

面向虚假新闻检测的社交媒体多模态数据集构建

高国鹏¹, 房耀东¹, 韩彦芳¹, 钱振兴², 秦川¹

(1. 上海理工大学光电信息与计算机工程学院, 上海 200093; 2. 复旦大学计算机科学技术学院, 上海 200433)

摘要: 社交媒体的出现正在改变着人们的生活, 通过社交媒体可以便捷地获取和分享新闻, 但同时助力了虚假新闻的滋生和传播, 从而严重影响社会安全和稳定。因此, 虚假新闻检测引起了研究者广泛关注。尽管存在多种基于深度学习的解决方案, 但这些方法需要大量的数据作为支撑。现有的虚假新闻数据集, 尤其是中文数据集不仅稀缺, 而且数据集中的新闻大多属于同一个类别。为了更好地检测虚假新闻, 构建了一个新的多模态的虚假新闻数据集 (MFND, multi-modal fake news dataset), 其中包含政治、经济、娱乐、体育、国际、科技、军事、教育、健康和社会生活这 10 个类别的中文和英文新闻数据。对提出的虚假新闻数据集的词频和类别进行分析, 并与现有的虚假新闻数据集在新闻数量、新闻类别、模态信息和新闻语种等方面进行了对比, 结果显示 MFND 在类别信息和新闻语种方面表现突出。另外, 利用现有的典型虚假新闻检测方法在 MFND 上进行训练和验证, 实验结果表明, 相较于现有主流的虚假新闻数据集, MFND 可以为模型提供 10% 左右的性能提升。

关键词: 社交媒体; 虚假新闻检测; 多模态; 多类别; 数据集

中图分类号: TP393

文献标识码: A

DOI: 10.11959/j.issn.2096-109x.2023060

Construction of multi-modal social media dataset for fake news detection

GAO Guopeng¹, FANG Yaodong¹, HAN Yanfang¹, QIAN Zhenxing², QIN Chuan¹

1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

2. School of Computer Science, Fudan University, Shanghai 200433, China

Abstract: The advent of social media has brought about significant changes in people's lives. While social media allows for easy access and sharing of news, it has also become a breeding ground for the dissemination of fake news, posing a serious threat to social security and stability. Consequently, researchers have shifted their focus towards

收稿日期: 2022-10-17; 修回日期: 2023-05-26

通信作者: 秦川, qin@usst.edu.cn

基金项目: 国家自然科学基金 (U20B2051, 62172280); 上海市自然科学基金 (21ZR1444600)

Foundation Items: The National Natural Science Foundation of China (U20B2051, 62172280), The Natural Science Foundation of Shanghai (21ZR1444600)

引用格式: 高国鹏, 房耀东, 韩彦芳, 等. 面向虚假新闻检测的社交媒体多模态数据集构建[J]. 网络与信息安全学报, 2023, 9(4): 144-154.

Citation Format: GAO G P, FANG Y D, HAN Y F, et al. Construction of multi-modal social media dataset for fake news detection[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 144-154.

fake news detection. Although several deep learning-based solutions have been proposed, these methods heavily rely on large amounts of supporting data. Currently, there is a scarcity of existing datasets, particularly in Chinese, and the collected news articles are often limited to the same category. To enhance the detection of fake news, a new multi-modal fake news dataset (MFND) was developed, which comprised Chinese and English news data from ten diverse categories: politics, economy, entertainment, sports, international affairs, technology, military, education, health, and social life. The word frequencies and categories of the proposed fake news dataset were analyzed and compared with existing fake news datasets in terms of number of news, news categories, modal information and news languages. The results of the comparison demonstrate that the MFND dataset excels in terms of category information and news languages. Moreover, training and validating existing typical fake news detection methods with MFND dataset, the experimental results show an improvement of approximately 10% in model performance compared to existing mainstream fake news datasets.

Keywords: social media, fake news detection, multi-modal, multi-category, dataset

0 引言

近年来,随着互联网的快速发展以及微博、推特等社交媒体的不断兴起,人们获取新闻的方式发生着巨大变化,越来越多的人开始通过社交媒体来阅读新闻。相较于传统报刊上的新闻,社交媒体上的新闻可以提供更加丰富的视听信息,同时凭借实时性高、成本低等优势,社交媒体逐渐成为新闻传播的主要途径。各大主流新闻媒体纷纷开设社交媒体账号,在社交媒体上进行新闻播报。值得注意的是,社交媒体缺乏完善的审核机制,任何媒体和个人都可以通过社交媒体进行“新闻”的制作和传播,这使得社交媒体上新闻的可靠性难以得到保证,社交媒体上呈现出真假新闻并存的局面,因此对于社交媒体上的虚假新闻检测成为研究的热点问题。

在虚假新闻的研究中,虚假新闻被赋予了不同的研究含义。一部分研究人员将人为的并且可以被证实为虚假的新闻当作虚假新闻来进行研究^[1];另一部分研究人员将讽刺性或者夸大性的新闻当作虚假新闻^[2]。为了更加明确地界定虚假新闻,Shu等^[3]指出虚假新闻是指故意制作、传播并且可以被证实为虚假的新闻文章或新闻报道。虚假新闻的传播有着严重的社会危害,严重影响着人们的生活和社会的稳定。美国皮尤研究中心(Pew Research Center)的调查报告指出,越来越多的美国用户通过社交媒体来阅读新闻。大约2/3的美国成年人认为,虚假新闻会对当前问题和事件的基本事实造成很大的误导。

美国麻省理工学院的一项研究^[4]表明,虚假新闻在社交媒体上的传播速度是真实新闻的6倍。为了最大限度地减小虚假新闻带来的有害影响,需要在新闻传播的前期进行假新闻的检测和辟谣。在早期,社交媒体还处于发展阶段,社交媒体上的新闻数据较少,可以依靠专家系统对新闻进行人为判定。随着互联网的快速发展,社交媒体上的新闻数量以指数级趋势增加,完全依靠人工进行虚假新闻的判定,无法及时、高效地从大量的新闻中筛选出虚假新闻。近年来,随着机器学习和深度学习的不断发展和完善,人工智能逐渐被应用到虚假新闻检测的研究中。新闻的文本内容传递着大量信息,针对新闻文本特征的提取出现在很多的研究中^[5-7]。Ma等^[8]考虑到新闻的时序性特点,利用循环神经网络(RNN)对动态范围内的新闻进行特征提取来实现虚假新闻检测。鉴于卷积神经网络(CNN)在图像领域的优异表现,Yu等^[9]使用CNN进行文字的语义特征提取,更好地学习到了其他场景下信息的特征表示。Bian等^[10]认为图神经网络可以学习到更高层次的特征,有利于虚假新闻的检测,为此提出了一种双向的图神经网络模型来进行虚假新闻检测。相较于纯文本的新闻,含有图片等多媒体数据的新闻更容易吸引人们的眼球,多模态新闻逐渐成为新闻的主要方式。多模态新闻有着丰富的视觉特征,可以为虚假新闻检测提供额外的特征表示,许多虚假新闻检测方法开始融合文本特征和图片特征^[11-13]。例如,Jin等^[14]使用LSTM和预训练的VGG网络进行文字和图片特征的提取

来开展虚假新闻检测任务; Tuan 等^[15]为了更好地融合多模态信息,使用基于 Transformer 的双向编码器(BERT)模型和预训练的 VGG 网络作为文本特征和图片特征提取器,并且将多个注意力机制层堆叠进行多模态特征的融合。

随着针对虚假新闻检测问题的不断深入研究,许多虚假新闻检测数据集被构建起来用于进行学习。这些数据集大部分是单模态的(纯文本的新闻),多模态的虚假新闻检测数据集相对较少。早期通过爬取推特、新浪微博等社交媒体网站构建出的多模态数据集中只保存了图片的链接,这些链接可能已经失效而无法获取到相关图片,这让数据集的完整性不能得到保证。目前现有的多模态数据集中,大多是与政治、经济相关的新闻,数据集中的新闻类别比较单一,并且部分多类别的数据集并未对类别信息进行标注。在生成对抗网络^[16]技术的不断发展和完善下,DeepFake 等深度伪造技术也在不断进步,多模态下的虚假新闻造假方式又更进一步。人们可以利用 DeepFake 技术生成与虚假新闻内容相关的图片,并且 DeepFake 技术使用的门槛逐渐降低。不同于传统的篡改,DeepFake 生成的篡改图片更加自然。

为了促进虚假新闻检测,尤其是多模态的虚假新闻检测的发展,本文基于社交媒体和事实核查网站构建了一个新的多模态的虚假新闻数据集:MFND(multi-modal fake news dataset)。MFND 由 10 000 条新闻组成,包含 5 000 条真实新闻、5 000 条虚假新闻以及与新闻相关的近 50 000 张图片。为了丰富新闻数据的类别信息,MFND 中收集了 10 个类别的新闻数据,各个类别的新闻数据量不均等,为了应对越来越复杂的虚假新闻形式,收集了一些由深度伪造技术生成的虚假新闻。同时,MFND 中不仅包含中文新闻,还收集了英文新闻,可以被用于多语态的虚假新闻检测。

1 相关数据集

虚假新闻的检测有着不同的策略和方法,研究人员根据研究所需构建相对应的虚假新闻数据集。总体来说,现有的虚假新闻检测数据集可以分为单模态数据集和多模态数据集两大类,表 1 展示了现有的单模态和多模态虚假新闻数据集,其中标签数量表示新闻是被二分类标记(真、假

两个标签)还是被多分类标记(真、大部分真、假等多个标签)的。

表 1 现有的单模态和多模态虚假新闻数据集
Table 1 Existing unimodal and multimodal fake news datasets

数据集分类	数据集名称	年份	标签数量
单模态	CREDBANK ^[17]	2015	2
	BuzzFeedNews ^[18]	2016	4
	LIAR ^[19]	2017	6
	NaistCovid ^[20]	2020	2
	FakeCovid ^[21]	2020	2
	CrossCOVID19 ^[22]	2021	2
多模态	MediaEval ^[23]	2016	2
	Weibo ^[14]	2017	2
	FacebookHoax ^[24]	2017	2
	TI-CNN ^[25]	2018	2
	Fakeddit ^[26]	2020	2、3、6
	MM-COVID ^[27]	2020	2

1.1 单模态数据集

单模态数据集由纯文本或者纯图片的新闻数据构成,主要为模型提供丰富的文字特征。为了对社交媒体上虚假新闻进行研究,2015 年,Mitra 等^[17]构建了一个大规模、无偏见的数据语料库 CREDBANK,主要包含新闻正文、话题、事件和相关专家的可信度判断。CREDBANK 中包含 6 000 万条推文,对应 1 000 多个真实的新闻事件,每个新闻事件都由多个专门的工作人员进行可信度标注,并且标注了判定的理由。Facebook 在 2016 年创建了 BuzzFeed News 虚假新闻数据集,该数据集收集了美国多个社交媒体账号在 Facebook 上发布的 2 282 个帖子,数据集中不仅包含正文内容,还包含新闻的分享数、评论数等社交上下文信息。对于数据的标注,Facebook 用人工对这些新闻进行多层次的新闻判定。Potthast 等^[18]对该数据集进行了进一步扩充和丰富,设置了 3 种形式的真假标签。数据集的大小在一定程度上影响着深度学习模型的检测准确度,Wang^[19]构建了一个大型的、公开的、纯文本的虚假新闻检测数据集 LIAR。LIAR 数据集中的数据主要来自事实核查网站 PolitiFact,这个网站中的新闻都是来自真实场景,包括推特上的帖子、电视辩论、

Facebook 帖子、新闻发布会和电视广告等。Wang 等从 PolitiFact 上收集了十年间的 12 836 条新闻数据。出于真实性考虑, 作者将 LIAR 的数据集设置了 6 个细粒度的标签, 分别为 pants-fire、false、barely-false、half-true、mostly-true 和 true; 同时考虑到数据的均衡性, 这 6 类标签的新闻数量基本均等。为了应对网络上所产生的谣言, 许多针对纯文本虚假新闻数据集被构建起来^[20-22]。

1.2 多模态数据集

多模态新闻是指新闻中含有除文字信息外其他模态的信息(如图片信息、视频信息等), 可以为虚假新闻检测模型提供额外的特征。Boididou 等^[23]在 2016 年提出了一个多模态的虚假新闻数据集, 数据集中包含 6 225 个真实的推特新闻和 9 596 个虚假的推特新闻, 同时包含与这些新闻相关的 193 个真实事件和 220 个被篡改或者滥用的图片与视频案例。Jin 等^[14]通过收集新浪微博平台上的数据构建了一个多模态的中文虚假新闻数据集, 数据集中包含约 40 000 条多模态新闻数据。该数据集中的虚假新闻来自微博官方平台对 2012—2016 年所验证过的虚假新闻帖子, 真实新闻主要来自中国主流媒体的新闻播报。微博数据集不仅提供了文字和图片信息, 还提供了新闻的社交背景信息。许多多模态虚假新闻数据集规模较小^[24-25], 为了提供一个有效而全面的数据集, Nakamura 等^[26]在 2020 年提出了 Fakeddit 数据集, 收集了 100 多万条来自多个类别的新闻样本, 并对新闻标注了 3 种形式的真伪性标签。针对社交媒体上广泛的谣言, Li 等^[27]于 2020 年提出了一个多模态、多语种、多维度的虚假新闻数据集, 数据集包含 6 种语言的 3 981 条虚假新闻和 7 000 多条真实新闻。

2 MFND

本节主要介绍面向社交媒体的虚假新闻检测数据集 MFND 构建流程(如图 1 所示), 包括新闻数据的收集、处理与存储以及对数据集的相关分析。

2.1 数据的收集

社交媒体是人们用来进行创作和分享的网络平台; 事实核查是对存疑的新闻信息进行核验, 并撰写核验报告, 而事实核查网站是用来展示需要被核验或者已经被核验过的新闻。这些平台和网站上有大量的新闻数据可以被用来进行研究。本文收集了新浪微博、推特等社交媒体和 Snopes 等事实核查网站上的新闻数据来构建 MFND。对于社交媒体网站, 收集了 2019 年 1 月至 2021 年 12 月这 3 年间的新闻数据; 对于事实核查网站, 收集 2021 年、2022 年这两年的新闻资讯。

社交媒体基本开放了可使用的数据接口(API), 利用这些 API 可以正确合法地获取新闻数据。社交媒体网络和事实核查网站设置了不同的约束条件进行数据收集: 社交媒体上的新闻, 按照时间和新闻类别或者关键词进行新闻的采集; 事实核查网站上的新闻, 按照新闻类别进行新闻样本的收集。社交媒体和事实核查网站上的新闻主要收集新闻样本中的标题、来源、正文以及新闻中可能存在的图片数据。由于根据新闻的链接进行新闻数据的存储, 部分链接会存在资源不可获取的情况, 为了数据收集的正常进行、方便后续处理, 统一将异常内容保存为“异常 404”。随着国家以及平台对谣言的整治, 虚假新闻尤其是中文的虚假新闻被辟谣后平台上的新闻会被撤销, 导致可获取的虚假新闻较少, 为了平衡真实

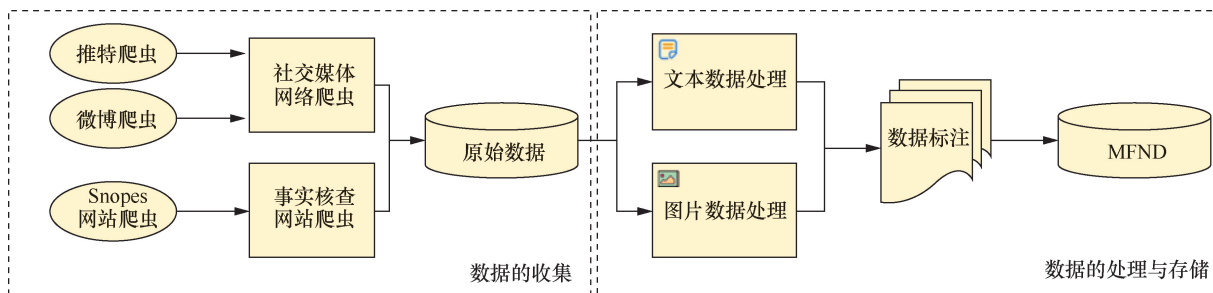


图1 MFND 构建流程

Figure 1 The flowchart of construction for MFND

新闻和虚假新闻,对公开的多模态数据集中的小部分新闻进行翻译或者抽取。

2.2 数据的处理与存储

(1) 文本数据的处理

初始收集到的新闻样本中包含大约两万条的新闻数据,这些原始数据包含收集到的异常数据,还有部分无法下载新闻所对应的图片,需要对这些原始数据进行进一步的处理。首先下载新闻中的图片链接,并将图片链接无法下载的对应新闻标记为异常数据。然后对数据进行过滤,剔除数据集中新闻内容为“异常 404”以及被标记为异常的新闻数据。表情符号因其简洁明了、形象有趣等特点,经常出现在社交媒体的推文中。对于这类特殊的语言符号,中文新闻下的表情符号使用“【表情】”这一字符串来进行文本替换,英文新闻下的表情符号使用“[emoji]”字符串来进行文本替换。

(2) 图片数据的处理

社交媒体会对用户上传的图像生成一组唯一的字符串作为标记。本文在对图像数据进行下载时,使用这组字符串作为图像的名称保存到本地,这会导致数据集中出现许多重复的图像。本文参考 Jin 等^[14]构建微博数据集的方法,使用基于局部感知哈希(LSH)的近似复制图像检测算法^[28]对数据集中重复的图像进行删除。虚假新闻中的图像大部分清晰度较差,本文使用百度智能云提供的图像清晰度增强模型,去除图像中的噪点、优化图像的纹理,实现图像清晰度的提升。对于数据集中尺寸较小的图像,使用百度智能云的图像无损放大方法,在不改变图像质量的前提下,对图像的长宽进行放大。

(3) 真实新闻和虚假新闻的标注

新华网、央视网、人民日报等权威媒体机构在发布新闻之前都会对内容事先核实,它们发布的新闻具有很高的可信度。因此将社交媒体收集到的新闻数据按照新闻来源进行数据抽取,对于来源是权威媒体或者官方媒体的新闻,将它们标记为真实新闻。社交媒体本身就带有简单的新闻检测功能,会将被举报为虚假的新闻进行核实和标注,对于这部分数据,将它们标注为虚假新闻。事实核查网站上的新闻经过了人工的核查,新闻的真伪都被进行了标注。网站中主要包含新闻的

标题、新闻的正文、新闻中可能存在的图片、新闻判定的阐释、新闻的真伪性和新闻报道的原始内容。事实核查网站中不是所有新闻都会被判定为真实新闻或者虚假新闻,部分新闻可能会被标记为“极有可能真实”等标签。在对事实核查网站收集到的新闻数据进行标签过滤时,只保存真实和虚假这两个标签的新闻。由于在数据收集时,是按照新闻类别来获取社交媒体和事实核查网站上的新闻的,所以按照收集的类别信息对真假新闻进行类别信息的标注;对于类别缺失的部分,使用文本分类模型对该新闻进行分类。

(4) 数据的存储

MFND 中的每条新闻数据包含了新闻的标题、新闻的来源、新闻的正文、图片的名称、新闻的类别和新闻的真伪标签这 6 部分的内容。新闻的来源字段说明了新闻的发布方,如人民日报;对于个人用户发布的新闻,使用用户的昵称作为新闻的来源。图片的名称字段中,只保存图片名不包含图片的存储路径,如果新闻中出现了多张图片,使用分号对图片名称进行拼接;如果新闻中不存在图片,将该字段设置为空。新闻的类别字段说明了新闻所属的类别,MFND 中收集了娱乐、经济、政治、国际、科技、军事、社会生活、教育、健康、体育这 10 个类别的新闻,其中英文新闻只包含娱乐、经济、政治、科技、军事、社会生活、教育、健康、体育这 9 个类别。新闻的真伪性字段只有两个标签——真实、虚假。MFND 中不仅包含了多模态的新闻数据(文本+图片),还包含了单模态的新闻数据(纯文本和纯图片)。对于纯文本或者纯图片的新闻数据,新闻的正文、类别、真伪 3 个字段一定不为空。

2.3 数据集的分析

表 2 展示了 MFND 的基本构成,它由 10 000 条中文新闻和英文新闻组成,其中中文新闻的数量为 8 000 条,英文新闻的数量为 2 000 条,真实新闻和虚假新闻所占的比例为 1:1。MFND 是一个多模态、多类别的虚假检测数据集,包含了政治、经济、娱乐、国际、科技、军事、社会生活、教育、健康和体育这 10 个类别下的新闻。从图 2 可以看出,MFND 中每个类别的新闻数量(纯文

字新闻和多模态新闻)不一致,新闻数据主要集中在娱乐和社会生活这两个类别中。

表2 MFND 新闻数量分析
Table 2 Analysis of the number of news in the MFND

语言	真实新闻	虚假新闻
中文	4 000	4 000
英文	1 000	1 000

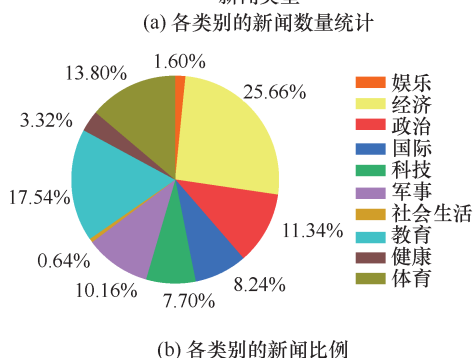
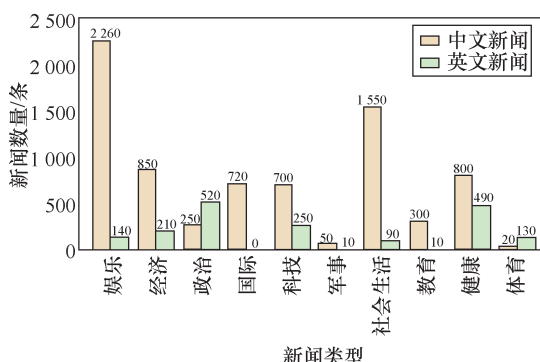


图2 MFND 中新闻类别的统计分析
Figure 2 Statistical analysis of news categories in the MFND

为了可视化数据集中真实新闻和虚假新闻的词频分布,对MFND中的所有中文新闻进行了词频分析。根据中文常用停用词表对新闻内容进行过滤,MFND中真实新闻和虚假新闻的10个高频词汇展示如下。

- 1) 虚假新闻: 孩子、中国、小女孩、帮忙、转发、朋友、扩散、附近、信息、大人。
- 2) 真实新闻: 中国、公司、发展、美国、电影、企业、市场、进行、生活、问题。
- 3) 整个新闻数据集: 中国、公司、发展、美国、孩子、电影、企业、市场、生活。

从中可以发现,真实新闻和虚假新闻在一定程度上具有相关性。

3 数据集对比验证

3.1 数据集对比

本文所构建的虚假新闻数据集的主要优势是新闻的类别信息丰富,数据集由娱乐、经济、政治、国际、科技、军事、社会生活、教育、健康和体育这10个常见类别的新闻组成,并且对每一条新闻的类别信息进行标注和保存。与其他大规模虚假新闻数据集相比,MFND的规模较小,同时没有包含新闻的社交上下文信息,社交上下文信息指的是新闻的用户评论、新闻的分享数、新闻的转发数等其他额外有效信息。这些问题是MFND的主要缺点,也给MFND的进一步改进提供了空间。

本文选取了现有的6个多模态数据集,分别是Weibo^[14]、MediaEval^[23]、FacebookHoax^[24]、TI-CNN^[25]、Fakeddit^[26]和MM-COVID^[27],与MFND进行量化比较,具体结果如表3所示。相较这6个多模态数据集,MFND在新闻的类别信息上有一定的优势。在这6个多模态数据集中,MM-COVID数据集提供了6种语言的新闻数据但是缺少中文新闻,其他5个多模态数据集只提供了单一语言的新闻数据,而MFND中包含了中文和英文两种语言的新闻。MediaEval、Weibo和Fakeddit这3个数据集的新闻数据是对多个类别下的新闻进行收集的,但是数据集中并没有进行相应的标注和说明。

3.2 数据集验证

为了有效检验MFND的性能,选择5种单模态虚假新闻检测和多模态虚假新闻检测方法,在MFND和Wang等^[29]开源的Weibo数据集(Weibo_Wang)上进行对比。

(1) 单模态虚假新闻检测方法

TextCNN^[30]: TextCNN使用简单的卷积神经网络来实现句子级别的语言分类,文本在输入之前需要使用预训练的语言模型进行分词操作。为了适用于中文虚假新闻分类,本文调整了模型的输入和输出,用TextCNN来进行中文虚假新闻的检测。

tanh-RNN^[8]: Ma等的这项工作使用循环神经网络对社交媒体上一段时间范围内的新闻进行文

表3 MFND 和现有的多模态虚假新闻数据集对比
Table 3 Comparison of our MFND with existing multi-modal fake news datasets

数据集	真实新闻	虚假新闻	模态信息	类别数量	社交上下文	语种
Weibo ^[14]	4 749	4 779	文本+图片	—	包含	1
MediaEval ^[23]	6 225	9 596	文本+图片	—	不包含	1
FacebookHoax ^[24]	6 577	8 923	文本+图片	1	包含	1
TI-CNN ^[25]	11 941	8 074	文本+图片	1	不包含	1
Fakeddit ^[26]	527 049	628 501	文本+图片	—	包含	1
MM-COVID ^[27]	7 192	3 981	文本+图片	1	包含	6
MFND	5 000	5 000	文本、图片、文本+图片	10	不包含	2

本特征的提取。由于 MFND 中没有保存新闻的时间信息，本文将每条新闻当作一个独立个体，忽略新闻的时间因素。RNN 具体的过程如式(1)和式(2)所示。

$$h_t = \tanh(Ux_t + Wh_{t-1} + b) \quad (1)$$

$$o_t = Vh_t + c \quad (2)$$

其中， U 、 W 、 V 分别表示输入层到隐藏层、隐藏层到隐藏层和隐藏层到输出层的权重矩阵； b 和 c 表示偏移量； $\tanh(\cdot)$ 表示非线性的激活函数； x_t 、 h_t 、 h_{t-1} 和 o_t 分别表示输入、 t 时刻的输出、上一时刻的输出和最终输出。

BERT^[31]：该模型因其强大的语言学习能力，成为当下最热的自然语言处理模型。Devlin 等^[31]指出使用预训练的 BERT 模型进行微调，可以搭建出效果优异的语言分类模型。本文参考这一建议，使用预训练的 BERT 模型进行微调，搭建了一个简单的文本二分类模型，将 BERT 作为主干网络来进行新闻文本特征的提取，然后加上一个全连接层来得到二分类任务的输出，处理过程如式(3)和式(4)所示。

$$H = \text{BERT}(\text{Embedding}(x)) \quad (3)$$

$$Y = f(W \times H + b) \quad (4)$$

其中， $\text{Embedding}(\cdot)$ 表示对输入进行预处理， H 表示 BERT 模型的输出， f 表示全连接层， W 表示全连接层中的权重矩阵， Y 表示模型的输出。

(2) 多模态虚假新闻检测方法

EANN^[29]：Wang 等的这项工作是基于事件类别的，主要包含 3 个模块：多模态特征提取模块、虚假新闻判断模块、事件判别模块。多模态特征提取模块使用 TextCNN 来获取文本特征，使用预

训练的 VGG-19 来获取图片特征。其提取出特征通过简单叠加给两个判断器进行判断。由于本次对比不做事件类别的判断，在对比实验中去除了事件判别模块，只保留了虚假新闻判别模块，如式(5)~式(7)所示。

$$R_T = \sigma(W_{\text{tf}} * R_{T_e}) \quad (5)$$

$$R_V = \sigma(W_{\text{vf}} * R_{V_{\text{vgg}}}) \quad (6)$$

$$R_F = R_T \oplus R_V \quad (7)$$

其中， R_{T_e} 和 $R_{V_{\text{vgg}}}$ 分别表示 TextCNN 模型处理后得到的文本特征表示和 VGG 处理后的图片特征表示； $\sigma(\cdot)$ 表示全连接层， W_{tf} 和 W_{vf} 表示全连接层中的参数权重； $\oplus$ 表示对特征向量进行拼接； R_T 、 R_V 和 R_F 表示文本特征、图片特征以及文本和图片融合特征。

SpokeFake^[32]：该模型将新闻中的文本内容和图片作为输入，用来进行虚假新闻的判断。它使用预训练的 VGG-19 网络进行图片特征的提取，使用预训练的 BERT 模型来提取新闻的文本特征，并对获得的多模态特征进行简单的融合。残差网络 (ResNet50) 可以更好地提取到图片的深层特征，本文修改了 SpokeFake 模型架构，使用 ResNet50 替换原先的图片特征提取的主干网络，式(8)和式(9)为数据处理过程。

$$T_f = \sigma(W \times \text{BERT}(x_M)) \quad (8)$$

$$V_f = \sigma(W \times \text{ResNet50}(x_G)) \quad (9)$$

其中， x_M 、 x_G 分别表示文本特征提取模块的输入、图片特征提取模块的输入； T_f 、 V_f 表示提取到的文本特征和图片特征； $\sigma(\cdot)$ 表示全连接层， W 表示全连接层中的参数权重。

(3) 虚假新闻检测评价指标

在评价指标方面,使用分类任务中常用的评价指标来评价各个方法在这两个数据集上的性能表现。这些指标具体定义如下。

准确率 (Accuracy) 表示模型预测正确的样本数量 (预测为真的正确样本和预测为假的正确样本数量之和) 占样本总量的百分数, 具体的定义如式(10)所示。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (10)$$

精确率 (Precision) 表示在预测为正样本的结果中预测正确的数量所占百分比, 式(11)为其具体定义。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (11)$$

召回率 (Recall) 表示预测为正样本的数量占原始正样本数量的百分比, 它的定义如式(12)所示。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (12)$$

F1 分数 (F1-score) 指标兼顾了精确率和召回率, 是两者的调和平均数, 式(13)展示了其具体定义。

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

其中, TP (True Positive) 表示样本本身为正样本预测结果也为正样本的样本数量; TN (True Negative) 表示样本预测结果为负样本但是实际上是正样本的样本数量; FP (False Positive) 表示样本被判定为正样本但事实上是负样本的样本数量; FN (False Negative) 表示样本被判定为负样本但是实际上是正样本的样本数量。这些分类任务中的指标, 值越高表示模型的性能越好。数据集方面, 按照八二法则进行数据集的划分, 数据集中 80% 的数据用来进行模型训练, 20% 的数据用作最后的性能评估。

(4) 实验参数设置

MFND 和 Weibo_Wang 数据集^[29]的对比主要关注中文新闻的检测性能。上面的对比方法中, 有些模型是针对英文虚假新闻检测设置的, 如在输入预处理部分, 分词模型是针对英文进行短语

分割提取, 这并不适用于中文新闻。为了让这些方法适用中文新闻的虚假新闻检测, 将这些对比方法的预处理部分统一更换成适用于中文的分词模型和停用词处理或词嵌入工具。这些对比方法的其他参数参考了相应的论文进行设置, 其中, 词向量大小表示文本被转化成向量后的维度大小; 优化器是一种优化策略, 用来管理和更新模型中可学习参数的值, 具体的参数设置如表 4 所示。所有的实验都是在 Ubuntu 20.04 系统和 NVIDIA 3080Ti 的显卡上进行训练和测试的, 主要使用的框架是 Pytorch 1.11 和 Tensorflow 2.8。

表 4 不同虚假新闻检测方法的参数设置
Table 4 Parameter settings for different fake news detection methods

检测方法	批量大小	优化器	词向量大小
tanh-RNN ^[8]	50	自适应学习	100
TextCNN ^[30]	50	随机梯度下降	768
BERT ^[31]	256	自适应矩估计	768
EANN ^[29]	100	自适应矩估计	32
SpokeFake ^[32]	256	自适应矩估计	768

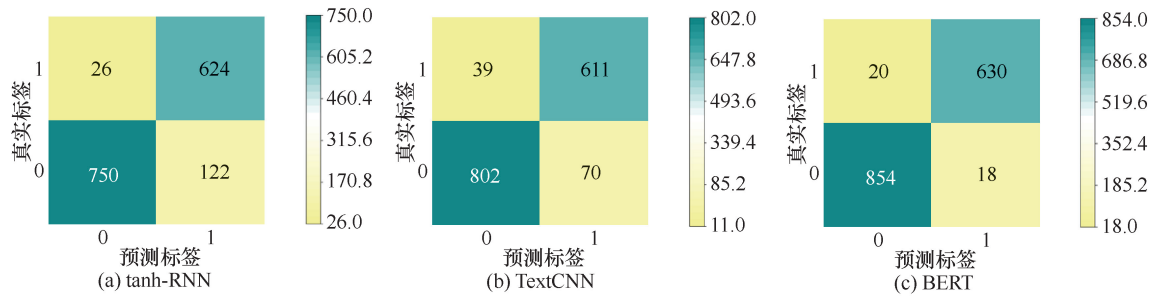
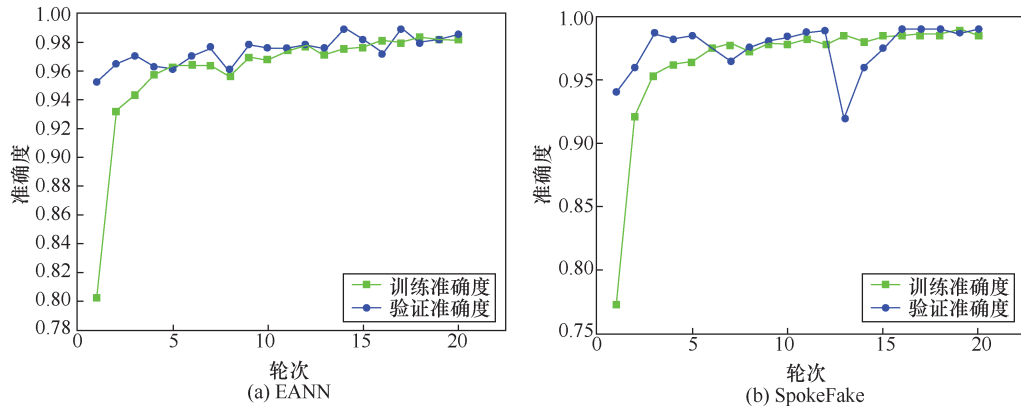
(5) 验证结果

表 5 展示了多种虚假新闻检测方法在不同数据集上的检测效果, 可以看出, 不管是单模态检测方法还是多模态检测方法, 相较 Weibo_Wang 数据集, 在 MFND 上都可以达到一个很好的检测效果。对于单模态的 3 种检测方法 tanh-RNN、TextCNN 和 BERT, 它们在 MFND 上可以达到 93% 左右的检测准确度, 其中 BERT 方法得到的检测效果最好。图 3(a)~图 3(c)分别展示了这 3 种单模态方法在 MFND 上的混淆矩阵, 图中真实标签和预测标签的 0 表示真实新闻、1 表示虚假新闻。从混淆矩阵图中可以看出, 总体来说, BERT 方法在真实新闻和虚假新闻的预测性能上优于 tanh-RNN 和 TextCNN 这两种方法。针对 tanh-RNN 和 TextCNN 这两种检测方法, TextCNN 在 MFND 上的预测性能更优, 但是 tanh-RNN 对于虚假类别的新闻检测效果比 TextCNN 好。图 4(a)、图 4(b)为 EANN 和 SpokeFake 这两种多模态方法在 MFND 上的训练和验证准确度, 由图可以看出, 在经历 15 轮次之后, 模型的训练和验证准确度趋于稳定。对于单

表5 多种虚假新闻检测方法在不同数据集上的检测效果

Table 5 The performance of multiple fake news detection methods on different datasets

方法		MFND				Weibo_Wang 数据集 ^[29]			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
单模态	tanh-RNN ^[8]	0.917	0.963	0.853	0.905	0.575	0.523	0.415	0.463
	TextCNN ^[30]	0.931	0.924	0.934	0.933	0.875	0.812	0.921	0.865
	BERT ^[31]	0.980	0.987	0.975	0.981	0.858	0.888	0.868	0.878
多模态	EANN ^[29]	0.982	0.969	0.995	0.982	0.803	0.866	0.731	0.793
	SpokeFake ^[32]	0.982	0.995	0.97	0.982	0.854	0.812	0.895	0.854

图3 单模态方法在 MFND 上的混淆矩阵
Figure 3 Confusion matrix for unimodal methods on the MFND图4 多模态方法在 MFND 上的训练和验证准确度
Figure 4 Training and validation accuracy of multimodal methods on MFND

模态方法而言, MFND 中每个新闻样本的长度相对较长,这使得可以为 RNN 类或者 CNN 类模型提供更好的文本特征。多模态的虚假新闻检测方法,因为其内容特征的丰富性,可以提供更多有用的信息用来进行判断,总的来说,多模态虚假新闻检测优于单模态虚假新闻检测。从表 5 可以看出, MFND 可以被用来对现有的中文虚假新闻数据集进行补充,从而得到更好的模型效果。由于该数据集中还包含部分英文新闻,使用 BERT 方法对 MFND 中的英文新闻进行检测,同时选取了英文虚假新闻数据集 LIAR^[18]中的部分数据进

行对比和测试。对于 BERT 方法的具体处理如下:批量大小为 32,分词模型的最大长度为 128。在 MFND 上, BERT 方法可以达到 0.67 的准确度而在 LIAR 数据集上只有 0.56 的准确度。相较中文新闻的检测, MFND 英文新闻检测的结果相对较差,可以作为其他数据集的补充。

4 结束语

本文研究并探讨了多模态虚假新闻检测以及多模态数据集相关的问题,构建了一个新的用于社交媒体的多模态虚假新闻数据集。本文所构

建的数据集是现有公开的数据集中为数不多的多语种、多类别的多模态虚假新闻数据集。

MFND 中由相同数量的真实新闻和虚假新闻组成,这些新闻来自 10 个新闻类别。通过对社交媒体网站和事实核查网站上的数据进行收集、筛选和标注,这些数据可以被用来进行细粒度的虚假新闻检测算法的开发。本文使用多个典型的虚假新闻检测方法在不同数据集上进行训练和测试,结果显示,经过 MFND 训练的模型检测性能得到提升。

随着新闻传播形式的不断丰富、造假手段的不断进步,虚假新闻呈现出不同的形式,需要对数据集进行不断迭代更新。数据集的规模很大程度上影响着检测模型的性能,本文构建的虚假新闻检测数据集在规模上有待扩充,同时数据集中新闻样本的正文长度跨度较大。不同的研究者研究虚假新闻的角度不同,所需要的新闻特征也不相同,而本文数据集的数据信息还不够全面。这些都为下一步研究和改进提供了空间和方向,在今后的研究中将不断对数据集及时进行更新和维护。

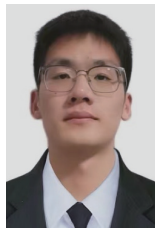
参考文献:

- [1] ALLCOTT H, GENTZKOW M. Social media and fake news in the 2016 election[J]. *Journal of Economic Perspectives*, 2017, 31(2): 211-36.
- [2] CONROY N K, RUBIN V L, CHEN Y. Automatic deception detection: methods for finding fake news[J]. *Proceedings of the Association for Information Science and Technology*, 2015, 52(1): 1-4.
- [3] SHU K, SLIVA A, WANG S H, et al. Fake news detection on social media: a data mining perspective[J]. *ACM SIGKDD Explorations Newsletter*, 2017, 19(1): 22-36.
- [4] VOSOUGHI S, ROY D, ARAL S. The spread of true and false news online[J]. *Science*, 2018, 359(6380): 1146-1151.
- [5] LI L, CAI G, CHEN N. A rumor events detection method based on deep bidirectional GRU neural network[C]//Proc of 2018 IEEE 3rd International Conference on Image, Vision and Computing. 2018: 755-759.
- [6] 郑洪浩, 郝一诺, 于洪涛, 等. 基于改进 Transformer 的社交媒体谣言检测 [J]. *网络与信息安全学报*, 2022, 8 (4): 168-174.
- [7] ZHEN H H, HAO Y N, YU H T, et al. Rumor detection in social media based on enhanced Transformer[J]. *Chinese Journal of Network and Information Security*, 2022, 8(4): 168-174.
- [7] CHENG M X, NAZARIAN S, BOGDAN P. Vroc: variational autoencoder-aided multi-task rumor classifier based on text[C]//Proc of the Web Conference 2020. 2020: 2892-2898.
- [8] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]//Proc of the 25th International Joint Conference on Artificial Intelligence. 2016: 3818-3824.
- [9] YU F, LIU Q, WU S, et al. A convolutional approach for misinformation identification[C]//Proc of the 26th International Joint Conference on Artificial Intelligence. 2017: 3901-3907.
- [10] BIAN T, XIAO X, XU T Y, et al. Rumor detection on social media with bi-directional graph convolutional networks[C]//Proc of the AAAI Conference on Artificial Intelligence. 2020: 549-556.
- [11] ZHANG H W, FANG Q, QIAN S S, et al. Multi-modal knowledge-aware event memory network for social media rumor detection[C]//Proc of the 27th ACM International Conference on Multimedia. 2019: 1942-1951.
- [12] AZRI A, FAVRE C, HARBI N, et al. Calling to CNN-LSTM for rumor detection: a deep multi-channel model for message veracity classification in microblogs[C]//Proc of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2021: 497-513.
- [13] 元鹏, 曹娟, 盛强. 语义增强的多模态虚假新闻检测[J]. *计算机研究与发展*, 2021, 58 (7): 1456-1465.
- [14] QI P, CAO J, SHENG Q. Semantics-enhanced multi-modal fake news detection[J]. *Journal of Computer Research and Development*, 2021, 58(7): 1456-1465.
- [14] JIN Z W, CAO J, GUO H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proc of the 25th ACM International Conference on Multimedia. 2017: 795-816.
- [15] TUAN N M D, MINH P Q N. Multimodal fusion with BERT and attention mechanism for fake news detection[C]//Proc of 2021 RIVF International Conference on Computing and Communication Technologies. 2021: 1-6.
- [16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proc of the 27th International Conference on Neural Information Processing Systems. 2014: 2672-268.
- [17] MITRA T, GILBERT E. Credbank: a large-scale social media corpus with associated credibility annotations[C]//Proc of the International AAAI Conference on Web and Social Media. 2015: 258-267.
- [18] POTTHAST M, KIESEL J, REINARTZ K, et al. A stylometric inquiry into hyperpartisan and fake news[J]. *arXiv Preprint arXiv: 1702.05638*, 2017.
- [19] WANG W Y. "Liar, liar pants on fire": a new benchmark dataset for fake news detection[J]. *arXiv Preprint arXiv: 1705.00648*, 2017.
- [20] GAO Z W, YADA S, WAKAMIYA S, et al. Naist covid: multilingual covid-19 Twitter and Weibo dataset[J]. *arXiv Preprint arXiv:2004.08145*, 2020.
- [21] SHAHI G K, NANDINI D. FakeCovid—a multilingual cross-domain fact check news dataset for COVID-19[J]. *arXiv Preprint arXiv:2006.11343*, 2020.
- [22] DU J S, DOU Y T, XIA C Y, et al. Cross-lingual covid-19 fake news detection[C]//Proc of 2021 International Conference on Data Mining Workshops. 2021: 859-862.
- [23] BOIDIDOU C, ANDREADOU K, PAPADOPOULOS S, et al. Verifying multimedia use at mediaeval 2015[C]//Proc of the MediaEval 2015 Multimedia Benchmark Workshop. 2015: 14-15.
- [24] TACCHINI E, BALLARIN G, DELLA VEDOVA M L, et al. Some like it hoax: automated fake news detection in social networks[J]. *arXiv Preprint arXiv: 1704.07506*, 2017.
- [25] YANG Y, ZHENG L, ZHANG J W, et al. TI-CNN: convolutional neural networks for fake news detection[J]. *arXiv Preprint arXiv:*

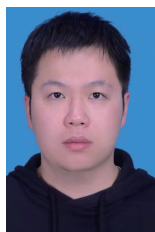
1806.00749, 2018.

- [26] NAKAMURA K, LEVY S, WANG W Y. Fakeddit: a new multi-modal benchmark dataset for fine-grained fake news detection[J]. arXiv Preprint arXiv: 1911.03854, 2019.
- [27] LI Y C, JIANG B H, SHU K, et al. MM-COVID: a multilingual and multimodal data repository for combating covid-19 disinformation[J]. arXiv Preprint arXiv: 2011.04088, 2020.
- [28] SLANEY M, CASEY M. Locality-sensitive hashing for finding nearest neighbors[J]. IEEE Signal Processing Magazine, 2008, 25(2): 128.
- [29] WANG Y Q, MA F L, JIN Z W, et al. Eann: event adversarial neural networks for multi-modal fake news detection[C]//Proc of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 849-857.
- [30] KIM Y. Convolutional neural networks for sentence classification[C]//Proc of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1746-1751.
- [31] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint arXiv: 1810.04805, 2018.
- [32] SINGHAL S, SHAH R R, CHAKRABORTY T, et al. SpotFake: a multi-modal framework for fake news detection[C]//Proc of 2019 IEEE 5th International Conference on Multimedia Big Data. 2019: 39-47.

[作者简介]



高国鹏（1998-），男，江苏淮安人，上海理工大学硕士生，主要研究方向为多媒体信息安全。



房耀东（1997-），男，江苏泰州人，上海理工大学硕士生，主要研究方向为多媒体信息安全。



韩彦芳（1974-），女，上海人，博士，上海理工大学讲师，主要研究方向为图像处理和模式识别。



钱振兴（1981-），男，江苏南通人，博士，复旦大学教授、博士生导师，主要研究方向为信息隐藏和 AI 安全。



秦川（1980-），男，安徽芜湖人，博士，上海理工大学教授、博士生导师，主要研究方向为多媒体信息安全和 AI 安全。