出分类结果。这篇工作的亮点在于首先树立了虚假新闻检测可解释性的研究方向，并提供了对应的模型。其次采用了层级注意力机制和共同注意力机制捕捉单词和句子间关系、新闻内容和评论间的关系。此方法最终的结果也超过了当时最先进的几个方法，并且解释性评估实验也体现了方法的先进之处，为后续研究贡献了很大的研究启发。

研究工作的另一个分支是使用CNN进行虚假新闻检测[35-38]。Yu等人[35]将一个事件的相关微博帖子按时间顺序分成若干组，每组通过段落向量方法[64]生成一个向量表示，之后所有向量表示形成一个矩阵作为网络的输入，CNN自动提取局部-全局的特征并学习潜在特征的高级交互。Karimi等人[36]提出了一种多源、多类别的检测模型，该模型结合了不同的源，以提高对不同程度虚假（包括真实、大部分真实、半真实、很不真实）的辨别能力。在这个过程中，文本内容用CNN提取特征，不同来源的信息被分别提取并融合在一起。Qian等人[37]利用用户对新闻反馈的历史数据进行虚假新闻检测。整个网络由一个用户响应生成器（User Response Generator，URG）和一个两级卷积神经网络

(Two-Level Convolutional Neural Networks，TCNN)组成，URG的目标是根据用户对真假新闻的反馈历史来学习用户对真假新闻的反应的生成模式，TCNN利用CNN来学习新闻在单词和句子级别的特征，最后将这两个模块融合以执行分类。Popat等人[38]提出从外部来源检索相关文章以提高预测能力，文章中相关的信息可以通过注意力机制得到。

5　性能论述

表3呈现了几个近年来常用数据集上搜集到的最先进的方法及性能。从泛用性上来讲，由于包含数据数量、题材类型、形式，近年来最为常用的，发表论文最多的数据集是BuzzFeedNews、LIAR、FNC-1和FakeNewsNet。还可以观察到，在微博数据集上的最先进方法RDM[65]已经在各个指标上达到了95%以上，但其他的数据集上的方法最高也仅刚超过90%，说明虚假新闻检测技术还有较大提升空间。在这些列出的先进的方法中，RDM利用强化学习模型寻找检测节点，以期望不必输入所有内容信息即可实现早期检测；GLAN[46]联合编码了源微博、回复信息和用户信息，构建异质图以实现虚假新闻检测，在两个Twitter数据集上取得了最好的效果；ED[66]从词汇层面、句法层面、语义层面和全文层面考察新闻内容，并依靠社会心理学和法医心理学的成熟理论，以实现仅针对于新闻内容的可解释虚假新闻检测；RoBERTa[67]评估了数据集偏差的问题，并探索了一系列基于预训练技术的虚假新闻检测语言模型，以及传统的和深度学习的模型，并首次从不同方面比较了它们的性能；USEF[52]将在虚假新闻挑战赛阶段一(FNC-1)中提出的立场检测任务(Stance Detection)与文本蕴含任务(Textual Entailment)联系起来，提出了结合统计学习和深度学习的模型，是现在FNC-1上最先进的模型；SAFE[68]关注于多模态形式的新闻，对文本和视觉信息提取特征并学习它们之间的相似性关联，最后根据文本、图像的特征或不匹配程度来判定新闻的真伪。

表 3　常用数据集当前最先进方法及指标

Tab. 3　State-of-the-art methods with performances on common datasets

数据集	模型名	精确率	召回率	F1值	准确率
微博[28]	RDM[65]	0.950	0.963	0.957	0.957
Twitter15[6]	GLAN[46]	—	—	0.906	0.905
Twitter16[28]	GLAN[46]	—	—	0.901	0.902
BuzzFeedNews[42]	ED[66]	0.857	0.902	0.879	0.879
LIAR[43]	RoBERTa[67]	0.63	0.62	0.62	0.62
FNC-1[30]	USEF[52]	—	—	0.636	—
FakeNewsNet[44]	SAFE[68]	0.873	0.92	0.896	0.856

6　展望

尽管近年来虚假新闻检测技术取得了巨大进展，但仍有一些方向有潜力进一步得到改善。

(1)早期检测：为了最大限度地减少虚假新闻的负面影响，尽早发现虚假新闻至关重要。尽管Chen等人[18]已经探索了早期检测，但性能仍需改进。在未来的研究中，早期检测可以基于历史信息聚焦于热点事件。

(2)恶意用户检测：虚假新闻藉由用户传播，因此只要事先对恶意用户进行标记，就可以切断虚假新闻的传播链，减少虚假新闻的影响。利用历史数据可以实现对恶意用户的检测，从而进一步提高虚假新闻的检测能力。

(3)多模态检测：多年以来，各个平台信息中出现了越来越多的多模态数据，其中视觉内容甚至远远多于文本内容。因此，整合视觉内容将成为虚假新闻检测的主导趋势。尽管有一些工作[25]对此领域进行了探索，但性能远远不能令人满意，因为对图像特征的处理只是简单地融合，而且视频数据还从未被利用。

7　结论

对虚假新闻的检测一直以来都是研究的热点问题。从早期手工提取特征到现在提取的有针对性的多样化的特征，从早期传统机器学习分类到深度神经网络，乃至从早期单一对象的检测到现在包括新闻内容特征、传播路径、用户信息等多方面联合检测，虚假新闻检测技术得到了深远的发展。

本文重点对虚假新闻检测技术从传播意图相关的特征的角度进行了深入的介绍。首先，本文给出了虚假新闻检测的问题定义和基本范式，讨论了基准数据集和最常用的评估指标。本文揭示了虚假新闻的一个本质特征，即与报道真实事件的真实新闻不同，假新闻通常是有意为之，有特定的传播

意图。虚假新闻的传播意图通常可分为误导公众、操纵舆论和吸引注意三类,本文将检测需要提取的特征与传播意图关联起来,对相关的方法作了对应分类和介绍。之后在提取特征的基础上介绍了包括传统机器学习和近年深度神经网络的虚假新闻检测方法,并对现有方法的性能作了简要的展示。在此基础上,本文最后提出了未来假新闻检测的几个方向。本文提供了一个全新的视角,可以指导研究者更好地理解这一领域。

参考文献

[1] TAKAYASU M, SATO K, SANO Y, et al. Rumor diffusion and convergence during the 3.11 earthquake: A twitter case study[J]. PLoS One, 2015, 10(4): e0121443.

[2] GUPTA A, LAMBA H, KUMARAGURU P, et al. Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy[C]//Proceedings of the 22nd International Conference on World Wide Web-WWW'13 Companion. Rio de Janeiro, Brazil. New York: ACM Press, 2013: 729-736.

[3] BALMAS M. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism[J]. Communication Research, 2014, 41(3): 430-454.

[4] JIN Zhiwei, CAO Juan, GUO Han, et al. Detection and analysis of 2016 US presidential election related rumors on twitter[M]//Social, Cultural, and Behavioral Modeling. Cham: Springer International Publishing, 2017: 14-24.

[5] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C]//Proceedings of the 20th International Conference on World Wide Web-WWW '11. Hyderabad, India. New York: ACM Press, 2011: 675-684.

[6] LIU Xiaomo, NOURBAKHSH A, LI Quanzhi, et al. Real-time rumor debunking on twitter[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne Australia. New York, NY, USA: ACM, 2015: 1867-1870.

[7] GUPTA M, ZHAO Peixiang, HAN Jiawei. Evaluating event credibility on twitter[C]//Proceedings of the 2012 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2012: 153-164.

[8] BIYANI P, TSIOUTSIOULIKLIS K, BLACKMER J. "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality[C]//AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016: 94-100.

[9] SUN Shengyun, LIU Hongyan, HE Jun, et al. Detecting event rumors on sina weibo automatically[M]//Web Technologies and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 120-131.

[10] 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-MDS'12. Beijing, China. New York: ACM Press, 2012: 1-7.

[11] JIN Zhiwei, CAO Juan, JIANG Yugang, et al. News credibility evaluation on microblog with a hierarchical propagation model[C]//2014 IEEE International Conference on Data Mining. Shenzhen, China. IEEE, 2014: 230-239.

[12] MA Jing, GAO Wei, WEI Zhongyu, et al. Detect rumors using time series of social context information on microblogging websites[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne Australia. New York, NY, USA: ACM, 2015: 1751-1754.

[13] WU Ke, YANG Song, ZHU K Q. False rumors detection on Sina Weibo by propagation structures[C]//2015 IEEE 31st International Conference on Data Engineering. Seoul. IEEE, 2015: 651-662.

[14] KWON S, CHA M, JUNG K, et al. Prominent features of rumor propagation in online social media[C]//2013 IEEE 13th International Conference on Data Mining. Dallas, TX, USA. IEEE, 2013: 1103-1108.

[15] PÉREZ-ROSAS V, KLEINBERG B, LEFEVRE A, et al. Automatic detection of fake news[J]. arXiv preprint arXiv:1708.07104, 2017.

[16] JIN Z, CAO J, ZHANG Y, et al. MCG-ICT at MediaEval 2015: Verifying multimedia use with a two-level classification model[C]//MediaEval, 2015.

[17] HASSAN N, LI Chengkai, TREMAYNE M. Detecting check-worthy factual claims in presidential debates[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne Australia. New York, NY, USA: ACM, 2015: 1835-1838.

[18] CHEN Tong, LI Xue, YIN Hongzhi, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[C]//Trends and Applica-

tions in Knowledge Discovery and Data Mining, 2018:. DOI:10.1007/978-3-030-04503-6_4.

[19] RASHKIN H, CHOI E, JANG J Y, et al. Truth of varying shades: Analyzing language in fake news and political fact-checking[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 2931-2937.

[20] WANG Xuezhi, YU Cong, BAUMGARTNER S, et al. Relevant document discovery for fact-checking articles[C]// Companion of the The Web Conference 2018 on The Web Conference 2018-WWW '18. Lyon, France. New York: ACM Press, 2018: 525-533.

[21] SHI Baoxu, WENINGER T. Fact checking in heterogeneous information networks[C]//Proceedings of the 25th International Conference Companion on World Wide Web-WWW '16 Companion. Montréal, Québec, Canada. New York: ACM Press, 2016: 101-102.

[22] WU You, AGARWAL P K, LI Chengkai, et al. Toward computational fact-checking[J]. Proceedings of the VLDB Endowment, 2014, 7(7): 589-600.

[23] YANG Shuo, SHU Kai, WANG Suhang, et al. Unsupervised fake news detection on social media: A generative approach[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 5644-5651.

[24] HOSSEINIMOTLAGH S, PAPALEXAKIS E E. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles[C]//Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2). 2018.

[25] JIN Zhiwei, CAO Juan, GUO Han, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM international conference on Multimedia. Mountain View California USA. New York, NY, USA: ACM, 2017: 795-816.

[26] WANG Yaqing, MA Fenglong, JIN Zhiwei, et al. EANN: event adversarial neural networks for multi-modal fake news detection[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London United Kingdom. New York, NY, USA: ACM, 2018: 849-857.

[27] JIN Zhiwei, CAO Juan, LUO Jiebo, et al. Image credibility analysis with effective domain transferred deep networks[J]. arXiv preprint arXiv:1611.05328, 2016.

[28] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]//IJCAI, 2016: 3818-3824.

[29] VOLKOVA S, SHAFFER K, JANG J Y, et al. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 647-653.

[30] HANSELOWSKI A, PVS A, SCHILLER B, et al. A retrospective analysis of the fake news challenge stance detection task[J]. arXiv preprint arXiv:1806.05180, 2018.

[31] RUCHANSKY N, SEO S, LIU Yan. CSI: A hybrid deep model for fake news detection[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore Singapore. New York, NY, USA: ACM, 2017: 797-806.

[32] WEN Weiming, SU Songwen, YU Zhou. Cross-lingual cross-platform rumor verification pivoting on multimedia content[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018.

[33] SHU Kai, CUI Limeng, WANG Suhang, et al. dEFEND: explainable fake news detection[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK USA. New York, NY, USA: ACM, 2019: 395-405.

[34] DE Sarkar S, YANG F, MUKHERJEE A. Attending sentences to detect satirical fake news[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 3371-3380.

[35] YU F, LIU Q, WU S, et al. A convolutional approach for misinformation identification[C]//IJCAI, 2017: 3901-3907.

[36] KARIMI H, ROY P, SABA-SADIYA S, et al. Multi-source multi-class fake news detection[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 1546-1557.

[37] QIAN Feng, GONG Chengyue, SHARMA K, et al. Neural user response generator: Fake news detection with collective user intelligence[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden. California: International Joint Conferences on Artificial Intelligence Organization, 2018.

[38] POPAT K, MUKHERJEE S, YATES A, et al. DeClarE:

debunking fake news and false claims using evidence-aware deep learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018.

[39] ZHOU X, ZAFARANI R. Fake news: A survey of research, detection methods, and opportunities[J]. arXiv preprint arXiv:1812.00315, 2018, 2.

[40] SHU Kai, SLIVA A, WANG Suhang, et al. Fake news detection on social media[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36.

[41] ZUBIAGA A, AKER A, BONTCHEVA K, et al. Detection and resolution of rumours in social media[J]. ACM Computing Surveys, 2018, 51(2): 1-36.

[42] CHU Zi, GIANVECCHIO S, WANG Haining, et al. Detecting automation of twitter accounts: Are you a human, bot, or cyborg?[J]. IEEE Transactions on Dependable and Secure Computing, 2012, 9(6): 811-824.

[43] WANG W Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017.

[44] SHU Kai, MAHUDESWARAN D, WANG Suhang, et al. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media[J]. Big Data, 2020, 8(3): 171-188.

[45] BOIDIDOU C, ANDREADOU K, PAPADOPOULOS S, et al. Verifying multimedia use at mediaEval 2015[J]. MediaEval, 2015, 3(3): 7.

[46] YUAN C, MA Q, ZHOU W, et al. Jointly embedding the local and global relations of heterogeneous graph for rumor detection[C]//2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019: 796-805.

[47] MA Ben, LIN Dazhen, CAO Donglin. Content representation for microblog rumor detection[C]//Advances in Computational Intelligence Systems, 2017. DOI: 10.1007/978-3-319-46562-3_16.

[48] HASSAN A, QAZVINIAN V, RADEV D. What's with the attitude?: Identifying sentences with attitude in online discussions[C]//EMNLP '10: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010: 1245-1255.

[49] RUBIN V, CONROY N, CHEN Yimin, et al. Fake news or truth? using satirical cues to detect potentially misleading news[C]//Proceedings of the Second Workshop on Computational Approaches to Deception Detection. San Diego, California. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 7-17.

[50] CUI Limeng, WANG Suhang, LEE D. SAME: sentiment-aware multi-modal embedding for detecting fake news[C]//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Vancouver British Columbia Canada. New York, NY, USA: ACM, 2019: 41-48.

[51] BOIDIDOU C, MIDDLETON S E, JIN Zhiwei, et al. Verifying information with multimedia content on twitter[J]. Multimedia Tools and Applications, 2018, 77(12): 15545-15571.

[52] SAIKH T, ANAND A, EKBAL A, et al. A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features[M]. Indian Institute of Technology Patna, Bihta, India; Indian Institute of Information Technology Kalyani, Kalyani, India, 2019.

[53] KOPPEL M, SCHLER J, BONCHEK-DOKOW E. Measuring Differentiability: Unmasking Pseudonymous Authors[J]. Journal of Machine Learning Research, 2007, 8(6).

[54] POTTHAST M, KIESEL J, REINARTZ K, et al. A stylometric inquiry into hyperpartisan and fake news[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018.

[55] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[56] JIN Zhiwei, CAO Juan, ZHANG Yongdong, et al. Novel visual and statistical image features for microblogs news verification[J]. IEEE Transactions on Multimedia, 2017, 19(3): 598-608.

[57] BLOM J N, HANSEN K R. Click bait: Forward-reference as lure in online news headlines[J]. Journal of Pragmatics, 2015, 76: 87-100.

[58] OLIVA A, TORRALBA A. Modeling the shape of the scene: A holistic representation of the spatial envelope[J]. International Journal of Computer Vision, 2001, 42(3): 145-175.

[59] QI Peng, CAO Juan, YANG Tianyun, et al. Exploiting multi-domain visual information for fake news detection[C]//

2019 IEEE International Conference on Data Mining (ICDM). Beijing, China. IEEE, 2019: 518-527.

[60] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[61] CIAMPAGLIA G L, SHIRALKAR P, ROCHA L M, et al. Computational fact checking from knowledge networks[J]. PLoS One, 2015, 10(6): e0128193. DOI:10.1371/journal.pone.0128193.

[62] WANG Biao, CHEN Ge, FU Luoyi, et al. DRIMUX: dynamic rumor influence minimization with user experience in social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(10): 2168-2181.

[63] SHU K, CUI L, WANG S, et al. dEFEND: Explainable Fake News Detection[C]//KDD. 2019.

[64] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]//International Conference on Machine Learning. PMLR, 2014: 1188-1196.

[65] ZHOU K, SHU C, LI B, et al. Early rumour detection[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 1614-1623.

[66] ZHOU X, JAIN A, PHO HA V V, et al. Fake news early detection: An interdisciplinary study[J]. arXiv preprint arXiv:1904.11679, 2019.

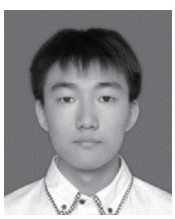
[67] KHAN J Y, KHONDAKER M, AFROZ S, et al. A benchmark study of machine learning models for online fake news detection[J]. Machine Learning with Applications, 2021, 4: 100032. 2021.

[68] ZHOU X, WU J, ZAFARANI R. Safe: similarity-aware multi-modal fake news detection (2020)[J]. Preprint. arXiv, 2020, 200304981.

作者简介

毛震东　男，1984年生，湖南岳阳人。中国科学技术大学特任研究员。主要研究方向为跨模态内容理解、计算机视觉、自然语言处理。

E-mail：zdmao@ustc.edu.cn



赵博文　男，1995年生，辽宁沈阳人。中国科学技术大学网络空间安全学院研究生。主要研究方向为跨模态内容理解。

E-mail：zbw0630@mail.ustc.edu.cn

白嘉萌　女，1995年生，陕西宝鸡人。中国科学院硕士研究生，主要研究方向为数据挖掘、推荐系统。

E-mail：baijiameng@iie.ac.cn

胡　博　男，1983年生，安徽阜阳人。中国科学技术大学信息学院副研究员。主要研究方向为社交网络和计算传播、推荐系统等。

E-mail：hubo@ustc.edu.cn