doi:10.3772/j.issn.1002-0470.2022.04.007

基于多模态深度融合的假消息检测①

景全亮② 范鑫鑫 王保利 毕经平 谭海宁

(中国科学院计算技术研究所 北京 100190)

(中国科学院大学 北京 100049)

摘 要 智能检测虚假信息是社交网络中需要解决的重要任务之一。本文旨在识别同时包含图像和文字的多模态虚假消息。目前,针对多模态的虚假消息检测已有一些成果,但现有模型通过直接拼接各模态特征方式实现多模态利用,忽略了图像和文件之间的关系,无法有效地学习消息中文字和图像的深度融合表示,导致该种类型的虚假消息检测方法表现不佳。本文提出基于预训练模型的多模态融合假消息检测方法,充分利用社交媒体中大量的含有多模态数据的消息,实现对假消息的有效检测,通过不同的训练任务加强模型融合多模态信息的能力,最终学习一个多模态信息的表示辅助假消息识别。在新浪微博真实数据集上的实验结果表明,本文提出的基于预训练的检测模型取得了比当前主流方法更优的效果,同时,本文采用的模型能够缓解训练集和测试集分布不均衡导致的检测准确率下降问题。

关键词 虚假信息检测;预训练;多模态融合;社交网络

0 引 言

近年来,随着社交媒体(新浪微博、Twitter、Facebook 等)应用的发展普及,获取信息的方式正在发生改变,人们花费在社交媒体上的时间越来越长[1],越来越多的人正在从社交媒体等渠道中获取信息,而不是从报纸、电视等传统、正规的渠道,例如,2016 年有 62% 的美国成年人在社交媒体上获得新闻,而在 2012 年该比例只占 49%。由于社交媒体等应用的开放性,每天都会有成千上万的消息在社交媒体中发表、传播,但是各机构并没有对各类信息进行有效的甄别,各类假消息层出不穷,对人们的生活造成了重大影响[2-3]。这已经成为各社交媒体、政府、社会面临的主要问题之一。

传统社交媒体的内容仅仅是文字信息,人们可以通过专家标注、分类方法[4-6]、图模型[7-9]等技术手段识别假消息。随着多媒体和计算机通信等技术的快速发展,社交媒体的内容越来越多样化,用户可以通过社交媒体发表文字、图片以及短视频信息,这吸引了越来越多人的关注,同时,由于人们可以随意对文本、图像、视频等多种信息进行伪造、拼接[10-11],这给假消息的检测带来了挑战。

本文的目标是检测同时包含了文本和图像的虚假消息。文本和图像提供了丰富的信息[10,12-13],为假消息的检测提供了各种技术途径。有些消息从文本特征即可判断真假[4,14-15],有些消息从图像内容即可识别真假[6-7],然而,有些消息需要使用图像和文本数据联合判断才能更加准确地判定是否为假消息[10,12,16]。

现阶段,基于传统的特征提取方法和基于深度学习的方法都已经被应用到假消息的检测任务中。文献[4]试图从消息的文本内容中提取特征进行假

① 国家重点研发计划(2017YFC0820700)和国家自然科学基金(61472403,62002343)资助项目。

② 男,1988 年生,博士生;研究方向:网络监测,假消息检测,大数据;联系人,E-mail:jingquanliang@ict.ac.cn。

(收稿日期:2021-03-12)

消息的检测,文献[17]利用人工提取的特征构建决策树模型实现假消息的识别。文献[11]利用引入注意力机制的循环神经网络(recurrent neural network,RNN)实现假消息的识别。在利用多类型数据方面,基于深度学习的方法能够提取更加相关的特征,取得了比传统方法更好的效果。文献[10]受自动编码器思想的启发,尝试通过学习文本和图像的共享表示形式,以此检测假消息。文献[12]通过基于注意力机制,利用视觉、文本和社交环境特征来预测假新闻。文献[16]使用一个额外的事件判别器来学习所有消息中所有事件之间共享的共同特征,基于此特征通过一个假消息的检测器判断消息的真假。

针对同时包含图像和文本的假消息检测,目前深度学习模型尚存在以下的缺陷或不足。首先,现有模型往往通过独立分支各自获取图像和文本特征,并将其拼接的方式实现各模态信息的利用,该种使用方式没有考虑文本和图像之间的关系,如文本和图像是否匹配等,从而降低了假消息检测的准确度,同时,现有的检测模型对于图像特征的提取比较粗糙,仅仅获取了整个图像的总体特征,没有对图像进行细粒度的处理,进一步影响检测准确性;其次,社交媒体中含有大量的图像和文本数据,该类数据包含的信息可以增强假消息识别的准确率,但是现有的方法仅仅基于标准的训练集,并没有充分利用社交媒体中的图像和文本数据,造成模型不能充分理解未包含训练集中特征的消息,导致对该类型假消息检测准确度低。

为了解决以上问题,亟需探索如何构建有效的模型融合文本和图像信息以便更加精确地识别假消息。本工作首先通过将文本和图像信息同时经由Transformer[18]模型处理和预训练,学习两者的融合表示;然后基于已标注数据集对预训练的模型进行参数调整,学习一个针对该任务的模型参数;最后通过该调整的模型识别假消息。

本文的主要贡献如下。

(1) 提出了一种融合社交媒体消息中文本和图像的模型,通过该模型可以有效学习文本和图像的融合表示。

(2) 所提融合模型充分利用了已有的海量社交媒体数据,提高了假消息识别的准确率,同时缓解了在数据分布不均衡时模型检测准确率下降过快的问题。

(3) 在真实数据集上进行了大量实验,实验结果表明,相较于当前主流方法,本文提出的假消息检测方法可以更有效地识别消息的真假。

本文剩余部分总结如下:第 1 节介绍了假消息检测相关工作,同时介绍了在多模态融合方面的研究进展;第 2 节介绍了本文模型所使用的大规模数据获取方法;第 3 节详细描述了本文提出的假消息检测框架和方法;第 4 节通过充分的实验对本研究中提出的方法进行了有效的验证,并分析实验结果;第 5 节总结了对本文的工作并展望未来发展方向和前景。

1 相关工作

本节将详细介绍目前主流的面向文本和图像的假消息检测相关工作。现阶段,假消息的检测方法主要可以分为两类,即基于单模态的方法和基于多模态的方法。

首先,在基于单模态的检测方法中,文献[4,14]基于文本的统计特征或者语义特征探索消息的可信性。文献[4]基于消息、用户、主题以及传播数据,构建决策树实现消息可信度的评估。文献[14]把假消息的检测问题转化为分类问题,基于支持向量机(support vector machine, SVM)的方法,利用从推文中提取的 45 个特征,包括推文内容、作者特征以及有关外部 URL 的信息等,对推文的可信度进行评分,依此识别虚假消息。文献[19]提出了一种在开放域中对非结构化文本进行可信度分析的通用方法,利用消息的语言风格和来源可靠性来评估其可信度。文献[11]利用深度学习的方法提取文本时空特征进行假消息的识别。文献[15]提出了一种基于递归神经网络的深度关注模型,选择性地学习文本的表示形式以进行谣言识别。该模型将注意力机制用在递归层面学习不同特征,并生成隐藏的表示,以捕获相关推文随时间变化的情况。以上现有

的各类方法一方面需要人工提取特征,且提取何种类型的特征需要领域专家的参与,耗时耗力。除此之外,需要人工提取的特征,比如传播数据、关注数等,往往在微博消息发表的初期是采集不到的,限制了该类方法的实时性;另一方面,仅仅通过文本信息的特定特征识别假消息,忽略了微博中包含的其他模态信息对检测的作用。

此外,最近的研究表明,视觉特征是用来检测假新闻非常重要的依据[1,6]。但是,关于验证社交媒体上多媒体内容的可信度的研究非常有限[10]。此外,文献[6,7]探索研究了微博内容中的视觉信息基本特征的提取,但是这些特征的获取仍是采用人工方式,不能代表视觉内容的复杂分布[10],因此通过这些特征并不能很好地识别假消息。

还有,社交上下文信息也为假消息的检测提供丰富的信息,比如消息传播方式、转发数、评论数和评论内容等。文献[20]探索利用消息的传播模式挖掘假消息出现时特定的特征。然而,消息传播此类数据的获取十分困难,且需要消息传播之后才能检测,无法做到实时或者准实时地进行真假识别。

仅基于文本或者图像数据进行假消息检测的方法,忽略了两者之间包含的隐形关联信息,因此,近几年通过融合图像和文本信息的检测方法逐渐被提出来。

现阶段,由于深度学习在算力、模型处理能力等各方面的提升,大部分多模态融合模型均是基于深度学习的思路,包括图像描述(image captioning)[20-21]和视觉问答(visual question answering, VQA)[22]。在基于多模态数据的假消息检测方面,文献[12]采用循环神经网络(RNN)融合图像、文本和社交上下文信息,其中,社交上下文信息是一些统计信息,包括正面词汇数量、负面词汇数量、URL 中包含的@符号数量、微博文本的情感得分、评论的数量等信息。对于给定的推文,首先让其文字和社交上下文信息采用长短期记忆网络(long short-term memory, LSTM)方式融合;然后将上一步获取的融合表示与采用预训练的卷积神经网络(convolutional neural network, CNN)方法获取的视觉特征融合。在融合过程中,LSTM 的每一个时间步长的输出都会采用注意力机制和视觉特征融合。文献[16]模型主要由 3 个主要部分组成,即多模态特征提取器、事件鉴别器和假新闻检测器。事件鉴别器采用对抗神经网络方式移除特定事件的特征,确保模型学习到推文中和事件无关的图像和文字的共享特征,通过学习识别虚假新闻的可辨别表示,提高假新闻检测的准确率。文献[10]采用了变分自动编码器(variational autoencoder)思想,模型由 3 个主要部分组成,即一个编码器、一个解码器和一个假新闻检测器,解决了在推文多模态数据之间学习共享表示这一挑战,以帮助假新闻检测。以上的相关方法存在的缺陷是:在已有带标签的数据集上训练,没有充分使用社交媒体中无标签的数据信息;同时也没有考虑针对图像的细粒度处理。

在融合模型方面,文献[23]提出了双向注意力来解决视觉和语言任务,提出了一种新的联合图像和文本特征的协同显著性的概念,使得两个不同模态的特征可以相互引导。此外,该文作者也对输入的文本信息,从多个角度进行加权处理,构建多个不同层次的图像问题联合注意力映射(image-question co-attention maps),即词级别(word-level)、短语级别(phrase-level)和问题级别(question-level)。最后,在短语级别,作者提出一种新颖的卷积-池化策略(convolution-pooling strategy)自适应地选择短语规模。文献[24]对模型和注意力机制进行了详细的探究,提出了经典的 BiDAF(双向注意流)模型,该模型计算了两种注意力,从上下文到问题,以及从问题到上下文。文献[18]在机器翻译任务中提出了 Transformer 模型,之后被应用于各类任务中。Bert[25]是 Google 在 NLP 方面的一个重要工作,使 NLP 预训练模型思想更加得成熟,可以说一定程度上改变了 NLP 领域的研究方式,之后基于预训练思想的各类模型出现[26-27]。总体的思想都是采用通用模型架构在语料库(Corpus)上预训练(pre-training);然后针对具体的任务,在通用模型架构上增加几层,固定通用模型的参数,微调(fine-tuning)增加的若干层参数。在跨模态信息融合方面,LXMERT[26]构建了一个多层的 Transformer 模型,它含有 3 个编码器:即一个对象关系编码器、一个语言编码器和一

个跨模态编码器。首先,采用对象关系编码器和语言编码器分别对文本和图像单独建模表示,然后将两种模态的结果与交叉模态转换器结合在一起。为了让模型具备联系视觉和语言语义的能力,用了大量的图像和句子对进行了模型预训练。文献 VisualBERT[27]采用了一组层叠的 Transformer 层,使用自我注意力机制把输入的一段文本和一张输入图像中的区域隐式地对齐起来。同时,作者还提出了两个在图像描述数据上的视觉-语言关联学习目标,用于 VisualBERT 的预训练。以上的模型主要是基于有标签数据集应用于 VQA、VCR 等任务,且大部分的模型都是通过两个单独分支对文本和图像分别处理,然后再对各自得到的结果融合。本文提出的模型借鉴了语言模型中的 Bert 思想,基于公众媒体平台上的大规模无标签数据实现自监督学习,实现文本和图像融合,通过预训练步骤实现在没有额外显式监督的条件下学习多模态的高阶特征,然后基于有标签数据微调模型,最终利用图像和文本的融合表示识别假消息。

在假新闻检测方面,先前的工作对图像的处理都是采用预训练的 CNN 模型,比如 VGG19,获取整张图像的特征。但最近的研究工作[28-30]均建议对图像进行细粒度处理,使用图像目标检测模型获取重点区域(regions of interest, ROI)作为图像的描述信息,然后把重点区域作为模型的输入。其中,文献[29]把图像检测模型和 Bert 模型结合,同时进行训练。从以上的研究中可以看出,基于图像重点区域的图像描述信息可以输入模型中,从而取得很好的效果。

本文提出的基于预训练思想的假消息检测方法将图像进行细粒度处理,获取图像各个重点区域,然后将图像各个重点区域和文本信息一同作为模型输入进行预训练,学习图像和文本的融合表示,进行假消息的识别。该方法不仅可以充分使用社交媒体网络中已有的图像和文本信息,同时也有效地缓解由数据不均衡导致的假新闻检测准确度不高的问题。

2 大规模图像-文本数据收集

本节主要介绍如何收集大量的同时含有图像和文本的数据集。目前,在自然语言处理领域,有非常多的文本语料可以使用,包括 BooksCorpus[31]、Wikipedia 和新闻语料[32]等;同时,在涉及图像和文本融合的任务中,现阶段,大部分的预训练模型[27-29,33]都是采用两个数据集:The Conceptual Captions(TCC)[34]和 SBU Captions[35],其中,TCC 数据集从互联网中的网页收集,含有 300 万张图片以及对图片的描述信息,SBU 数据集含有 100 万张图片以及对应的标题。本文的目标是识别同时含有图像和文字的假消息,由于社交网络中不同用户发表的图片和文字消息在语言风格、内容等方面有较大的差异,检测模型不能直接应用于以上数据集,因此需要在社交媒体中收集大量的高质量的同时含有图像和文字的数据作为预训练集。

基于以上需求,本文设计了社交媒体数据收集方法,下面以新浪微博为例,说明采集数据的具体过程。

数据采集。微博用户达到数亿级别,每个用户发表信息的质量参差不齐。为了确保采集数据的质量,从权威用户发布的信息中采集数据,文献[16]使用的微博数据集中的真消息都是从微博权威用户中获取的,比如人民日报、新华网等,因此以本数据集中的权威用户为基础,爬取该类用户的数据。本文所采集的数据年份为 2010 年 9 月至 2020 年 4 月,采集的原始数据数量为 18 万条。

数据过滤。在收集数据的过程中,为了获取高质量的数据,根据图像的内容和文本的内容对数据进行过滤。针对图像,把图像低于 300×300 像素的数据丢弃,同时,也将丢弃不能被模型识别的 GIF 动态图;针对文本信息,把文本低于 10 个字的数据丢弃。为了确保文本信息的质量,会过滤一些特殊的符号,比如@、空格等信息。最终,过滤之后,收集了大约 13 万条同时包含图像和文字的数据。

3 基于预训练模型的假新闻识别

图 1 是模型的整体框架,本文借鉴自然语言处理领域中 Bert 模型思想,使用 Transformer 作为基础的结构。Bert 中学习的是文本之间的相互关系,本

文和 Bert 不同的是，本文的模型需要学习文本、图像以及文本和图像之间的关联关系，因此在模型数据输入阶段，本文会将图像看作文本，同时把图像和文本的表示输入模型中。图像和文本采用不同的编码器分别进行编码，其中，图像的编码通过图像检测模型 Faster-RCNN[36] 获取，该模型会对一张图像进行分割，提取重要的区域；模型中文本的每一个输入代表一个字。图像和文本输入模型经过多层 Transformer 之后，模型会融合两种模态的数据，最终学习一个文本和图像的融合表示。

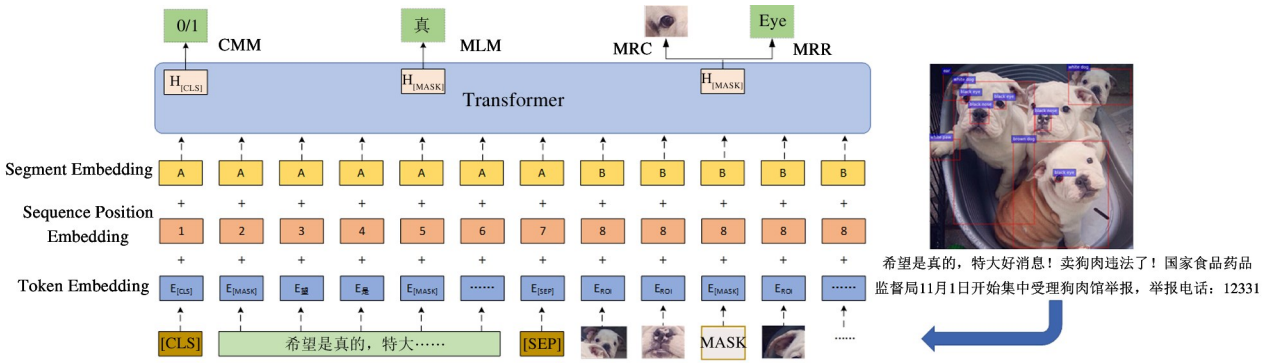

图 1　假消息检测模型框架图

本文采用模型在训练时包括两个阶段：预训练阶段和微调阶段。两个阶段数据输入一致，都包括图像和文本，不同的是在微调阶段仅需一个目标任务即可。本节将详细介绍以上两个阶段，其中，在预训练阶段将说明采用何种预训练任务使模型获取好的模型预训练参数，从而可以在微调阶段获取较优的模型参数以进行假消息的识别。

3.1　模型输入

模型输入包含文本和图像两部分，下面分别予以说明。

文本嵌入表示。首先需要构造模型的文本输入，本文采用中文全词覆盖（whole word masking）的方法处理文本信息[37]。

文本数据采用上述方法处理完成之后，整个文本就分成了词的序列。在文本序列的起始位置添加特殊字符[CLS]，在序列的结束位置添加特殊字符[SEP]。字符[CLS]的作用是在模型输出时作为图像和文本的共享表示，字符[SEP]的作用是作为图像和文本的分隔符。之后，如式（1）～（3）所示，需要做字符嵌入 we_i、字符位置嵌入 wpe_i 和输入类型的嵌入 wte_i，通过各个嵌入层，把各信息映射至向量，其中输入类型表示输入的是文本还是图像。

$$we_i = WordEmbedding(w_i) \tag{1}$$

$$wpe_i = PositionEmbedding(i) \tag{2}$$

$$wte_i = TokenEmbedding(wt_i) \tag{3}$$

式中 w_i 代表了第 i 个位置的词语，wt_i 代表了输入类型。最后采用和 Bert 中相同的策略，每一个字符的嵌入表示是字符嵌入、字符位置嵌入和输入类型的嵌入的加和。

图像嵌入表示。与现有工作不同，本文直接采用通过预训练的 CNN 模型提取图片的特征。本文应用预训练好的 Faster-RCNN 模型[36] 提取 n 个候选框（RoI），该预训练模型基于 ResNet-101 实现，使用了 Visual Genome 数据集预训练。RoI 用其特征和对应的坐标位置表示，把提取出来的 n 个 RoI 的特征标识为 $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_n\}$，每一个 $\boldsymbol{c}_i$ 是一个 2048 维度的向量，该维度是 Faster-RCNN 模型提供的向量维度；每一个 RoI 的位置标识为 $\{p_1, p_2, \cdots, p_n\}$，每一个元素代表 RoI 的具体位置信息：

$$p_i = \left(\frac{p_i^{tx}}{W}, \frac{p_i^{ty}}{H}, \frac{p_i^{bx}}{W}, \frac{p_i^{by}}{H}\right) \tag{4}$$

其中，(p_i^{tx}, p_i^{ty}) 和 (p_i^{bx}, p_i^{by}) 分别代表某一个 RoI 在图像中左上角和右下角的坐标，通过该坐标可以确定 RoI 在图像中的位置，在这里，取 n 的值为 10。

图像信息的表示生成和文本信息处理过程类似，可以把 n 个 RoI 看做 n 个单词。需要对这 n 个 RoI 进行特征嵌入、位置嵌入、类型嵌入、图像坐标位置嵌入。其中，针对特征嵌入，由于已经获取了每

一个区域的特征,特征映射的作用是把特征向量采用多层感知机方式映射到和文本相同维度的向量空间。与文本处理不同的是,本文同时应用了 RoI 在图像中的具体坐标位置信息:

$$ve_i = MLP(\boldsymbol{c}_i) \tag{5}$$

$$vpe_i = VisualPositionEmbedding(i) \tag{6}$$

$$vte_i = TokenEmbedding(vt_i) \tag{7}$$

$$vce_i = CorEmbedding(p_i) \tag{8}$$

最终,每一个图像的嵌入表示是特征嵌入、位置嵌入、类型嵌入和图像坐标位置嵌入的总和,即:

$$v_i = (ve_i + vce_i)/2 + vpe_i + vte_i \tag{9}$$

位置和类型嵌入表示。无论是文本还是图像数据都使用位置嵌入信息,其目的是为了表示每一个元素在序列中的位置,其中,文本信息有着严格的顺序,按照从小到大的顺序排序。对于图像输入,由于每一个图像之间没有严格的顺序关系,因此在图像的位置嵌入中,位置变量都设置了相同的固定值;同时,类型表示输入的是文本还是图像,是为了区分多模态信息。针对文本信息的类型嵌入,类型变量全部取 0,即 $wt_i = 0$;针对图像信息的类型嵌入,类型变量全部取 1,即 $vt_i = 1$。

3.2 预训练任务

本小节将详细介绍模型在预训练过程中所采用的预训练任务。本文主要采用了 4 种预训练任务,分别是掩码语言模型(masked language modeling, MLM)、掩码区域分类(masked ROI classification, MRC)、掩码区域特征回归(masked ROI regression, MRR)和多模态匹配(cross-modality matching, CMM)。

掩码语言模型。在文本输入模型时会遮掩一部分词,在模型的最终输出时预测这些被遮掩的词,其目的是为了捕捉句内不同单词之间的关系。与 Bert[25] 模型不同的是,在预测这些被遮掩词的时候,不但利用了文本中非遮掩的词,同时也利用了先前提取的 n 个 RoI 信息,基于此种方式,可有效捕获视觉和语言内容之间的依赖关系。在执行遮掩时,文本中的词会随机按照 15% 的概率遮掩,具体地,如果某个词汇被选中遮掩,那么有 3 种遮掩方式:(1)该词以 80% 的概率被一个特殊字符[MASK]代替;(2)该词以 10% 的概率替换为任意的词;(3)该词以 10% 的概率保持不变。在预测时,本文采用常用的交叉熵作为损失函数:

$$\mathfrak{L}_{mlm} = -E_{x\in D}\sum_{j=1}^{M} CE(s(\boldsymbol{w}_o^j),\ h_k(\hat{\boldsymbol{w}}_o^j)) \tag{10}$$

式中 D 代表训练数据集,$\boldsymbol{w}_o^j$ 代表文本中被遮盖的 M 个词中的第 j 个,$s(\boldsymbol{w}_o^j)$ 为真实标签值。$\hat{\boldsymbol{w}}_o^j$ 对应于 Transformer 模型中针对该位置的输出向量。通过添加一个多层感知机以预测正确的词语,多层感知机的输入即 $\hat{\boldsymbol{w}}_o^j$,输出为 $h_k(\hat{\boldsymbol{w}}_o^j)$。

掩码区域分类。通过遮掩视觉特征并预测视觉分类信息,让模型理解视觉,达到让视觉信息和文本信息匹配对齐的目的。由于预测视觉分类信息是同时基于未被遮掩的文本信息和视觉信息,促进了视觉信息和语言信息的融合。遮掩视觉特征信息时,和掩码语言模型类似,会随机按照 15% 的概率遮掩视觉特征。在这里,同样有 3 种遮掩的方式:(1)该视觉特征以 80% 的概率被 0 代替;(2)该词以 10% 的概率替换为任意的其他特征;(3)该词以 10% 的概率保持不变。在预测时需要用到分类的标签信息,此信息从 Faster R-CNN[34] 中获取,同样采用交叉熵作为损失函数:

$$\mathfrak{L}_{mrc} = -E_{x\in D}\sum_{j=1}^{N} CE(l(\boldsymbol{v}_o^i),\ h_v(\hat{\boldsymbol{v}}_o^i)) \tag{11}$$

其中,$\boldsymbol{v}_o^i$ 代表被遮盖的 N 个 RoI 中的第 i 个,$l(\boldsymbol{v}_o^i)$ 为真实标签值。$\hat{\boldsymbol{v}}_o^i$ 对应于 Transformer 模型中针对该位置的输出向量,通过添加一个多层感知机以预测正确的分类,多层感知机的输入即 $\hat{\boldsymbol{v}}_o^i$,输出为 $h_v(\hat{\boldsymbol{v}}_o^i)$。

掩码区域特征回归。该任务和 MRC 的目的相同,都是为了能够让模型学习理解视觉信息,让视觉信息和文本信息匹配对齐。MRR 和 MRC 相比,可以更加精确地学习视觉信息。该任务的目标是针对遮掩的视觉区域,能够预测具体的特征。在实现的过程中,本文会在 Transformer 模型的输出之后,添加一个全连接层,该层输出维度和视觉特征的输入维度一致,在这里使用的损失函数是 L2 损失函数。

$$\mathfrak{L}_{mrr} = -E_{x\in D}\sum_{i=1}^{N} \| (m(\boldsymbol{v}_o^i) - r_v(\hat{\boldsymbol{v}}_o^i)) \|_2^2 \tag{12}$$

其中,$m(\boldsymbol{v}_o^i)$ 为 $\boldsymbol{v}_o^i$ 的特征嵌入,$r_v(\hat{\boldsymbol{v}}_o^i)$ 为通过一个

多层感知机预测的图像特征。

多模态匹配。除了以上 3 个关于文本和视觉的任务之外,本文还设置了一个多模态的匹配任务,该任务的目的是为了让模型学习文本信息和视觉信息是否匹配。在训练的过程中,针对每一条包含图像和文本的训练数据,本文以 0.5 的概率替换训练条目的视觉信息为其他任意视觉信息,使文本和视觉信息不匹配,以此生成负样本。模型会训练一个分类器对是否匹配做出预测,在模型输入章节,在文本的前面添加一个特殊字符[CLS];在训练时,会在该特殊字符的输出后面添加一个全连接层,得到一个分类结果,采用二分类交叉熵作为损失函数。

$$\mathfrak{L}_f = -E(y_f \log(h_f(\hat{\boldsymbol{x}}_o)) + (1-y_f)\log(1-h_f(\hat{\boldsymbol{x}}_o))) \tag{13}$$

其中,$\hat{\boldsymbol{x}}_o$ 代表特殊字符[CLS]的模型输出,$h_f(\hat{\boldsymbol{x}}_o)$ 为通过添加一个多层感知机以预测多模态信息是否匹配的输出值,y_f 为真实标签值。

该模型的完整目标函数定义如下:

$$\mathfrak{L} = \lambda_1 \mathfrak{L}_{mlm} + \lambda_2 \mathfrak{L}_{mrc} + \lambda_3 \mathfrak{L}_{mrr} + \lambda_4 \mathfrak{L}_f \tag{14}$$

其中,λ_1、λ_2、λ_3、λ_4 代表各个损失的权重,其值分别设置为 1、6.6、6.6、1。

3.3 模型执行

本文所提检测模型主要包括模型预训练及模型调整。

模型预训练。针对输入的文本,首先采用 Bert[25] 中提供的WordPieceTokenizer[38] 的分词方式实现句子单词级的切分,然后使用中文分词工具实现对句子词语级别的划分,最终基于这两个切分的列表实现中文全词覆盖,本模型使用的中文分词工具是 Jieba 分词工具。针对输入的图像,使用在 Visual Genome[36] 上预训练的 Faster R-CNN[39] 模型对图像处理,不同于文献[39]的做法,针对每一张图像,其获取的候选框数量是一个动态变化的数值,而本文固定获取 10 个候选框(RoI),这样有助于对输入模型时的数据进行预处理操作,不用对候选框少的图像进行补全对齐操作。在模型结构参数方面,采用了 12 层的 Transformer 模型,隐状态向量维度为 768 维,中间向量维度大小为 3076 维。在预训练过程中使用了多个预训练任务,因此有多个损失。模型训练时,最终损失的大小是所有损失的总和。训练的过程使用 Adamw 作为模型优化器,学习率为 1e-4,批数量大小设置为 50,训练轮数为 65。

模型调整。调整过程就是应用从微博中获取的人工标注假新闻数据集,对模型进行训练,以便让模型能够适应假新闻识别的任务。模型调整的过程中仅仅判断消息的真假,不再执行预训练任务,由于本文仅执行假新闻检测任务,没有其他任务,因此没有采用在 LXMERT[26]、VisualBERT[27] 和 VL-BERT[29] 等其他研究中采用的仅仅微调模型中几层神经网络参数的策略,而是对模型的所有参数进行修改。在模型调整的过程中,设置学习率为 1e-5,批数量大小设置为 40,训练 100 轮。

4 实验评估

本节将对所提方法的有效性进行验证及分析。首先,介绍测试使用的数据集,并说明对比的基准方法;然后,对实验结果进行分析,验证本文所提模型的有效性。

4.1 数据集

当前,同时含有图像和文本的用于假消息检测的数据集主要有 2 个:Tweet 数据集和新浪微博数据集。Tweet 数据集的隐私政策,无法获取数据,因此本文主要在新浪微博数据集上进行实验评估。下面从数据集大小和数据特点等方面分别介绍这个数据集。

在假消息的检测方面,新浪微博数据集已经被诸多研究工作使用[10,12,16],其从官方渠道采集数据,例如人民日报、新华网等,数据集的爬取时间为 2012 年 5 月至 2016 年 1 月,后续本文把该数据标识为 weibo-T。针对该数据集,首先移除了没有同时包含图像和文本的微博,然后移除重复的图片和低质量的图片,以确保数据集的质量。由于该数据集爬取的截止时间为 2016 年 1 月,之后又有许多假消息产生,为了进一步验证模型的性能,本文又进一步从新浪微博官方渠道(https://service.account.weibo.com/index?type=5&status=4&page=1)中爬取了数据,该渠道鼓励普通用户报告可疑帖子,并由专门

的人员检查帖子的真实性。本文爬取数据的截止时间为2020年5月,后续把该数据集标识为weibo-O。在预处理此数据集时,遵循和以往工作[12]中相同的步骤,首先删除了低质量的图像,以确保整个数据集的质量,然后统计正样本和负样本的数量,最后将整个数据集按照7:1:2的比例分为训练集、验证集和测试集。在生成数据的过程中,为了确保各集合中的数据不会重复,本文设计了数据集生成算法,如算法1所示。为了验证训练数据和测试数据的相关性对检测模型的影响,算法在具体执行的过程中,需要相关系数参数,取值设置为从0.2到1且步长为0.1的9个数值,这样就生成了9对数据集,数据集的详细信息如表1所示。表中数据用斜杠分割,分别表示在某个相关系数下的假新闻和真新闻的数量。为了公平比较,对weibo-T和weibo-O两部分数据集分别进行测试,以验证所提模型的可行性。

算法1 对比数据集生成算法

输入:所有的带标签的数据集D,相关系数c,数据比例R;

输出:模型调整集、测试集。

1. P, N←*CountNumber*(D,R);/* 计算需要生成的正样本数量、负样本数量 */
2. *shuffle*(D);/* 随机打乱数据集D,每次生成的集合有所不同,在一定程度上验证模型的泛化能力 */
3. PCL,PSet ← *generatePositive*(D,P,N,R);/* 除了生成正样本PSet之外,还有返回一个包含正样本内容的列表PCL,列表的每一项是经过中文分词之后的结果 */
4. for e in (D/*PSet*): /* 排除已使用的数据集 */
5. s←*calCoffeicent*(e,*PCL*);/* 计算和已经加入了调整集合中的文本相似度 */
6. if s is greater than c:
7. generateNegitive(e) /* 生成一条负样本 */
8. end for

表1 新浪微博数据集详情

名称＼系数	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
weibo-T	279/156	672/403	1129/662	1550/899	2068/1265	2390/1551	2706/1928	2960/2281	3580/4055
weibo-O	833/269	2029/620	3050/983	3833/1286	4760/1706	5229/2034	5709/2443	6054/2827	6299/3293

4.2 对比方法

为了广泛验证本文模型,选择了两类方法进行对比,即单模态方法和多模态方法。

单模态方法。由于数据集包含图像和文本两种模态,每一类模态都可以单独使用作为假消息检测的依据,因此,可看作是单模态方法。

基于文本的检测方法(Text)。该方法仅仅使用文本信息作为检测依据。使用CNN模型来提取文本特征作为检测的依据,在使用时把每一个词编码为32维的向量,经过CNN提取特征得到结果之后,接一个全连接层,全连接层采用的维度大小也是32维,然后采用softmax方式得到预测结果。CNN模型的参数设置采用和文献[14]相同的配置,使用20个过滤器(filter),每一个过滤器的窗口大小(window size)从1到4。

基于图像的检测方法(Vis)。该模型仅仅使用图像信息判断是否为假消息。使用预训练好的VGG-19对图像进行处理,获取图像特征,然后接一个32维的全连接层获取最终的预测结果。

多模态方法同时使用图像和文本信息来检测是否为假消息,目前利用多模态对假新闻进行识别的方法主要有两个,即EANN[16]和MVAE[10]。

EANN。该框架利用神经网络学习未见事件的可传递特征。它由3个主要组件组成,即多模态特征提取器、假新闻检测器和事件鉴别器。多模态特征提取器提取微博中文本和图像的共有特征,其与假新闻检测器配合使用,以学习用于识别假新闻的显著特征表示。同时,事件鉴别器通过去除事件特定特征来学习事件不变表示。该模型也可以只使用两个组件来检测假新闻,即多模态特征提取器和假新闻检测器。因此,同MVAE[10]一样,为了进行公平的比较,实验中使用了一个不包括事件鉴别器的EANN变体。

MVAE。该方法为解决在推文中学习各模态之

间相关性的挑战，提出了一种多模态变分自编码器模型，模型由 3 个主要部分组成：编码器、解码器和假消息检测器。基于文本和图像的重建方式，联合训练编码器、解码器和假消息检测器，最终得到多模态数据（图像和文本）的共享表示，依此进行假消息检测。

4.3　结果分析

本节中，进行了 2 组实验来验证本文提出的假消息检测模型的有效性。第 1 组实验是通过在已有数据集和本文采集的数据集上进行，数据集的详细信息如表 1 中相关系数为 1 的列所示。该实验会计算模型检测准确率、召回率等指标，判断模型的有效性。表 2 展示了本文所提方法以及对比方法的实验结果，针对数据中包含的假消息和真实消息，分别列出了各检测方法检测结果的准确率、召回率和 F1 分数。从表中可以看到，总体来说，本文所提方法在检测准确率上要优于各对比方法。

表 2　在新浪微博数据集上的实验结果

数据集	方法	准确率	精确率	召回率	F1
weibo-T	Text	0.828	0.87	0.83	0.84
	Vis	0.638	0.74	0.64	0.67
	EANN	0.846	0.88	0.85	0.85
	MVAE	0.822	0.86	0.82	0.83
	本文方法	**0.873**	**0.89**	**0.87**	**0.88**
weibo-O	Text	0.851	0.85	0.85	0.85
	Vis	0.631	0.63	0.63	0.63
	EANN	0.828	0.830	0.83	0.83
	MVAE	0.817	0.86	0.82	0.83
	本文方法	**0.862**	**0.86**	**0.86**	**0.86**

在 2 个数据集中，仅通过文本识别假消息的准确率要明显高于仅通过图像识别。也就是说在数据集中，相对于图像数据，文本信息提供了更加丰富的语义特征来辅助识别假消息。在 weibo-T 数据集中，本文所提方法和基线方法相比，检测准确率提升了 2.7%，从 84.6% 提升到了 87.3%，F1 分数从 85% 提高到了 88%；在 weibo-O 数据集中也表现出了类似的趋势，检测准确率和 F1 分数也有了提升，其中检测准确率从 85.1% 提升到了 86.2%，F1 分数从 85% 提高到了 86%。

在第 2 组实验中，为了验证训练数据和测试数据的相关性对检测模型的影响，本文采用算法 1 生成的训练数据集合对模型参数训练调整，并用对应的测试集测试训练好的模型。在这里用本文所提模型和 EANN 做比较，最终结果如图 2 所示。从图中可以看到，在 2 个数据集中，本文提出的模型全面优于 EANN 方法。在 weibo-T 数据集中，随着相关系数的增加，本文所提模型的准确率从 69.8% 提高到了 87.3%，EANN 模型的准确率从 62.8% 提高到了 84.6%，通过对比可以发现，随着相关系数的增加，由于测试集和验证集中能够匹配到的词语在增多，所以经过测试集训练的模型，在验证集上的检测准确率也逐渐上升，符合直观的理解。同时，从图中可以看到，在相关系数相同的条件下，本文所提模型识别假消息的准确率也高于其他模型，在weibo-T数

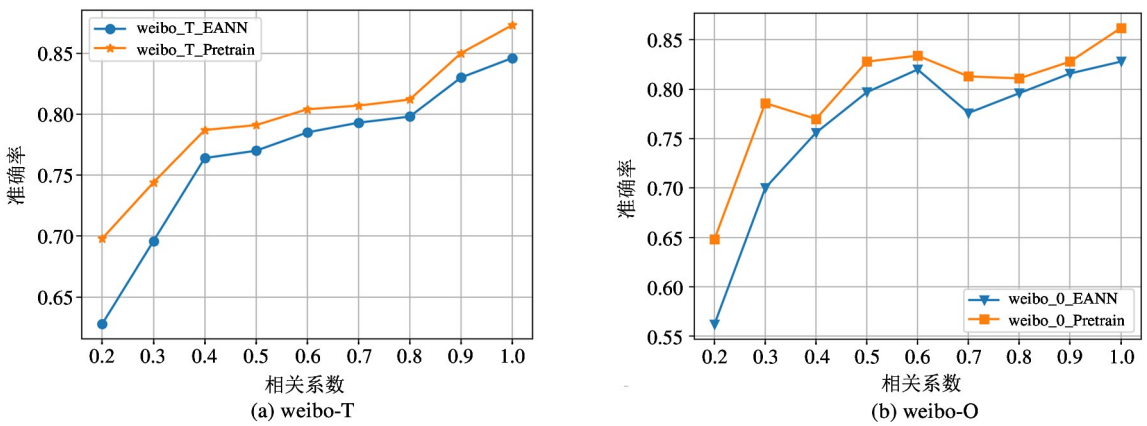

图 2　模型检测准确率对比示意图

据集中,准确率的变化幅度在1.4% ~7% 之间;在weibo-O 数据集中,准确率的变化幅度在 1.2% ~ 8.6% 之间。通过该实验可以证明,当测试集和验证集中的数据分布不均衡时,本文所提方法有明显优势。上述现象出现是由于用户发表微博消息中文本的多样性,导致训练集和测试集中的数据可能存在较大差异性,同时现有的模型并没有很好地学习文本之间的关系,从而导致用训练集训练的模型不能很好地对测试集中的数据进行检测,模型效果不佳。

5 结 论

本文提出了一种基于预训练方式的假消息检测方法。基于该方法可以充分利用社交媒体中已有的大量多模态数据,基于多个预训练任务有效地融合消息中图像和文本信息,最终,基于多模态的融合表示有效地识别假消息。实验结果表明,本文提出的假消息检测方法在准确度方面优于现有的检测方法,并缓解了在数据内容分布不均衡时造成的模型检测准确率下降问题。

未来的工作将进一步考虑基于多模态的假消息识别方法,并从以下几个方面进行尝试:(1) 在实际应用场景中,越来越多的用户发表的内容中包含视频信息,而目前大多数的方法都是建立在文本或者图像之上,没有对视频数据分析处理,基于视频和文本信息的假消息识别值得更多的关注;(2) 将用户对微博的评论信息引入,进一步提升假消息检测的准确信。

参考文献

[1] SHU K, SLIVA A, WANG S, et al. Fake news detection on social media: a data mining perspective[J]. *ACM SIGKDD Explorations Newsletter*, 2017, 19(1): 22-36

[2] LAZER D M J, BAUM M A, BENKLER Y, et al. The science of fake news[J]. *Science*, 2018, 359(6380): 1094-1096

[3] VOSOUGHI S, ROY D, ARAL S. The spread of true and false news online[J]. *Science*, 2018, 359(6380): 1146-1151

[4] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C]//Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 2011: 675-684

[5] WU K, YANG S, ZHU K Q. False rumors detection on Sinaweibo by propagation structures[C]//2015 IEEE 31st International Conference on Data Engineering, Seoul, Korea, 2015: 651-662

[6] JIN Z, CAO J, ZHANG Y, et al. Novel visual and statistical image features for microblogs news verification[J]. *IEEE Transactions on Multimedia*, 2016, 19(3): 598-608

[7] GUPTA M, ZHAO P, HAN J. Evaluating event credibility on twitter[C]//Proceedings of the 2012 SIAM International Conference on Data Mining, Anaheim, USA, 2012: 153-164

[8] JIN Z, CAO J, JIANG Y G, et al. News credibility evaluation on microblog with a hierarchical propagation model[C]//2014 IEEE International Conference on Data Mining, Shenzhen, China, 2014: 230-239

[9] JIN Z, CAO J, ZHANG Y, et al. News verification by exploiting conflicting social viewpoints in microblogs[C]//The 30th AAAI Conference on Artificial Intelligence, Phoenix, USA, 2016: 2972-2978

[10] KHATTAR D, GOUD J S, GUPTA M, et al. Mvae: multimodal variational autoencoder for fake news detection[C]//The World Wide Web Conference, San Francisco, USA, 2019: 2915-2921

[11] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, USA, 2016: 3818-3824

[12] JIN Z, CAO J, GUO H, 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, USA, 2017: 795-816

[13] 李璐旸,秦兵,刘挺. 虚假评论检测研究综述[J]. 计算机学报, 2018, 41(4): 946-968

[14] GUPTA A, KUMARAGURU P, CASTILLO C, et al. Tweetcred: real-time credibility assessment of content on twitter[C]//International Conference on Social Informatics, Barcelona, spain, 2014: 228-243

[15] CHEN T, LI X, YIN H, et al. Call attention to rumors:

deep attention based recurrent neural networks for early rumor detection[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, Australia, 2018: 40-52

[16] WANG Y, MA F, JIN Z, 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, UK, 2018: 849-857

[17] 段大高, 谢永恒, 盖新新, 等. 基于神经网络的微博虚假消息识别模型[J]. 信息网络安全, 2017(9): 134-137

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, Long Beach, USA, 2017: 5998-6008

[19] POPAT K, MUKHERJEE S, STRÖTGEN J, et al. Credibility assessment of textual claims on the web[C]// Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, USA, 2016: 2173-2178

[20] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3156-3164

[21] KARPATHY A, FEI-FEI L. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3128-3137

[22] AGRAWAL A, LU J, ANTOL S, et al. VQA: visual question answering[J]. *International Journal of Computer Vision*, 2017, 123(1):4-31

[23] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]// Advances in Neural Information Processing Systems, Barcelona, Spain, 2016: 289-297

[24] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[EB/OL]. https://arxiv.org/pdf/1611.0160.pdf: arXiv, (2018-06-21),[2020-12-20]

[25] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. https://arxiv.org/pdf/1810.04805.pdf:arXiv, (2019-05-24),[2020-12-10]

[26] TAN H, BANSAL M. Lxmert: learning cross-modality encoder representations from transformers[EB/OL]. https://arxiv.org/pdf/1908.07490.pdf:arXiv, (2019-12-03),[2020-12-10]

[27] LI L H, YATSKAR M, YIN D, et al. Visualbert: a simple and performant baseline for vision and language[EB/OL]. https://arxiv.org/pdf/1908.03557.pdf: arXiv, (2019-08-09),[2021-01-12]

[28] LI G, DUAN N, FANG Y, et al. Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training[EB/OL]. https://arxiv.org/pdf/1908.06066.pdf:arXiv, (2019-12-02),[2020-12-20]

[29] SU W, ZHU X, CAO Y, et al. Vl-bert: pre-training of generic visual-linguistic representations[EB/OL]. https://arxiv.org/pdf/1908.08530v4.pdf:arXiv, (2020-02-18),[2020-12-20]

[30] CHEN Y C, LI L, YU L, et al. Uniter: learning universal image-text representations[EB/OL]. https://arxiv.org/pdf/1909.11740.pdf:arXiv, (2019-09-25),[2020-12-20]

[31] ZHU Y, KIROS R, ZEMEL R, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books[C]// Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 19-27

[32] XU L, ZHANG X, LI L, et al. CLUE: a Chinese language understanding evaluation benchmark[EB/OL]. https://arxiv.org/pdf/2004.05986.pdf:arXiv, (2020-11-05),[2020-12-20]

[33] ALBERTI C, LING J, COLLINS M, et al. Fusion of detected objects in text for visual question answering[EB/OL]. https://arxiv.org/pdf/1908.05054.pdf: arXiv, (2019-11-03),[2020-12-20]

[34] SHARMA P, DING N, GOODMAN S, et al. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018: 2556-2565

[35] ORDONEZ V, KULKARNI G, BERG T L. Im2Text: describing images using 1 million captioned photographs[C]// Advances in Neural Information Processing Systems,

Granada, Spain, 2011: 1143-1151

[36] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6077-6086

[37] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT[EB/OL]. https://arxiv.org/pdf/1906.08101.pdf: arXiv, (2019-10-29), [2020-12-20]

[38] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation[EB/OL]. https://arxiv.org/pdf/1609.08144.pdf: arXiv, (2016-10-08), [2020-12-10]

[39] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems, Montreal, Canada, 2015: 91-99

Multi-modal deep fusion based fake news detection method

JING Quanliang, FAN Xinxin, WANG Baoli, BI Jingping, TAN Haining

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

(University of Chinese Academy of Sciences, Beijing 100049)

Abstract

Intelligent detection on fake news has become one of most important tasks in social networks. This work aims to identify multi-modal fake news that contains both image and text information. To date some existing efforts have been devoted to addressing this issue, however, they realize multi-modal utilization by simply splicing various modal features, ignoring the multi-modal integration (fusion representation) in proper manner, which leads to poor performance on detection. To circumvent the above problems, a pre-train-oriented multi-modal fusion method to detect fake news is proposed, which makes full use of large numbers of messages containing multi-modal data in social media to achieve effective detection of fake news. For the method, a representation for multi-modal information is learned to enhance the capability of multi-modal fusion through training different tasks, in this way, fake news can be effectively identified in a high accuracy. The multi-facets experiments using real-world dataset SinaWeibo show that the proposed method is rational and outperforms other baselines. Furthermore, the model used in this paper can effectively alleviate the problem of fast degradation of detection accuracy caused by uneven distribution of training set and test set.

Key words: fake news detection, pre-training, multi-modal fusion, social media