基于双分支网络联合训练的虚假新闻检测

郭铃霓[1],黄 舰[1],吴兴财[1],杨振国[1],刘文印[1,2]
1.广东工业大学 计算机学院,广州 510006
2.鹏城实验室网络空间安全研究中心,广东 深圳 518000

摘 要:虚假新闻在社交媒体上的广泛传播,给社会带来了不同程度的负面影响。针对虚假新闻早期检测任务中,社交上下文信息不充分的问题,提出一种基于双分支网络联合训练的虚假新闻检测模型。该模型由最大池化网络分支(max pooling branch,MPB)和广义均值池化网络分支(generalized mean pooling branch,GPB)组成。MPB采用卷积神经网络对新闻文章进行文本特征提取,GPB引入了可训练的池化层,学习新闻文章潜在的语义特征。同时,在每个分支网络中,对新闻标题和正文之间进行语义关联性度量。最终,对两个分支网络联合训练后的结果进行决策融合,判断新闻的真实性。实验结果表明,提出的模型在准确率、召回率、F1值评测指标上均优于基线模型,F1值达到94.1%,比最优的基线模型提升了6.4个百分点。

关键词:虚假新闻早期检测;联合训练;双分支网络;语义关联性度量
文献标志码:A **中图分类号:**TP389.1 doi:10.3778/j.issn.1002-8331.2101-0001

Fake News Detection Based on Joint Training Two-Branch Network

GUO Lingni[1], HUANG Jian[1], WU Xingcai[1], YANG Zhenguo[1], LIU Wenyin[1,2]
1.School of Computer, Guangdong University of Technology, Guangzhou 510006, China
2.Cyberspace Security Research Center, Peng Cheng Laboratory, Shenzhen, Guangdong 518000, China

Abstract: The wide spread dissemination of fake news on social media has brought negative impact to the society. The key problem in detecting fake news at an early stage is that the news is just published lacking social context. To this end, the joint training two-branch network is proposed for detecting fake news. Specifically, the two-branch network is consisted of two branches, i.e., the max pooling network branch(MPB) and the generalized mean pooling network branch(GPB). MPB uses convolutional neural network to extract textual features of news articles, and GPB introduces a trainable pooling layer to learn the latent semantic features of news articles. Meanwhile, the semantic relevance between news title and the body text is measured in each branch. Finally, the results from the two branches are fused to judge the authenticity of news. The experimental results show that the proposed model outperforms the baseline models in terms of the evaluation metrics of accuracy, recall and F1-score, and achieves 94.1% on F1-score, which is 6.4 percentage points higher than other baselines.

Key words: fake news early detection; joint learning; two-branch network; sematic correlation metric

互联网时代,推特、微博、微信等在线社交媒体平台的快速发展,给读者获取新闻资讯提供了便利,也为虚假新闻的滋生和传播提供了土壤。《2019年网络谣言治理报告》(http://society.people.com.cn/n1/2019/1226/c1008-31524533.html)指出,2019年期间,微信平台共发布17 881篇辟谣文章,辟谣文章阅读量1.14亿次。其中,医疗健康、食品安全、社会科学是虚假新闻的高发领域。

虚假新闻的泛滥,给社会和人们的日常生活带来不同程度的负面影响。例如,新冠肺炎疫情期间,各种虚假新闻层出不穷,包括但不限于:“盐水漱口防病毒”

基金项目:国家自然科学基金(62076073);广东省基础与应用基础研究基金(2020A1515010616);广东省重点领域研发计划项目(2019B010136001);广东省重点科技计划项目(LZC0023)。

作者简介:郭铃霓(1996—),女,硕士研究生,CCF学生会员,研究方向为虚假新闻检测、跨模态检索,E-mail:LingniGuo@outlook.com;黄舰(1995—),男,硕士研究生,CCF学生会员,研究方向为虚假新闻检测;吴兴财(1993—),男,硕士研究生,研究方向为机器学习;杨振国(1988—),男,博士,副教授,CCF普通会员,研究方向为在线事件检测、域适应;刘文印(1966—),通信作者,男,博士,教授,研究方向为网络空间安全、区块链。

收稿日期:2021-01-04 **修回日期:**2021-05-13 **文章编号:**1002-8331(2022)15-0153-09

“喝板蓝根可以预防新型冠状病毒”“双黄连口服液能抑制新型冠状病毒”等。诸如此类的虚假新闻,导致相关商品遭哄抢脱销,不仅误导群众,还扰乱市场经济。Vosoughi等人指出,相比于真实新闻,虚假新闻传播更快、更频繁[1]。因此,对虚假新闻进行检测,具有重要意义。

最初的虚假新闻检测主要依赖于官方辟谣网站,由相关领域的多位专家对新闻的真实性进行研判。这种方式需要专家知识,不仅耗费大量的人力物力,而且时效性差。近年来,基于机器学习和深度学习的虚假新闻自动检测技术得到了发展。目前的虚假新闻检测方法大致可以分为基于内容的检测方法和基于社交上下文的检测方法。两种方法的区别在于是否使用社交上下文信息。例如,新闻在社交媒体上的传播路径、社交用户彼此之间的关系网络、社交用户的参与情况(点赞、转发、评论)等。社交上下文信息越丰富,越有利于虚假新闻检测。然而,基于社交上下文的虚假新闻检测方法不适用于虚假新闻早期检测,当新闻在新闻渠道上发布但尚未在社交媒体上传播时,社交上下文信息不充分。虚假新闻早期检测具有实际意义,当虚假新闻曝光的次数越多,并且反复出现在社交用户视野中时,用户越容易相信其真实性。一旦用户认为虚假新闻是真的,就很难改变他们的认知。基于新闻内容的检测,由于不需要考虑社交上下文信息,数据容易获取且能够实现虚假新闻的早期检测,受到越来越多的关注。已有研究通常把新闻文本内容作为整体,进行虚假新闻检测,较少考虑到新闻标题和正文之间的语义关联性。如果一则新闻并非真实发生,为了吸引读者,通常会采用猎奇、煽动性的标题,往往与正文内容无关。虽然带有“标题党”的新闻文章通常并不可靠,但并非所有这些新闻文章都是虚假新闻,这促使去探索虚假新闻和“标题党”之间的关系。

为解决上述问题,本文提出一种基于双分支网络联合训练的虚假新闻检测模型(jointly training two-branch network,JTTN),该模型由两个分支子网络构成,分别是最大池化网络分支(max pooling network branch,MPB)和广义均值池化网络分支(generalized mean pooling network branch,GPB)。MPB采用卷积神经网络进行文本特征提取,GPB在卷积神经网络的基础上,引入了可训练的池化层,两个分支网络联合训练,共同学习新闻内容潜在的语义特征。在每个分支子网络中,对新闻的标题和正文之间进行语义关联性度量。最终,对两个分支子网络联合训练后的结果进行决策融合,输出模型的预测结果。

1 相关工作

1.1 基于内容的虚假新闻检测

基于内容的虚假新闻检测方法指利用新闻的内容进行检测,包括文本信息(标题、正文、网页链接),视觉信息(图片、表情包),音频信息等。现有研究大多集中在新闻的文本内容上,从中提取统计特征或语义特征。Castillo等人[2]基于新闻文本内容的语言特征来检测虚假新闻,他们设计了一份语言特征列表,如问号、表情符号、情绪正负词和代词,来衡量推特上信息的可信度。Popat[3]发现,文章的语言风格对于理解文章的可信度起着至关重要的作用。然而,基于语言风格的特征不具备语义信息,很容易被操纵。Feng等人在文献[4]中使用上下文无关的语法规则识别虚假信息。Ma等人[5]首先探索了通过捕获时间语言特征来用深度神经网络表示新闻的可能性。Chen等人[6]将注意力机制引入到循环神经网络中,以集中捕获独特的时间语言特征。随着多媒体技术的发展,虚假新闻试图利用带有图像或视频的多媒体内容来吸引和误导读者,以便迅速传播。Qi等人[7]从图像角度出发,通过探索图像物理层面和语义层面的不同特征,提出了一个多域视觉神经网络模型来融合频域和像素域的视觉信息,从而进行虚假新闻检测。该模型对不同数据集的泛化能力仍需进一步验证。Xue等人[8]进一步挖掘虚假图片的信息,对图片的像素域特征和频域特征进行融合后,作为视觉特征。同时,引入了图片的物理属性,最后通过集成学习联合视觉特征和物理特征,实现虚假新闻图片检测。

大部分基于内容的虚假新闻检测工作通常把新闻标题和正文作为一个整体来进行语义和风格特征的分析,较少工作直接从“标题党”的角度出发,即分析标题和正文之间存在的差异和关联性。虽然已经有针对“标题党”检测任务的研究工作[9],但该工作的重点在于识别新闻是否存在“标题党”的现象。因此,基于“标题党”检测的思想,本文重点探索新闻标题和正文之间的语义关联性,利用最大均值差异(maximum mean discrepancy,MMD)[10]进行度量。结合深度神经网络和不同的池化操作进行联合训练,自动提取文本的潜在特征,以检测新闻的真实性。

1.2 基于社交上下文的虚假新闻检测

基于社交上下文的虚假新闻检测方法通过探索与新闻相关的社交上下文信息来检测虚假新闻,即新闻在社交媒体上的传播方式以及用户的参与情况等。社交用户和新闻之间的互动所建立的社会联系,为新闻提供了丰富的社交上下文信息。社交上下文信息代表了用户在社交媒体上对新闻的参与情况[11],例如关注者数量、评论、点赞、话题标签和分享转发的网络结构。Wu

等人[12]利用社交媒体上的用户资料和新闻传播路径来分类虚假新闻。Liu等人[13]将新闻的传播路径作为多元时间序列来建模，结合RNNs和CNNs网络来检测虚假新闻。然而，在虚假新闻的早期检测阶段，即新闻在新闻渠道上发布但尚未在社交媒体上传播时，不能依靠新闻的传播信息，因为它们并不存在[14]。Ma等人[15]基于树状结构的递归神经模型来学习推文的表示。Jin等人[16]使用了手工提取的关注者数量、转发量等社交上下文特征。尹鹏博等人[17]通过对用户的历史微博进行分析，结合用户属性和微博文本，采用C-LSTM模型实现谣言检测。沈瑞琳等人[18]提出基于多任务学习的微博谣言检测方法，利用情感分析任务辅助谣言检测，在一定程度上解决了深度学习中带标签数据不足的问题，但模型对相关的辅助数据具有依赖性。

社交上下文信息通常是非结构化数据，需要通过大量的手工劳动来收集。同时，社交上下文特征需要经过一段时间的积累才能提取出来，不能及时检测新出现的虚假新闻。在新闻还没在社交媒体上传播开来之前，需要使用基于内容的检测方法，因为在这个阶段还不存在丰富的社交上下文信息。因此，本文基于新闻内容本身，通过挖掘潜在的信息来进行虚假新闻检测。

2　方法

本文提出的基于双分支网络联合训练的虚假新闻检测模型结构如图1所示，模型由两个分支子网络组成，分别是MPB和GPB。每个分支子网络包含了三个模块：(1)文本特性提取器；(2)标题正文间语义关联性度量；(3)虚假新闻分类器。首先，文本特性提取器分别提取新闻文章的标题特征和正文特征，并使用MMD来度量它们之间的语义关联性，然后将两个特征进行加权融合，作为虚假新闻分类器的输入。最后，对两个分支子网络联合训练的分类结果进行决策融合，输出模型的预测结果(真实或虚假)。MPB采用最大池化进行下采样，GPB采用广义均值池化进行下采样。

2.1　文本特征提取

给定一篇由标题 T(title)和正文 B(body text)组成的新闻文章 $A=\{T,B\}$，在不同的分支子网络中，采用不用的文本特征提取方法。在MPB中，本文使用卷积神经网络Text-CNN[19]来学习新闻的特征表示。Text-CNN利用不同窗口大小的多个卷积核来捕获文本的特征信息。对于标题 T 中的每个字，对应的 d 维词嵌入向量表示为 $\boldsymbol{x}_t^l\in\mathrm{R}^d, l=1,2,\cdots,n_t$。本文使用下标 t 来标识标题 T，使用下标 b 来标识正文 B。新闻标题的词嵌入向量序列可表示为：

$$T_{1:n_t}=x_1\oplus x_2\oplus\cdots\oplus x_{n_t} \tag{1}$$

其中，$T_{1:n_t}\in\mathrm{R}^{n_t\times d}$，$\oplus$ 表示拼接操作，n_t 为新闻标题的长度。窗口大小为 h 的卷积核以标题中 h 个词的连续序列 $\{\boldsymbol{x}_t^{i:(i+h-1)}\}_{i=1}^{n-h+1}$ 作为输入，对其进行卷积操作，输出特征映射 $C_t=\{c_t^i\}_{i=1}^{n-h+1}$。以从第 i 个字开始的连续序列为例，卷积操作可表示为式(2)：

$$c_t^i=\sigma(\boldsymbol{w}_t\cdot\boldsymbol{x}_t^{i:(i+h-1)}+b_t) \tag{2}$$

$$\boldsymbol{x}_{i:(i+h-1)}=\boldsymbol{x}_i\oplus\boldsymbol{x}_{i+1}\oplus\cdots\oplus\boldsymbol{x}_{i+h-1} \tag{3}$$

其中，$\boldsymbol{x}_t^{i:(i+h-1)}\in\mathrm{R}^{h\times d}$，$\oplus$ 表示拼接操作，w_t 为卷积核，b_t 为偏置项，$\sigma(\cdot)$ 是ReLU激活函数。对卷积操作后得到的特征映射进行最大池化操作，实现降维。池化层对特征映射 c_t^i 取最大值，从中提取出最重要的信息。每

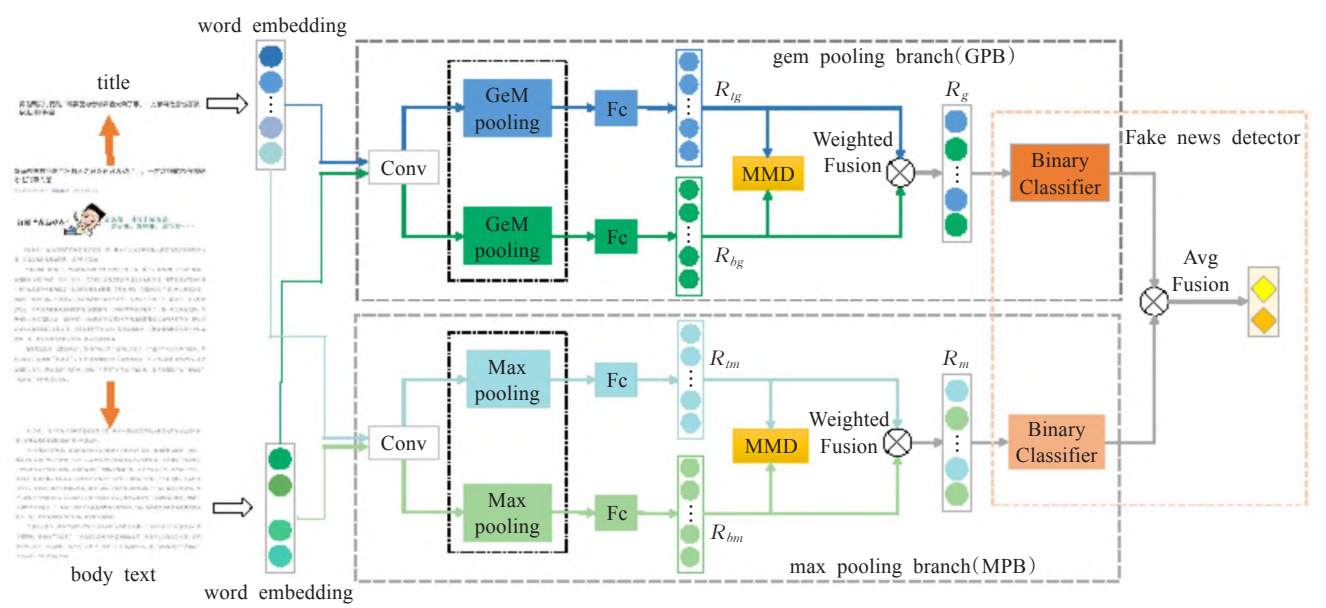


图1　JTTN模型

Fig.1　JTTN model

个特征映射经过最大池化后,可表示为:

$$\hat{C}_{tm}=\max\left\{c_t^i\right\}_{i=1}^{n-h+1} \tag{4}$$

最后,将池化后的结果输入全连接层,得到标题的特征表示为:

$$R_{tm}=W_{tm}\hat{C}_{tm}+b_{tm} \tag{5}$$

其中,R_{tm} 的下标 tm 表示标题特征通过MPB子网络获得,W_{tm} 表示权重矩阵,$\hat{C}_{tm}\in\mathrm{R}^k$,$k$ 表示不同窗口大小的卷积核数目。

类似地,对于长度为 n_b 的新闻正文 B,经过 d 维词嵌入后,可表示为:

$$B_{1:n_b}=x_1\oplus x_2\oplus\cdots\oplus x_{n_b} \tag{6}$$

采用跟上述新闻标题相同的特征提取方式,新闻正文特征可表示为:

$$R_{bm}=W_{bm}\hat{C}_{bm}+b_{bm} \tag{7}$$

Text-CNN的池化层采用最大池化操作,在减少模型参数量的同时能保证特征的位置和旋转不变性,但是忽略了文本特征的位置信息。Radenović等人在文献[20]中提出了一种可训练的广义均值池化层(GeM pooling layer),并证明其能够显著提高检索性能。广义均值池化介于最大池化和均值池化之间,二者是其特殊形式。

因此,在GPB子网络中,基于Text-CNN的网络结构,采用广义均值池化代替原来的最大池化方式,来捕获不同粒度的特征信息。对于公式(2)得到的每个特征映射 c_t^i,分别进行广义均值池化操作。计算公式可表示为:

$$\hat{C}_{tg}=\left\{f_{tg}^i\right\}_{i=1}^{n-h+1} \tag{8}$$

$$f_{tg}^i=\left(\frac{1}{\left|c_t^i\right|}\sum_{x\in c_t^i}x^{p_i}\right)^{\frac{1}{p_i}} \tag{9}$$

当 $p_i=1$ 时,广义均值池化相当于均值池化,当 $p_i\to\infty$ 时,广义均值池化相当于最大池化。相比于最大池化,广义均值池化包含可学习的参数 p_i,对输入的样本先求 p 次幂,然后取均值,再进行 p 次开方。

将池化后的结果输入到全连接层,得到新闻标题的特征表示为:

$$R_{tg}=W_{tg}\hat{C}_{tg}+b_{tg} \tag{10}$$

其中,R_{tg} 的下标 tg 表示标题的特征表示通过GPB子网络获得,W_{tg} 为权重矩阵,b_{tg} 为偏置项。

类似地,对于新闻正文 B,通过GPB子网络获得的特征表示为:

$$R_{bg}=W_{bg}\hat{C}_{bg}+b_{bg} \tag{11}$$

2.2 标题正文间语义关联性度量

一篇完整的新闻通常由标题(短文本)T 和正文(长文本)B 组成。受到“标题党”检测任务的启发,发现虚假新闻发布者为了吸引更多读者阅读和传播虚假信息,通常会使用夸大、猎奇、色情的标题来吸引眼球,新闻的正文内容往往与标题不匹配。但仅仅检测“标题党”还不够,因为一些真实新闻也会存在“标题党”现象。因此,在上述文本特征提取过程中,使用两个分支网络,充分挖掘新闻的语义信息。接下来,本文使用最大均值差异来度量新闻标题和正文之间的语义关联性。最大均值差异是迁移学习,尤其是域适应中使用最广泛的一种损失函数,主要用来度量在再生希尔伯特空间中两个分布的距离。

假设一篇新闻的标题和正文来自于两个文本语义分布,分别表示为 X_T 和 X_B。如果标题跟正文描述同一件事情,在语义上相关,则认为它们所在的分布相同,该新闻倾向于是真实新闻。反之,该新闻倾向于是虚假新闻。本文使用MMD来度量标题和正文两个分布间的距离,距离定义为:

$$MMD(X_T,X_B)=\left\|\frac{1}{|X_T|}\sum_{x_t\in X_T}\phi(x_t)-\frac{1}{|X_B|}\sum_{x_b\in X_B}\phi(x_b)\right\| \tag{12}$$

其中,$\sigma(\cdot)$ 表示映射函数,用于把原变量映射到再生希尔伯特空间。如果一篇新闻是虚假新闻,则它的标题和正文之间的MMD距离要比真实新闻大,关联性更小。本文目的在于最大化虚假新闻的标题和正文之间的MMD距离。如果这个值足够小,就认为两个分布相同,否则就认为它们不相同。MMD距离损失函数可以表示为:

$$\zeta_{\mathrm{mmd}}(\theta_T,\theta_B)=MMD^2(X_T,X_B) \tag{13}$$

其中,$\theta_T=\{\theta_{tm},\theta_{tg}\}$ 表示新闻标题特征提取过程中所需参数,$\theta_B=\{\theta_{bm},\theta_{bg}\}$ 表示新闻正文特征提取过程中所需参数。

2.3 虚假新闻分类器

到目前为止,通过文本特征提取器,分别获得新闻标题和正文的特征表示。在MPB中,标题 T 的特征表示为 R_{tm},正文 B 的特征表示为 R_{bm}。在GPB中,标题 T 的特征表示为 R_{tg},正文 B 的特征表示为 R_{bg}。在每个分支网络中,分别对标题特征和正文特征进行加权融合,融合后的特征作为虚假新闻检测器的输入,然后连接含Softmax函数的全连接层来预测新闻的真假。虚假新闻分类器可表示为 $L_d(\cdot;\theta_d)$,θ_d 表示分类器中的所有参数。对于第 i 篇新闻 a_i,虚假新闻检测器的最终输出记为 $p_\theta(a_i)$,表示该新闻是虚假新闻的概率。

$$p_\theta(a_i)=L_d(R_m^i,R_g^i;\theta_d) \tag{14}$$

$$R_m=\lambda_1R_{tm}+\lambda_2R_{bm} \tag{15}$$

$$\boldsymbol{R}_g=\lambda_3\boldsymbol{R}_{tg}+\lambda_4\boldsymbol{R}_{bg} \tag{16}$$

其中，$\boldsymbol{R}_m$、$\boldsymbol{R}_g$ 分别表示一篇文章在MPB和GPB中融合后的特征。λ_1、λ_2、λ_3、λ_4 分别表示加权权重。虚假新闻检测器的目的在于识别某一篇文章是否是虚假新闻。用 Y 表示新闻文章集合 A 的真实标签集合，使用交叉熵损失函数作为虚假新闻检测器的分类损失：

$$\zeta_{\text{class}}(\theta_d)=$$
$$-E_{(a_i,y)\sim(A,Y)}\left[y\log(p_\theta(a_i))+(1-y)\log(1-p_\theta(a_i))\right] \tag{17}$$

其中，a_i 表示某一篇文章，y 表示该文章对应的真实标签。目的在于寻找最优的参数 $\hat{\theta}_d$ 来最小化分类损失，这个过程可以表示为：

$$\hat{\theta}_d=\arg\min_{\theta_d}\zeta_{\text{class}}(\theta_d) \tag{18}$$

2.4　双分支联合训练

为了从不同范围和粒度捕获新闻文章的文本信息，采用了双分支网络联合训练的方法，分别为MPB和GPB。在每个分支网络里，基于Text-CNN和不同的池化方式，提取新闻的标题和正文特征，然后利用MMD度量标题和正文之间的语义距离并约束两个分布的特征。最后，两个分支网络联合训练，输出虚假新闻检测任务的预测结果。这样做的目的在于，(1)检测虚假新闻；(2)充分探索新闻的标题和正文之间的语义关联性。模型最终的损失函数可以表示为：

$$\zeta_{\text{final}}(\theta_{tm},\theta_{bm},\theta_{tg},\theta_{bg},\theta_d)=$$
$$\alpha\zeta_{\text{class}}(\theta_d)-\beta(\zeta_{mmd_m}(\theta_{tm},\theta_{bm})+\zeta_{mmd_g}(\theta_{tg},\theta_{bg})) \tag{19}$$

其中，$\zeta_{\text{class}}(\cdot)$ 表示交叉熵分类损失。$\zeta_{mmd_m}(\cdot)$ 表示在MPB中，标题和正文间的语义关联损失。θ_{tm}、θ_{bm} 分别表示在MPB中标题和正文特征提取过程中所需要的参数。类似地，$\zeta_{mmd_g}(\cdot)$ 表示在GPB中标题和正文之间的语义关联损失。θ_{tg}、θ_{bg} 分别表示在GPB中标题和正文特征提取过程中所需要的参数。目的在于最小化最终的损失函数，该过程可以表示为：

$$(\hat{\theta}_{tm},\hat{\theta}_{bm},\hat{\theta}_{tg},\hat{\theta}_{bg},\hat{\theta}_d)=$$
$$\arg\min_{\theta_{tm},\theta_{bm},\theta_{tg},\theta_{bg},\theta_d}\zeta_{\text{final}}(\theta_{tm},\theta_{bm},\theta_{tg},\theta_{bg},\theta_d) \tag{20}$$

其中，θ_{tm}、θ_{bm}、θ_{tg}、θ_{bg}、θ_d 表示MPB子网络、GPB子网络以及分类器中包含的参数，例如卷积核、权重矩阵、偏置项等。通过反向传播算法对上述参数进行更新，其优化过程见算法1。每轮训练都采用Adam优化器，通过自适应调整学习率来优化网络的收敛速度。在网络训练过程中，采用Early Stop策略，当模型的性能无明显变化时，停止训练。

算法1　JTTN

输入：新闻文章 $A=\{(T_i,B_i)\}_{i=1}^N$，新闻标签 $Y=\{y_i\}_{i=1}^N$，学习率 η

输出：网络参数 $\theta_{tm},\theta_{bm},\theta_{tg},\theta_{bg},\theta_d$

1. 随机初始化网络参数：$\theta_{tm},\theta_{bm},\theta_{tg},\theta_{bg},\theta_d$

2. whilenot convergence do /*当网络未收敛时*/

3.　for each epochdo/*对于每一轮迭代，执行以下步骤*/

4.　for each mini-batch do/*对于每一个批处理，执行以下操作*/

5.　更新分类器参数：$\theta_d\leftarrow\theta_d-\eta\cdot\alpha\frac{\partial\zeta_{\text{class}}}{\partial\theta_d}$

6.　更新MPB子网络提取标题特征过程中所需的参数：$\theta_{tm}\leftarrow\theta_{tm}-\eta(\alpha\frac{\partial\zeta_{\text{class}}}{\partial\theta_{tm}}-\beta\frac{\partial\zeta_{mmd_m}}{\partial\theta_{tm}})$

7.　更新MPB子网络提取正文特征过程中所需的参数：$\theta_{bm}\leftarrow\theta_{bm}-\eta(\alpha\frac{\partial\zeta_{\text{class}}}{\partial\theta_{bm}}-\beta\frac{\partial\zeta_{mmd_m}}{\partial\theta_{bm}})$

8.　更新GPB子网络提取标题特征过程中所需的参数：$\theta_{tg}\leftarrow\theta_{tg}-\eta(\alpha\frac{\partial\zeta_{\text{class}}}{\partial\theta_{tg}}-\beta\frac{\partial\zeta_{mmd_g}}{\partial\theta_{tg}})$

9.　更新GPB子网络提取正文特征过程中所需的参数：$\theta_{bg}\leftarrow\theta_{bg}-\eta(\alpha\frac{\partial\zeta_{\text{class}}}{\partial\theta_{bg}}-\beta\frac{\partial\zeta_{mmd_g}}{\partial\theta_{bg}})$

10.　end for

11.　end for

12. end

13. 返回网络参数：$\theta_{tm},\theta_{bm},\theta_{tg},\theta_{bg},\theta_d$

3　实验与结果分析

3.1　数据集

为评估本文所提出的模型性能，研究采用Wang等人公开的新闻数据集[21]进行实验。该数据集(https://github.com/yaqingwang/WeFEND- AAAI20)收集了从2018年3月到2018年10月，微信公众号发布的新闻文章。公开的新闻数据集包含了微信公众号名称(新闻发布者)、新闻标题、新闻链接、新闻封面链接、用户反馈报告以及新闻标签(fake or real)这六项信息。为了能够探索新闻标题和正文之间的语义关联性，从而进行虚假新闻检测，在该数据集的基础上，做进一步的信息收集和数据清洗。根据数据集公开的新闻链接和封面链接，通过网络爬虫技术爬取了每一篇新闻对应的文章正文，封面图片以及文章内部的图片。由于受到微信营运平台的监管和读者的反馈举报，很多新闻都已经失效，特别是虚假新闻。通常情况是新闻文章被删除或者公众号被封号，这导致不能爬取到所有完整的数据。因此，对于已经失效的新闻文章，只保留它们的标题信息。最终得到的数据统计信息如表1所示。本文使用新闻的标题和正文数据作为模型的输入。

表1 新闻数据集统计信息

Table 1 Statistics of news dataset

统计项	虚假新闻	真实新闻	总计
新闻文章	4 225	16 503	20 728
标题	4 225	16 503	20 728
正文	918	8 011	8 929
图片	10 068	118 115	128 183

3.2 对比实验

为了验证本文方法的有效性,选取了目前虚假新闻检测任务常用的方法作为基线方法进行对比。

(1)CNN_T:CNN_T只使用新闻标题作为输入,由于缺乏正文,所以在JTTN模型的基础上去掉标题和正文之间的语义关联性度量,然后使用双分支网络进行特征提取,再通过分类器进行二分类。

(2)CNN_B:CNN_B只使用新闻正文作为输入,其余设置与CNN_T相同。

(3)LSTM:LSTM使用一层LSTM作为文本特征提取器,通过对RNN在每个时间步长的输出进行平均,得到潜在表示,然后将这些表示输入到全连接层进行预测。建立在LSTM特征提取器之上的全连接层输出新闻是虚假新闻的概率。

(4)HAN[22]:HAN基于新闻的内容,构建了一个层次注意力神经网络框架来进行虚假新闻检测。它对新闻内容进行编码,采用“词-句子-文章”的层次化结构来表示一篇文章,在句子级别关注词层次,在文档级别关注句层次。

(5)Att-RNN[16]:Att-RNN利用注意机制来融合文本、视觉和社会上下文特征。实验中,去除视觉和社会上下文信息,其余部分相同。

(6)EANN[23]:EANN由三个主要部分组成,多模态特征提取器、虚假新闻检测器和事件鉴别器。多模态特征提取器从帖子中提取文本和视觉信息,它与虚假新闻检测器一起学习可识别的特征表示来检测虚假新闻。事件鉴别器负责删除任何特定于事件的特征。由于输入只有文本信息,所以去除了视觉特征提取器和事件鉴别器。

(7)SAFE[24]:SAFE基于Text-CNN提取新闻文本特征,通过计算新闻文章文本和视觉信息之间的相似性,实现虚假新闻检测。该模型以完整的新闻文章作为输入,设置了与文献[24]相同的超参数。

3.3 评估方法与参数设置

本文使用准确率(Accuracy)、精确度(Precision)、召回率(Recall)、F1值(F1-score)作为评估指标。通常F1值越高,说明分类器性能越好。实验采用PyTorch深度学习框架来构建虚假新闻检测模型并进行模型训练。根据新闻的发布日期,按照7∶1∶2的比例划分数据集,70%作为训练集、10%作为验证集、20%作为测试集。其中,最新发布的新闻文章作为测试数据。在参数设置方面,新闻标题长度n_t设为32,正文长度n_b设为300,不足部分用0填充,超出部分删除。标题和正文的嵌入维度d均设为300,最后加权融合后的特征维度为128维。Text-CNN有三种卷积核,大小分别为2、3、4,每种卷积核的数量为200。在网络训练过程中,使用Adam优化器,设置批处理大小为256,迭代次数为200,学习率为1E−3。MMD中的映射函数$\phi(\cdot)$为高斯核函数。

3.4 实验结果及分析

表2显示了本文方法跟其他方法的实验对比结果。实验结果表明,针对虚假新闻检测任务,本文提出的方法在准确率、精确度和F1值上均优于其他方法,取得了最好的分类性能。针对实验结果,有以下几点分析:(1)从CNN_T和CNN_B的实验结果可以看出,将新闻标题和正文一起作为模型的输入来检测虚假新闻,其性能优于仅使用标题或者正文作为模型的输入。由此可验证新闻标题正文间语义关联性度量的有效性。(2)HAN采用了词层级和句层级的注意力机制,目的在于提取出文章中贡献最大的词和句子。这种解决方法,对文本分类能起到很好的效果,但不适用于虚假新闻检测,因为虚假新闻也是围绕一个主题展开描述。仅仅依靠文章最重要的信息,无法有效地检测虚假新闻,导致虚假新闻预测结果的F1分数偏低。(3)LSTM擅长处理时序信息,在文本任务中,它能够更好地联系上下文信息提取特征,但虚假新闻检测任务更注重语义风格等的局部特征,对时序特征不会过分依赖,因此使用Text-CNN进行特征提取的EANN模型能够更好地提取文本的局部特征,在虚假新闻检测任务上表现得更好。(4)SAFE通过引入额外的全连接层来扩展Text-CNN,以自动提取每篇新闻文章的文本特征。与之不同的是,本文方法引入了可训练的池化层,通过训练网络自动调节参数,进一步学习新闻潜在的文本特征,故其总体性能优于

表2 JTTN模型与其他方法的实验结果对比

Table 2 Comparison of experimental results between JTTN model and other methods

方法	Accuracy	Precision	Recall	F1-score
CNN_T	0.981	0.936	0.875	0.905
CNN_B	0.986	0.944	0.913	0.928
LSTM	0.961	0.852	0.750	0.798
HAN	0.903	0.813	0.779	0.796
Att-RNN	0.953	0.891	0.620	0.731
EANN	0.977	**0.955**	0.810	0.877
SAFE	0.985	0.944	0.908	0.925
JTTN	**0.988**	0.931	**0.951**	**0.941**

SAFE。(5)本文的模型使用双分支网络进行联合训练，能够充分地挖掘新闻文章潜在的语义风格特征，从而捕获虚假新闻与真实新闻的差异。另外，基于“标题党”检测的思想，通过度量新闻标题和正文之间的语义关联性，能更好地检测出虚假新闻。

3.5　不同关联性度量方法对比

为了分析不同关联性度量方法对实验结果的影响，共设计了4种变体：(1)去掉标题和正文之间的语义关联性度量(-MMD)。(2)使用CORAL[25]作为度量方法(CORAL)。(3)使用余弦相似度作为度量方法(COS)。(4)使用最大均值差异作为度量方法，即本文提出的方法(MMD)。实验结果如表3所示，结果表明，在4种变体中，使用最大均值差异作为度量方法的实验结果最好，使用余弦相似度作为度量方法的效果次之。实验结果也表明新闻标题和正文之间的语义关联性度量对虚假新闻检测任务的有效性。

表3　不同关联性度量方法的实验结果

Table 3　Results of different correlation measurement methods

度量方法	Accuracy	Precision	Recall	F1-score
-MMD	0.974	0.904	0.891	0.897
CORAL	0.982	0.923	0.908	0.915
COS	0.984	0.933	0.909	0.921
MMD	**0.988**	0.931	**0.951**	**0.941**

最大均值差异比余弦相似度表现更优的原因在于：余弦相似度假设在语义特征空间中，两个特征向量对应位置的元素特征是对齐的，但这种假设过于严格，在异构源向量中通常是无效的。而最大均值差异是将两个特征向量映射到再生希尔伯特空间中，通过核学习方法，来度量两个分布之间的距离，并不要求两个特征向量间的元素特征对齐，更适用于度量标题和正文间的语义关联性。

3.6　单分支网络与双分支网络实验结果对比

为了探究双分支网络联合训练模型是否比单分支网络训练更有效，本文设计了单分支网络与双分支网络的对比实验。以本文设计的JTTN模型为基础，分别去掉其中的一个分支，作为单分支网络。实验结果如图2所示，其中，MPB、GPB分支表示单分支网络，JTTN表示双分支网络。

从图2的结果可以看出，双分支网络的准确率和F1值均比单分支网络高。双分支网络的F1值分别比MPB和GPB高出了0.016和0.015。证明了双分支网络联合训练比单分支网络单独训练效果更好。

3.7　参数分析

在损失函数计算公式(19)中，α和β被用来权衡交叉熵分类损失(α)和语义关联损失(β)之间的相对重

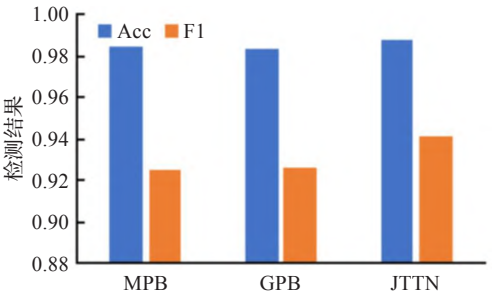


图2　单分支网络与双分支网络实验结果对比

Fig.2　Comparison of single-branch network and two-branch network experimental results

要性。为了评估α和β对模型性能的影响，设计了相关实验，分别设置α和β的值从0递增到1，步长设置为0.2。在α和β不同的取值下，模型的检测结果(准确率和F1值)如图3所示。可以看出，相比于α，不同的β值对模型性能的影响较为显著。当β的取值较大时，模型的准确率和F1值较高，分类器效果较好。由此，可以验证模型中标题正文间语义关联性度量的可行性和有效性。图3(a)中，准确率的变化范围从0.982到0.988，α和β的不同取值对准确率的影响不明显。图3(b)中，F1值的变化范围从0.91到0.95，相差了0.04。从实验结果可知，当$\alpha=0.2;\beta=0.4$或者$\alpha=0.4;\beta=1$时，也就是说，当$\alpha:\beta\approx1:2.3$时，模型取得最好的效果。

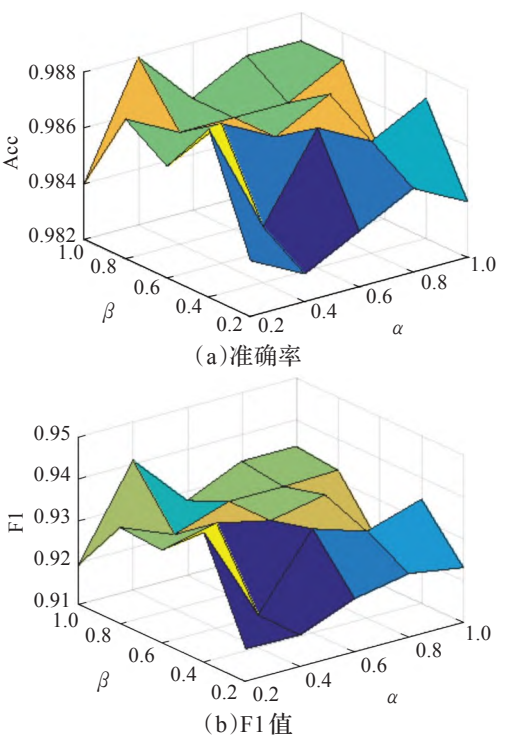


图3　损失函数参数分析

Fig.3　Parameter analysis of loss function

3.8　收敛性分析

图4展示了本文提出的模型在训练过程中，最终的损失函数值(loss)随迭代次数(epoch)变化的情况。网

络经过约20次迭代训练后,逐渐收敛到相对平稳的趋势。由此可以验证本文提出的模型的有效性以及损失函数计算的可行性。

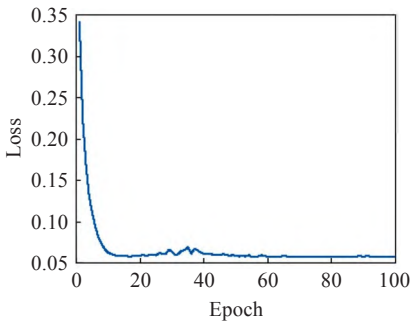
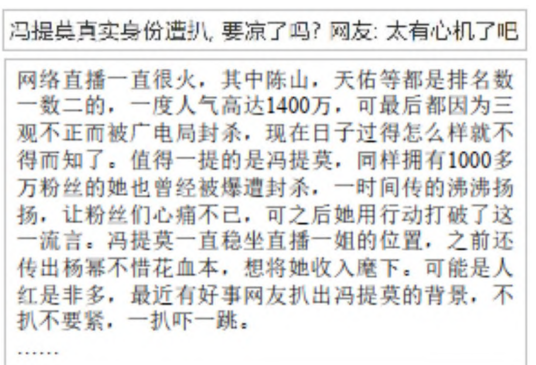


图4　损失函数

Fig.4　Loss function

3.9　案例分析

对于模型分类错误的新闻,找出具有代表性的例子进行分析,探究分类错误的原因,如图5所示。图5(a)表示真实新闻被预测为虚假新闻的例子,从文章内容可以看出,它的标题使用了问号,且引用网友的话,让读者迫切地想知道文章主角的真实身份到底是什么。很明显,这符合“标题党”的现象。文章正文前半部分对标题提出的人物身份进行描述,但后半部分,却转向描述别的人物,偏离了标题。基于上述两点,本文模型把它识别成了虚假新闻。图5(b)表示虚假新闻被预测为真实新闻的例子,文章标题表明已找到“马航MH370”失联


冯提莫真实身份遭扒, 要凉了吗? 网友: 太有心机了吧

网络直播一直很火，其中陈山，天佑等都是排名数一数二的，一度人气高达1400万，可最后都因为三观不正而被广电局封杀，现在日子过得怎么样就不得而知了。值得一提的是冯提莫，同样拥有1000多万粉丝的她也曾经被爆遭封杀，一时间传的沸沸扬扬，让粉丝们心痛不已，可之后她用行动打破了这一流言。冯提莫一直稳坐直播一姐的位置，之前还传出杨幂不惜花血本，想将她收入麾下。可能是人红是非多，最近有好事网友扒出冯提莫的背景，不扒不要紧，一扒吓一跳。

……

(a)真实新闻被预测为虚假新闻的例子


马航MH370.....找到了

2014年3月8日01:20，马航MH370在马来西亚和越南的交接处与胡志明管控区失去联系，飞机上239名乘客和机组人员至今下落不明。

2014年3月15日，网友@Kathleen_琳琳 发微博称“马航失事地点在柬埔寨森林中”。从评论看来，当时的网友显然对该说法不抱信任，并指责博主造谣，随后该条微博也被新浪微博定性为“不实信息”。

万万没想到的是，在4年多后的今天，在互联网几乎已经快要遗忘这架飞机的时候，英国技术专家Ian Wilson宣称在Google地图上发现了疑似马航MH370飞机的残骸，竟然位于东南亚柬埔寨的一个密林深处！

……

(b)虚假新闻被预测为真实新闻的例子

图5　识别错误的新闻例子

Fig.5　Examples of wrong results

飞机,正文部分也举例证明标题的说法,很难区分真假,所以本文模型将其预测为真实新闻。

4　结束语

文本所提出的基于双分支网络联合训练的虚假新闻检测方法,通过采用双分支网络结构来挖掘新闻标题和正文潜在的语义特征,同时,度量标题和正文之间的语义关联性,实现虚假新闻的早期检测。本文模型取得了较好的性能,准确率、F1值分别高达0.988、0.941。实验结果表明,基于双分支网络进行联合训练的方法具有可行性和有效性。目前本文仅使用新闻的文本类型(单模态)作为模型的输入,未来的工作将集中在增加模型的输入数据类型,采用更多的社交媒体信息,如图片、视频等作为模型的输入,实现多模态虚假新闻检测。

参考文献:

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

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

[3] POPAT K.Assessing the credibility of claims on the web[C]//Proceedings of the 26th International Conference on World Wide Web Companion,2017:735-739.

[4] FENG S,BANERJEE R,CHOI Y.Syntactic stylometry for deception detection[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2:Short Papers),2012:171-175.

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

[6] 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.Cham:Springer,2018:40-52.

[7] QI P,CAO J,YANG T,et al.Exploiting multi-domain visual information for fake news detection[C]//2019 IEEE International Conference on Data Mining(ICDM),2019:518-527.

[8] XUE J,WANG Y,XU S,et al.Mvfnn:multi-vision fusion neural network for fake news picture detection[C]//International Conference on Computer Animation and Social Agents.Cham:Springer,2020:112-119.

[9] BOURGONJE P,SCHNEIDER J M,REHM G.From clickbait to fake news detection:an approach based on detecting the stance of headlines to articles[C]//Pro-

ceedings of the 2017 EMNLP Workshop:Natural Language Processing Meets Journalism,2017:84-89.

[10] BORGWARDT K M,GRETTON A,RASCH M J,et al. Integrating structured biological data by kernel maximum mean discrepancy[J].Bioinformatics,2006,22(14):49-57.

[11] 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.

[12] WU L,LIU H.Tracing fake-news footprints:characterizing social media messages by how they propagate[C]// Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining,2018:637-645.

[13] LIU Y,WU Y F B.Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]//Thirty-Second AAAI Conference on Artificial Intelligence,2018.

[14] ZHOU X,JAIN A,PHOHA V V,et al.Fake news early detection:a theory-driven model[J].Digital Threats:Research and Practice,2020,1(2):1-25.

[15] MA J,GAO W,WONG K F.Rumor detection on twitter with tree-structured recursive neural networks[C]// Association for Computational Linguistics,2018.

[16] 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,2017:795-816.

[17] 尹鹏博,潘伟民,彭成,等.基于用户特征分析的微博谣言早期检测研究[J].情报杂志,2020,39(7):81-86.
YIN P B,PAN W M,PENG C,et al.Research on early detection of Weibo rumors based on user characteristics analysis[J].Journal of Intelligence,2020,39(7):81-86.

[18] 沈瑞琳,潘伟民,彭成,等.基于多任务学习的微博谣言检测方法[J].计算机工程与应用,2021,57(24):192-197.
SHEN R L,PAN W M,PENG C,et al.Microblog rumor detection method based on multi-task learning[J].Computer Engineering and Applications,2021,57(24):192-197.

[19] KIM Y.Convolutional neural networks for sentence classification[J].arXiv:1408.5882,2014.

[20] RADENOVIĆ F,TOLIAS G,CHUM O.Fine-tuning CNN image retrieval with no human annotation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,41(7):1655-1668.

[21] WANG Y,YANG W,MA F,et al.Weak supervision for fake news detection via reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2020:516-523.

[22] YANG Z,YANG D,DYER C,et al.Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2016:1480-1489.

[23] 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,2018:849-857.

[24] ZHOU X,WU J,ZAFARANI R.Safe:similarity-aware multi-modal fake news detection[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham:Springer,2020:354-367.

[25] SUN B,SAENKO K.Deep coral:correlation alignment for deep domain adaptation[C]//European Conference on Computer Vision.Cham:Springer,2016:443-450.