CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 44 卷第 9 期 2022 年 9 月
Vol. 44, No. 9, Sep. 2022

文章编号:1007-130X(2022)09-1702-09

融合评论的多任务联合谣言检测方法*

王　繁[1,2],郭军军[1],余正涛[1,2]

(1.昆明理工大学信息工程与自动化学院,云南 昆明 650500;
2.昆明理工大学云南省人工智能重点实验室,云南 昆明 650500)

摘　要:目前,针对微博领域的谣言检测方法主要基于微博正文,同时辅以用户评论特征、传播特征等信息进行判定。然而已有方法没有考虑用户评论质量会直接影响谣言检测的性能,质量低的评论甚至会引入无用甚至负面的特征,进而对谣言检测的性能带来更大的影响。针对该问题,基于用户评论和谣言检测的关联性,首次提出一种考虑评论有效性,并基于多任务联合学习的谣言检测方法。首先将谣言检测作为主任务,用户评论相关性检测为辅助任务;然后采用门控机制和注意力机制过滤和选择有效的用户评论特征;最后基于自主构建的 3 万条疫情微博谣言数据集进行实验。实验结果表明,对用户评论进行筛选不仅可以提升谣言检测性能,还能对用户评论质量进行判定。

关键词:谣言检测;联合学习;用户评论;评论有效性

中图分类号:TP391　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2022.09.021

A multi-task joint rumor detection method combining comments

WANG Fan[1,2], GUO Jun-jun[1], YU Zheng-tao[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500;
2. Yunnan Key Laboratory of Aritficial Intelligence, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: At present, the rumor detection method for microblog field is mainly based on the microblog text itself, supplemented by information such as user comment characteristics and propagation characteristics. However, the current methods ignore the quality of user comments that may directly affect the performance of rumor detection and introduce useless or even negative features, exerting an impact on the performance of detection. In response to this problem, based on the relevance of user comments and rumor detection, a rumor detection algorithm that considers the effectiveness of comments is proposed. It considers the effectiveness of microblog comments while determining rumors, and rumor detection is implemented based on the the multi-task joint learning method. Firstly, rumor detection is taken as the main task, and user comment correlation detection is taken as the auxiliary task. Secondly, the gating mechanism and the attention mechanism are used to filter and select effective user comment features. Finally, experiments on the self-constructed dataset with 30,000 epidemic microblog rumors show that the screening of user comments can not only improve the performance of rumor detection, but also realize the judgment of the quality of user comments.

Key words: rumor detection; joint learning; user comment; comment validity

* **收稿日期**:2020-12-23;**修回日期**:2021-04-12
基金项目:国家重点研发计划(2018YFC0830101, 2018YFC0830100);国家自然科学基金(61866020, 61972186, 61762056, 61472168);云南省科技厅面上项目(2019FB082)
通信作者:郭军军(guojjgb@163.com)
通信地址:650500 云南省昆明市昆明理工大学信息工程与自动化学院
Address: Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, P. R. China

1　引言

微博用户量急剧增加,信息发布门槛低,网络运营平台缺乏及时、有效的监督机制,使得虚假信息、网络谣言等充斥网络。谣言借助微博独有的特点进行广泛传播,对社会、企业和个人都造成了极大的不良影响。基于微博数据的谣言检测,通过挖掘微博中的有效特征,开发准确的检测和干预技术有助于缓解谣言传播的负面影响。

谣言具有特殊性,为有意误导读者而撰写的,也可能掺杂着真实内容而导致文本特征不足,因此单从新闻内容很难辨别真假。如图1所示,深色用户评论信息具有来自社交媒体人群的丰富信息,包括观点、立场和情绪,对谣言的发现和甄别具有一定的指导意义;浅色用户评论信息对微博谣言判定并没有影响,有的甚至毫不相关,因此用户评论的质量不同对谣言的判定所起作用也不同。目前国内外研究人员针对谣言的检测主要通过探索新闻正文文本特征和用户社交环境实现。Ruchansky等人[1]使用混合的深度学习框架同时对新闻文本、用户响应和用户特征进行建模,为假新闻检测提供了全新的思路;Guo等人[2]利用神经网络对用户评论进行层次化建模,以检测用户的虚假评论;Wu等人[3]通过对抗网络从新闻内容的语义信息中捕获差异化的可信度特征,并将其融合以获取信息可信度评估。但是,这些方法对社交媒体数据之间的关联性信息利用不足,用户评论信息参差不齐,内含的噪声信息会对谣言检测带来影响。此外,建立多任务联合学习模型来训练2个任务是提高网络谣言检测效果的一种有效而新颖的方法。Kochkina等人[4]提出的方法模拟了2个任务之间的信息共享和表示强化,为每个任务扩展了有价值的特征;Wu等人[5]通过过滤共享特征并作用于特定任务,实现假新闻检测。然而典型的多任务学习方法中,共享特征未经筛选就平等地用到各任务中,导致一些无用特征干扰甚至误导检测。如何既考虑微博正文与用户评论之间的联系,又考虑它们之间的差别,同时过滤和选择用户评论中的关键特征以提高谣言检测准确率,是当前微博谣言检测任务亟待解决的难题之一。本文期望通过多任务联合学习的方式利用用户评论的有效特征指导模型进一步提升分类效果。

本文设计了一种带有过滤机制的多任务联合学习方法,从微博正文和用户评论的角度,通过引入共享特征过滤选择机制丢弃无效特征和选择有利特征来提升谣言检测的性能。此外,为了更好地捕获远程依赖关系并提高模型的并行度,本文还应用Transformer编码器模块[6]对2个任务的输入表示进行编码。实验结果表明,该方法的性能优于基线方法,在微博谣言检测中初步取得了较好的结果。本文的主要贡献如下:

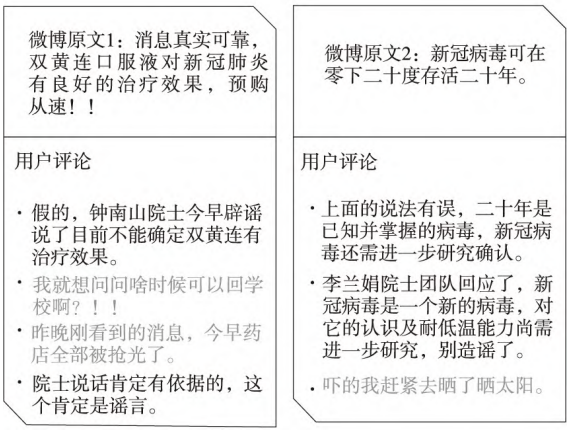

Figure 1　Relevance and difference between microblog content and user comments

图1　微博正文与用户评论的关联与差异

(1)提出一种多任务选择和信息过滤机制实现多任务融合,设计了一个融合用户评论筛选的多任务联合学习模型,并首次引入用户评论相关性检测作为辅助任务来改善最终检测性能。

(2)提出的模型通过门控机制和注意力机制来过滤和选择多任务间的共享特征流实现对用户评论的有效筛选,从而提升模型的检测效果。

基于3万条真实微博谣言检测数据集进行实验,对本文方法的性能进行全面评估。实验结果表明,本文方法对微博谣言检测是有效的。

2　相关研究

检测的目标是在早期或者使用可解释的因素有效地识别错误信息。谣言检测最直接的方法就是检查文本中主要内容的真实性,以判断事件的真实性。目前谣言检测方法大多是基于新闻内容和社交环境[7],包括文本特征、用户信息和用户响应等。

基于文本特征的方法旨在充分挖掘新闻内容特征,主要包括新闻文本、标题、图片和视频特征等。Potthast等人[8]探索了极端片面新闻与假新闻之间的写作风格,提出一种评估文本相似性的方法;Guo等人[9]认为由人群引发的新闻评论情绪(社会情绪)

在谣言检测中也起着重要作用,提出了一种双重情感特征框架来挖掘出版者情绪与社会情绪之间的关系。另外,典型的假新闻检测被认为是一个文本分类问题,探索潜在的文本层次结构[10-12]可能促进假新闻的检测。Karimi 等人[11]提出一种分层的文本层次结构来探讨真实新闻与假新闻之间的层次结构差异;Wang 等人[12]提供了一个新的、公开的假新闻数据集并设计了一种新的混合卷积神经网络来整合元数据和文本。随着计算机视觉 CV(Computer Vision)和自然语言处理 NLP(Natural Language Processing)领域的迅速发展,Abavisani 等人[13]提出了多模态融合方法,引入交叉注意力模块结合图像和文本信息实现检测任务,从视觉元素中提取视觉特征,以捕获假新闻的不同特征,可以很好地检测带有部分真实新闻内容的虚假新闻。

基于社交环境的方法旨在利用用户社交活动作为辅助信息来进行网络谣言检测。用户在社交媒体平台上的活动可以衍生出很多的社交语境特征,主要有基于用户的特征、基于网络的特征和基于响应的特征。可疑、低可信度用户的特征更多表现为:账号未经验证,账号创建时间较短,用户描述长度较短。从用户配置文件中提取用户特征[14,15]是一种假新闻检测的有效手段。Lu 等人[16]基于用户个人信息特征构建图感知共同注意网络来提升谣言检测性能。基于网络的特征是通过构造特定的网络来提取的,例如交互网络[17]和传播网络[18,19]。Shu 等人[17]利用出版社、新闻和用户之间的三元关系搭建交互网络实现假新闻分类;Monti 等人[18]提出了利用几何深度学习来学习假新闻中的特定传播模式的方法;Shu 等人[19]探索、验证了真、假新闻分层传播网络的结构、时间和语言特点。

基于响应的特征代表了用户的社会反应,包括立场和话题等。段大高等人[20]从微博评论的角度定义支持性、置信度和内容相关性 3 个特征来构建支持向量机算法判别消息真伪;Shu 等人[21]开发了文本评论联合注意网络,通过建立新闻句子和用户评论之间的相互影响来学习特征表示,并通过注意权重来学习句子和评论的可解释程度;Wu 等人[22]提出了一种自适应交融网络实现文本和评论之间的情感联想和语义冲突的交互融合,建立特征关联以提高谣言检测的性能。但是,以上方法都只注重对用户评论特征的挖掘及交互融合,忽略了用户评论质量对谣言检测也具有一定的影响,甚至会引入无用甚至负面的特征误导检测结果。

不同于上述已有方法,本文从评论信息有效利用的角度实现微博谣言检测,首次提出了一种融合评论的筛选多任务联合学习方法,融合用户评论的同时采用门控和注意力机制有效地过滤和选择用户评论特征,以提高微博谣言检测性能。

3 融合评论的联合学习谣言检测方法

3.1 检测方法框架

针对用户评论信息差异较大,评论质量影响谣言检测性能的问题,本文提出一个融合评论的多任

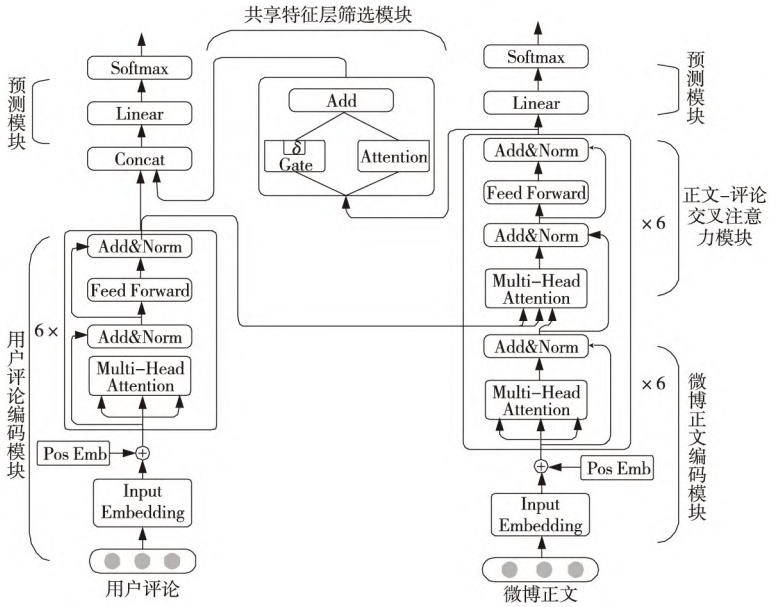

Figure 2 Framework of multi-task joint rumor detection method combined with comments

图 2 融合评论的多任务联合谣言检测方法框架

务联合学习方法 CMT-G&A(Comment Multi-task-Gate & Attention),其框架如图 2 所示。该方法主要包括 4 个模块,分别是微博正文-用户评论编码模块、正文-评论交叉注意力模块、共享特征层筛选模块和事件预测模块。

3.2　微博正文-用户评论编码模块

3.2.1　微博正文编码模块

微博正文编码模块用于提取微博正文的文本特征。设 E_1 为某一事件下的一条微博正文,每条正文长度为 l_1,$C=\{c_1,c_2,\cdots,c_N\}$ 是一组响应 E_1 的用户评论,每条用户评论长度为 l_2。本文使用 Transformer 编码模块[6]对微博正文特征嵌入进行编码。为了能够利用词在序列中的位置信息,在编码模块中将位置编码添加到词嵌入表征中,它与词嵌入表征具有相同的维数。编码模块核心是自注意力机制,具体如式(1)~式(3)所示:

$$\boldsymbol{E}_{\mathrm{Con}}=E(x_1,x_2,\cdots,x_n) \tag{1}$$

$$\boldsymbol{E}_{\mathrm{Con}}=\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V} \tag{2}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

其中,$\boldsymbol{E}_{\mathrm{Con}}$ 为微博正文输入文本的词嵌入表征;$x_1,x_2,\cdots,x_n$ 表示微博正文中的词;$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbf{R}^{l_1\times d}$ 分别为查询向量、键向量和值向量;d 表示微博正文和用户评论每个字(词)通过预训练搜狗新闻语料库[23]提取的词向量维度;$\sqrt{d_k}$ 为缩放因子。

相较于仅执行单一的注意力,本文使用不同的权重矩阵将输入信息投影到多个不同的向量空间(注意力头数),共同关注来自不同位置的不同表示子空间的信息是有益的,也即多头注意力。多头注意力通过不同的线性投影对 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 进行 h 次线性投影,然后对 h 次投影结果并行执行缩放点积注意计算,最后将这些注意结果串联起来再次获得新的表示。多头注意力可以使参数矩阵形成多个子空间,让矩阵学习多方面的信息。如式(4)和式(5)所示:

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{4}$$

$$\boldsymbol{H}_{\mathrm{Con}}=MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,head_2,\cdots,head_h)\boldsymbol{W}^{\mathrm{O}} \tag{5}$$

其中,$\boldsymbol{H}_{\mathrm{Con}}\in\mathbf{R}^{l_1\times d}$ 为微博正文编码模块的输出;$\boldsymbol{W}_i^{\boldsymbol{Q}}\in\mathbf{R}^{d\times d_k}$,$\boldsymbol{W}_i^{\boldsymbol{K}}\in\mathbf{R}^{d\times d_k}$,$\boldsymbol{W}_i^{\boldsymbol{V}}\in\mathbf{R}^{d\times d_k}$,$\boldsymbol{W}^{\mathrm{O}}\in\mathbf{R}^{d\times d}$ 为训练的参数;$d_k=d/h$。

3.2.2　用户评论编码模块

用户评论编码模块与微博正文编码模块相似,都采用 Transformer 编码模块[6]对用户评论特征嵌入进行编码,如式(6)~式(8)所示:

$$\boldsymbol{E}_{\mathrm{Com}}=E(x_1,x_2,\cdots,x_n) \tag{6}$$

$$\boldsymbol{E}_{\mathrm{Com}}=\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V} \tag{7}$$

$$\boldsymbol{H}_{\mathrm{Com}}=Transformer\_encoder(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \tag{8}$$

其中,$\boldsymbol{E}_{\mathrm{Com}}$ 为用户评论输入文本的词嵌入表征;$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbf{R}^{l_2\times d}$ 分别为查询向量、键向量和值向量;$\boldsymbol{H}_{\mathrm{Com}}\in\mathbf{R}^{l_2\times d}$ 为用户评论编码模块的输出。

3.3　正文-评论交叉注意力模块

用户评论包含一些针对微博正文的有用信息,能对谣言的检测起到促进作用。为了融合用户评论信息来促进谣言检测,本文仍采用 Transformer 编码模块[6]来提取正文-评论交叉注意力特征。不同之处在于该体系结构中,查询向量 $\boldsymbol{Q}$ 是微博正文编码模块的输出 $\boldsymbol{H}_{\mathrm{Con}}$,而键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$ 是用户评论编码模块的输出 $\boldsymbol{H}_{\mathrm{Com}}$,如式(9)和式(10)所示:

$$head_i=Attention(\boldsymbol{Q}_{\mathrm{Con}}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}_{\mathrm{Com}}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}_{\mathrm{Com}}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{9}$$

$$H_{\mathrm{Shared}}=MultiHead(\boldsymbol{Q}_{\mathrm{Con}},\boldsymbol{K}_{\mathrm{Com}},\boldsymbol{V}_{\mathrm{Com}})=Concat(head_1,head_2,\cdots,head_h)\boldsymbol{W}^{\mathrm{O}} \tag{10}$$

其中,$\boldsymbol{Q}_{\mathrm{Con}}\in\mathbf{R}^{l_1\times d}$,$\boldsymbol{K}_{\mathrm{Com}}\in\mathbf{R}^{l_2\times d}$,$\boldsymbol{V}_{\mathrm{Com}}\in\mathbf{R}^{l_2\times d}$;$\boldsymbol{H}_{\mathrm{Shared}}\in\mathbf{R}^{l_1\times d}$ 为微博正文和用户评论通过多头注意力所学的共同特征;$\boldsymbol{W}_i^{\boldsymbol{Q}}\in\mathbf{R}^{d\times d_k}$,$\boldsymbol{W}_i^{\boldsymbol{K}}\in\mathbf{R}^{d\times d_k}$,$\boldsymbol{W}_i^{\boldsymbol{V}}\in\mathbf{R}^{d\times d_k}$,$\boldsymbol{W}^{\mathrm{O}}\in\mathbf{R}^{d\times d}$ 为训练的参数;$d_k=d/h$。

3.4　共享特征层筛选模块

用户评论大多是根据微博事件而产生的,在判断用户评论是否与该微博事件相关的过程中,微博正文特征的使用能够有效促进用户评论相关性检测。为了根据特定任务选择有价值的和合适的特征,本文在共享层后面设计了一个特征筛选模块。共享特征筛选模块由 2 个单元组成,如图 3 所示,分别为门控筛选单元和注意力筛选单元。门控筛选单元用于过滤一些无用特征,注意力筛选单元用于关注用户评论相关性检测任务中有价值的共享特征。

门控单元采用一个单一的门控单元过滤共享特征中无用的特征。与 LSTM(Long Short-Term Memory)[24]的遗忘门机制相似,其共享特征通过 sigmoid 激活函数作为一种门控状态,再与共享特征进行点乘运算通过 tanh 激活函数作为当前状态的输出,如式(11)和式(12)所示:

$$\boldsymbol{g}=\delta(\boldsymbol{W}\cdot\boldsymbol{H}_{\mathrm{Shared}}+\boldsymbol{b}) \tag{11}$$

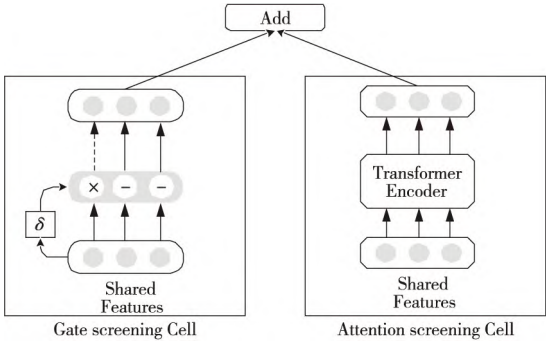

Figure 3 Module of shared feature screening

图 3 共享特征筛选模块

$$\boldsymbol{G}=\tanh(\boldsymbol{g}\odot\boldsymbol{H}_{\text{Shared}}) \tag{12}$$

其中，$\boldsymbol{H}_{\text{Shared}}\in\mathbf{R}^{l_1\times d}$ 为 2 个任务的共同特征；$\boldsymbol{g}\in\mathbf{R}^{l_1\times d}$ 为门控共享单元状态；$\boldsymbol{G}\in\mathbf{R}^{l_1\times d}$ 为共享特征 $\boldsymbol{H}_{\text{Shared}}$ 经过门控机制过滤后的特征；$\boldsymbol{W}\in\mathbf{R}^{l_1 d\times l_1 d}$ 和 $\boldsymbol{b}\in\mathbf{R}^{l_1\times d}$ 为可训练的参数；δ 为 sigmoid 激活函数；$\odot$ 表示点乘操作。

注意力筛选单元以 $\boldsymbol{H}_{\text{Shared}}$ 作为输入同样采用 transformer 编码模块[6]来获得更加有用的特征，如式(13)～式(15)所示：

$$\boldsymbol{H}_{\text{Shared}}=\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V} \tag{13}$$

$$\boldsymbol{A}_{\text{Shared}}=Transformer\_encoder(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \tag{14}$$

$$\boldsymbol{F}_{\text{Com}}=\boldsymbol{G}\oplus\boldsymbol{A}_{\text{Shared}} \tag{15}$$

其中，$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}\in\mathbf{R}^{l_1\times d}$；$\boldsymbol{A}_{\text{Shared}}\in\mathbf{R}^{l_1\times d}$ 为共享特征 $\boldsymbol{H}_{\text{Shared}}$ 经过注意力机制选择后的特征。最后将过滤后的输出特征 $\boldsymbol{G}$ 与经过选择后的输出特征 $\boldsymbol{A}_{\text{Shared}}$ 相加作为共享特征层筛选模块的输出 $\boldsymbol{F}_{\text{Com}}$。

3.5 事件预测模块

用户评论编码模块提取的特征与共享特征层筛选模块的输出特征进行拼接后，本文应用 softmax 函数分别实现对不同任务的分类，给出特定任务的概率分布预测，如式(16)～式(18)所示：

$$\bar{y}_1=Softmax(\boldsymbol{W}_1\boldsymbol{F}_1+\boldsymbol{b}_1) \tag{16}$$

$$\bar{y}_2=Softmax(\boldsymbol{W}_2\boldsymbol{H}_{\text{Shared}}+\boldsymbol{b}_2) \tag{17}$$

$$\boldsymbol{F}_1=[\boldsymbol{F}_{\text{Com}};\boldsymbol{H}_{\text{Com}}] \tag{18}$$

其中，$\bar{y}_1$ 是用户评论相关性检测任务预测结果；$\boldsymbol{F}_1\in\mathbf{R}^{(l_1 d+l_2 d)\times 2}$ 为门控筛选单元与注意力筛选单元相加后的特征 $\boldsymbol{F}_{\text{Com}}$ 与用户评论特征 $\boldsymbol{H}_{\text{Com}}$ 的拼接；$\bar{y}_2$ 是谣言检测任务预测结果；$\boldsymbol{W}_1\in\mathbf{R}^{(l_1 d+l_2 d)\times 2}$，$\boldsymbol{W}_2\in\mathbf{R}^{l_1 d\times 2}$；$\boldsymbol{b}_1$，$\boldsymbol{b}_2$ 是偏置项，为训练参数。

得到用户评论相关性检测任务和谣言检测任务的预测后，对模型进行训练以最小化所有任务的预测和真实分布的交叉熵，如式(19)和式(20)所示：

$$\zeta=\sum_{i=1}^{2}\lambda_i L(\bar{y}_1,y_{\text{Com}})+(1-\lambda_i)L(\bar{y}_2,y_{\text{Con}}) \tag{19}$$

$$L(\bar{y}_i,y_i)=y_i\log\bar{y}_i+(1-y_i)\log(1-\bar{y}_i) \tag{20}$$

其中，ζ 为 2 个任务损失的加权和，λ_i 为平衡损失参数，y_{con} 为微博正文的真实标签，y_{com} 为用户评论相关性真实标签。

4 实验

4.1 数据集和评估指标

本文使用的数据集是从新浪微博平台获取的 2020 年疫情相关的热门微博，共 201 条谣言微博及 11 233 条用户评论，378 条真实微博及 20 334 条用户评论。在实际生活中，正常信息量通常远大于谣言信息量，因此在构建疫情数据集时真实信息与谣言信息的比例大概为 2∶1。谣言微博的选取主要以新浪微博上的微博小助手官方辟谣平台为依据，挑选其中转发数超过 50、评论数超过 20 的热门微博，经过筛选与预处理后形成 json 格式文件。数据收集完成后，首先对微博事件标注标签，1 为谣言微博，0 为真实微博；然后针对某一个微博事件下的所有评论，同样标注标签，1 表示该条用户评论与微博描述的事件相关，0 表示与微博描述的事件不相关。

数据其它预处理主要包括：(1)去除文本内容中的多余符号、超链接和特殊字符；(2)去除微博中相同的用户评论信息。

本文使用准确率(A)、精确度(P)、召回率(R)和 $F1$ 分数($F1$)对谣言识别结果进行评价。

4.2 实验设置

模型的超参数配置方面，本文通过预先训练的搜狗新闻语料库[23]来表示微博正文和用户评论一个字(词)的 300 维词嵌入，这是一个包含 36 万字/词的搜狗新闻预训练语料库。将微博正文和用户评论进行 jieba 分词处理后，将处理完成的字或词构建词典，最后通过构建的词典依次提取预训练词向量，其中作者把微博正文和用户评论构建为一个词典。微博正文-用户评论编码模块、正文-评论交

叉注意力模块和共享特征层筛选模块中的 Transformer 编码模块[6]头数设置为 2,最长用户评论长度设置为 30,最长微博正文长度设置为 60,对于长度不足的采用 0 向量填充。数据批次设置为 64,词表大小设置为 20 000,学习率设为 5e−5,学习率衰减为 0.9,*dropout* 为 0.5,参数采用 Adam 优化器[25]更新。用户评论相关性检测任务损失平衡参数 $\lambda_1=0.4$,谣言检测任务损失平衡参数 $\lambda_2=0.6$。本文将数据集分割为训练集、验证集和测试集,分别包含 24 173,3 614 和 3 780 条用户评论。

4.3　基线模型分析

为了验证融合评论的多任务联合学习模型对微博谣言检测任务的有效性,本文采用几种非常典型的分类模型来比较编码方式的差异性。另外,本文还将 CMT-G&A 与当前最先进的方法进行比较。

为了更加公平地比较,本文设置学习率、微博正文和用户评论长度、*dropout* 等参数与主模型一致;LSTM 与 GRU 隐藏向量大小使用最佳性能参数,设置为 128;CNN 卷积核数量设置为 256,卷积核尺寸为(2,3,4)。本文使用的基线模型具体如下所示:

(1)BGRU(Bi-directional Gate Recurrent Unit):将微博正文和用户评论分别通过双向 GRU[26]提取文本特征表示,用户评论特征不经过过滤直接参与评论相关性检测辅助任务,用户评论特征与微博正文特征拼接后作为共享特征经过全连接层来实现谣言检测主任务预测。

(2)BLSTM(Bi-directional Long Short-Term Memory):将微博正文和用户评论分别通过双向 LSTM[26]提取文本特征表示,用户评论特征直接参与评论相关性检测辅助任务,用户评论特征与微博正文特征拼接后作为共享特征经过全连接层进行谣言检测主任务预测。

(3)RCNN (Region-CNN):将微博正文和用户评论分别通过双向 GRU[26,27]提取文本特征,通过最大池化分别对两者特征进行降维,用户评论直接通过全连接层进行评论相关性检测辅助任务,用户评论与微博正文特征拼接后作为共享特征经过全连接层进行谣言检测主任务预测。

(4)BLSTM-ATT(Bi-directional Long Short-Term Memory ATTention):将微博正文和用户评论分别通过双向 LSTM[26]提取文本特征,通过注意力机制[28]关注各自任务目标更关键的信息,抑制其他无用信息,用户评论与微博正文特征拼接后作为共享特征经过全连接层进行谣言检测主任务预测。

(5)BGRU-ATT(Bi-directional Gate Recurrent Unit ATTention):将微博正文和用户评论分别通过双向 GRU[26,27]提取文本特征,通过注意力机制[28]关注各自任务目标更关键的信息,共享特征经过全连接层进行谣言检测任务预测。

(6)CMT(Comment Multi-Task):首先将微博正文和用户评论分别通过微博正文-用户评论编码模块提取各自文本特征;再通过正文-评论交叉注意力模块提取共享特征进行谣言检测主任务预测;最后共享特征不经过共享特征筛选模块而是直接与用户评论拼接后进行用户评论相关性辅助任务预测。

(7)MT-trans-G-A(Multi-Task-Gate-Attention):Wu 等人[5]设计的多任务共享特征筛选框架,引入位置检测任务和虚假新闻检测任务来检测假新闻。

(8)dEFEND(Explainable Fake News Detection):Shu 等人[21]开发了正文评论联合注意网络,通过建立新闻句子和用户评论之间的相互联系去学习特征表示,并通过注意权重学习句子和评论的可解释程度。

(9)CMT-G&A:在 CMT 模型基础上通过加入门控机制和注意力机制来过滤和选择共享特征,以实现对微博谣言的检测,为本文所提的主要模型。

从表 1 所述的模型实验结果可以发现,本文基线模型 CMT 的准确率、精确率和 $F1$ 值都超过了其他所有基线模型,表明其他基线模型在融合用户评论的谣言检测任务中预测效果略有不足;而基线模型 CMT 引入 Transformer 编码模块[6]对 2 个任务的输入进行编码,利用其长距离依赖和并行性,提高了模型的性能,表明了本文模型编码方式的有效性;当前较先进模型中,MT-trans-G-A 的准确率、精确率、召回率和 $F1$ 值相比本文主要模型的要低,其原因可能是 MT-trans-G-A 更加注重多任务间共享特征的筛选,忽略了用户评论特征,而用户评论特征对提升谣言检测任务性能更加有效;dEFEND 则表现出了更好的预测结果,表明共同注意力能很好地挖掘微博正文和用户评论的关联性,相比本文主要模型结果较低的原因是微博用户量更大、用户评论数据更加嘈杂,评论质量是关键因素。CMT-G&A 在引入门控机制和注意力机制后预测性能有很好的提升,相较于 CMT 准确率提升了 6.1%,精确率提升了 17.7%及 $F1$ 值提升

了 7.7%。但是,召回率却低于基线模型,原因可能是本文为了更好地模拟真实场景,数据集构建过程中谣言事件相较于真实事件更少,导致模型更加偏向于预测真实事件;同时分词错误也是影响模型性能的重要因素,由于分词错误导致模型无法准确识别很多关键词的类别,进而影响模型预测结果。本文模型在另外 3 个性能指标上都达到了最优结果,表明本文模型是有效的。

Table 1 Performance comparison of baseline models

表 1 基线模型性能比较

基线模型	准确率/%	精确率/%	召回率/%	$F1$ 值
BGRU	83.07	73.82	78.53	76.10
BLSTM	82.11	67.28	**92.68**	77.96
RCNN	87.21	77.46	88.46	82.59
BLSTM-ATT	84.74	97.15	57.08	71.91
BGRU-ATT	88.07	77.28	92.41	84.17
CMT	88.78	80.33	89.06	84.47
MT-trans-G-A	85.76	86.99	68.67	76.75
dEFEND	89.95	86.58	83.69	85.11
CMT-G&A	**94.92**	**98.03**	86.94	**92.15**

4.4 消融实验分析

本文为了验证不同模块的有效性,将 CMT-G&A 模型分解成几个简化的模型,评价指标的得分情况如表 2 所示,最优结果用粗体表示。简化模型具体如下所示:

(1)S-task:只将微博正文通过 Transformer 编码[6]提取文本特征进行谣言检测主任务预测,用户评论相关性检测任务不参与模型训练。

(2)CMT-G:与 CMT 的不同之处在于,先将通过门控机制过滤无关特征后的共享特征与用户评论特征拼接,然后再进行用户评论相关性检测辅助任务预测。

(3)CMT-A:与 CMT 的不同之处在于,先将通过多头注意力机制选择有效特征后的共享特征与用户评论特征进行拼接,然后再进行用户评论相关性检测辅助任务预测。

Table 2 Experimental results of simplified models

表 2 简化模型实验结果

简化模型	准确率/%	精确率/%	召回率/%	$F1$ 值
S-task	89.61	96.75	72.22	82.70
CMT	88.78	80.33	**89.06**	84.47
CMT-G	91.43	93.59	80.62	86.62
CMT-A	91.54	96.36	78.32	86.41
CMT-G&A	**94.92**	**98.03**	86.94	**92.15**

表 2 展示了模型 CMT-G&A 与简化模型的实验结果。CMT-G&A 的结果在准确率、精确率和 $F1$ 值上都明显优于其他 4 种模型。CMT 相较于基于微博正文的单任务谣言检测模型 S-task,性能上有一些降低,原因可能在于融合用户评论后的共享特征确实有一些无用甚至有害特征干扰了检测。从 CMT-G 和 CMT-A 的实验结果可以看出,在加入门控筛选单元或注意力筛选单元后,模型的准确率、精确率和 $F1$ 值相比 S-task 的有较明显的提升。CMT-G&A 融合门控筛选单元与注意力筛选单元后性能最优,表明多任务联合学习间的共享特征分别通过门控机制过滤和注意力机制选择后对谣言检测任务有促进作用。可见,在融合用户评论的谣言检测中,用户评论的质量确实对谣言检测性能有一定影响。本文提出的融合评论的筛选多任务联合学习模型不仅能有效地挖掘微博事件中用户评论的有效信息,而且多任务中共享特征的过滤和选择能有效地促进微博谣言的检测。

4.5 超参实验分析

4.5.1 词嵌入维度对检测性能的影响

在深度学习中,模型的参数设置对实验结果也会有很大的影响,通过调节模型中的一些重要参数能更大程度地提升模型性能。为了验证随机初始化和预训练词向量对模型效果的影响,本文做了如下实验:

针对随机初始化词向量,分别设置维度为 300,512 和 768;对于预训练词向量,选择搜狗新闻语料库[23]训练的 sou-gou 词向量;为了公平起见,本文选择 S-task、CMT、CMT-G&A 3 个典型模型进行对比,实验结果如图 4 所示。

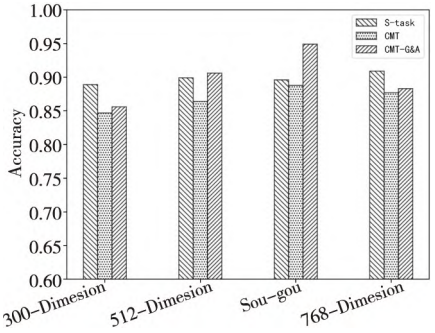

Figure 4 Sensitivity analysis about word embedding

图 4 词嵌入敏感性分析

在从图 4 可以看出,CMT-G&A 和 CMT 在预训练的词向量上表现出了更好的性能,S-task 模型对于随机初始化或预训练词向量变化不明显,同时随机初始化词向量的维度过大和过小对模型的性

能也有较大的影响。本文后续采用预训练的 sougou 词向量继续开展实验。

4.5.2　头数目对检测性能的影响

Transformer[6]中的自注意力机制能够捕获长距离依赖,并且能够学习到句子内部结构和语法,通过设置多个头可以使模型关注不同方面的信息。为了验证不同多头注意力对模型性能的影响,本文还做了如下实验:对于主要模型 CMT-G&A,对自注意力机制设置不同头数,实验结果如图 5 所示。从图 5 可以看到,头数设为 2 时模型的预测性能表现最好。其原因在于,头数过多会造成注意力冗余,参数过多反而影响模型的性能;头数过少又会导致注意力特征提取不充分,模型表达能力不足。本文后续将头数设为 2 继续开展实验。

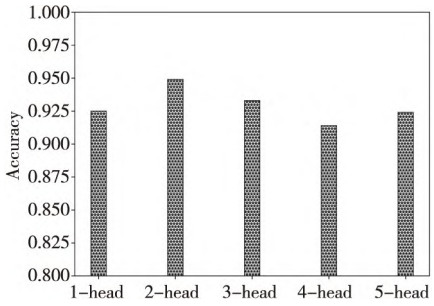

Figure 5　Sensitivity analysis about multi-head attention

图 5　多头注意力敏感性分析

4.6　重要词汇分析

在对同一微博事件进行预测时,用户评论中每个单词的重要权值也不同,为了更加直观地表示 CMT-G&A 模型如何从用户评论中学到有用的信息,同时验证用户评论信息对谣言检测任务的影响,本文从数据集中取出一个微博事件,统计了用户评论对应微博事件的单词权重并进行热力图展示,如图 6 所示。

从图 6 可知,模型针对同一事件在融合 2 条不同的用户评论时,对用户评论中词的关注度是不一样的。颜色深的部分表示当前用户评论中对于微博事件注意力权重较高的词,颜色越深权重越高。不考虑特殊符号例如“,”“。”等,用户评论中“不信谣”“谣言”等关键词表现出了较高的权重。实验表明用户评论也为微博事件的判定提供了一些重要的线索,融合用户评论对谣言检测是有效的,可以很大程度上帮助我们识别网络谣言。

5　结束语

本文针对微博谣言检测任务中文本特征不足,用户评论整体质量不高的问题,提出了一种融合评论的筛选多任务联合学习方法,通过用户评论与微博事件之间的关联性,将谣言检测任务作为主任务,用户评论相关性检测任务作为辅助任务,并通过联合学习同时学习和更新主任务模型和辅助任务模型的参数。一系列实验结果表明,融合评论的多任务联合学习方法不仅能较好地融合用户评论信息,而且用户评论的过滤和选择更好地提升了谣言检测任务的性能。

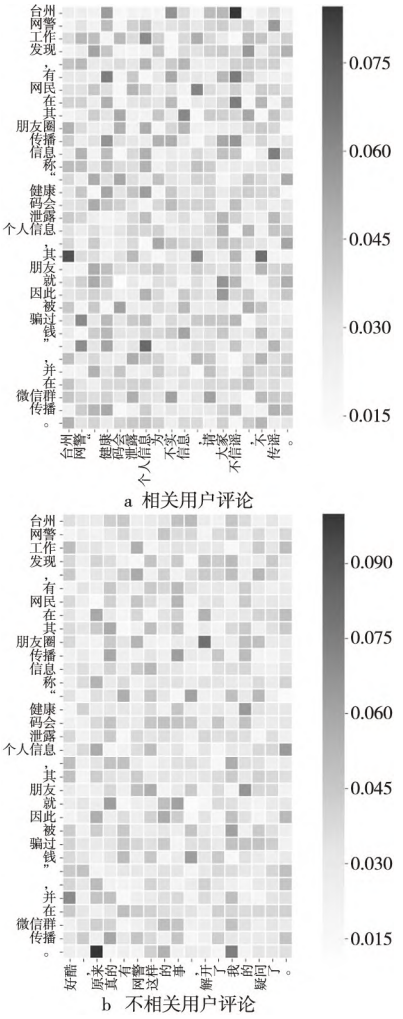

Figure 6　User comment weight visualization

图 6　用户评论权重可视化

未来将探索未标注的辅助特征,例如多媒体内容(图片、视频)往往比单独的文字信息更容易引起注意,后续尝试利用这些多模态特征来进一步提升谣言检测任务的性能。

参考文献:

[1] Ruchansky N, Seo S, Liu Y. CSI: A hybrid deep model for fake news detection[C]//Proc of the 26th ACM International Conference on Information and Knowledge Management, 2017:797-806.

[2] Guo H,Cao J,Zhang Y,et al. Rumor detection with hierarchical social attention network[C]//Proc of the 27th ACM International Conference on Information and Knowledge Management,2018:943-951.

[3] Wu L W,Rao Y,Nazir A,et al. Discovering differential features:A dversarial learning for information credibility evaluation[J]. Information Sciences,2020,516:453-473.

[4] Kochkina E,Liakata M, Zubiaga A. All-in-one:Multi-task learning for rumour verification[J]. arXiv:1806.03713,2018.

[5] Wu L W, Yuan R, Jin H L,et al. Different absorption from the same sharing:Sifted multi-task learning for fake news detection[J]. arXiv:1909.01720,2019.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of International Conference on Neural Information Processing Systems, 2017:5998-6008.

[7] Zhou X,Zafarani R,Shu K,et al. Fake news:Fundamental theories,detection strategies and challenges[C]//Proc of the 20th International Conference on Web Search and Data Mining,2019:836-837.

[8] Potthast M,Kiesel J,Reinartz K,et al. A stylometric inquiry into hyperpartisan and fake news[J]. arXiv:1702.05638,2017.

[9] Guo C, Cao J,Qiang S, et al. Exploiting emotions for fake news detection on social media[J]. arXiv:1903.01728,2019.

[10] Karimi H, Roy P,Tang J L,et al. Multi-source multi-class fake news detection[C]//Proc of the 27th International Conference on Computational Linguistics,2018:1546-1557.

[11] Karimi H, Tang J. Learning hierarchical discourse-level structure for fake news detection[J]. arXiv:1903.07389,2019.

[12] Wang W Y. Liar,liar pants on fire:A new benchmark dataset for fake news detection[C]//Proc of the 55th Annual Meeting of the Association for Computational Linguistics,2017:421-426.

[13] Abavisani M, Wu L W, Hu S W, et al. Multimodal categorization of crisis events in social media[C]//Proc of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, 2020:14679-14689.

[14] Shu K, Zhou X Y, Wang S H, et al. The role of user profiles for fake news detection[C]//Proc of the 2019 IEEE/ACM International Conference on Social Networks Analysis and Mining, 2019:436-439.

[15] Castillo C,Mendoza M,Poblete B. Information credibility on Twitter[C]//Proc of the 20th International Conference on World Wide Web,2011:675-684.

[16] Lu Y J, Li C T. GCAN:Graph-aware co-attention networks for explainable fake news detection on social media[J]. arXiv:2004.11648, 2020.

[17] Shu K, Wang S H, Liu H. Beyond news contents:The role of social context for fake news detection[C]//Proc of the 12th ACM International Conference on Web Search and Data Mining, 2019:312-320.

[18] Monti F,Frasca F,Eynard D,et al. Fake news detection on social media using geometric deep learning[J]. arXiv:1902.06673,2019.

[19] Shu K, Mahudeswaran D, Wang S, et al. Hierarchical propagation networks for fake news detection:Investigation and exploitation[C]//Proc of the International AAAI Conference on Web and Social Media, 2020:626-637.

[20] Duan Da-gao,Wang Chang-sheng,Han Zhong-ming,et al. A rumor detection model based on weibo's reviews[J]. Computer Simulation,2016,33(1):386-390. (in Chinese)

[21] Shu K, Cui L M, Wang S H, et al. dEFEND:Explainable fake news detection[C]//Proc of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019:395-405.

[22] Wu L W, Rao Y. Adaptive interaction fusion networks for fake news detection[J]. arXiv:2004.10009,2020.

[23] Li S, Zhao Z, Hu R F, et al. Analogical reasoning on Chinese morphological and semantic relations[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics,2018:138-143.

[24] Graves A,Jürgen S. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks,2005,18(5-6):602-610.

[25] Kingma D,Ba J. Adam:A method for stochastic optimization[C]//Proc of the 3rd International Conference on Learning Representations,2015:1-15.

[26] Ma J, Gao W, Mitra P,et al. Detecting rumors from microblogs with recurrent neural networks[C]//Proc of International Joint Conference on Artificial Intelligence, 2016:3818-3824.

[27] Li Li-zhao,Cai Guo-yong,Pan Jiao. A microblog rumor events detection method based on C-GRU[J]. Journal of Shandong University (Engineering Science),2019,49(2):106-110. (in Chinese)

[28] Liao Xiang-wen,Huang Zhi,Yang Ding-da,et al. Rumor detection in social media based on a hierarchical attention network[J]. Scientia Sinica Informationis,2018,48(11):1558-1574. (in Chinese)

附中文参考文献:

[20] **段大高,王长生,韩忠明,等. 基于微博评论的虚假消息检测模型[J]. 计算机仿真**,2016,33(1):386-390.

[27] **李力钊,蔡国永,潘角. 基于 C-GRU 的微博谣言事件检测方法[J]. 山东大学学报(工学版)**,2019,49(2):106-110.

[28] **廖祥文,黄知,杨定达,等. 基于分层注意力网络的社交媒体谣言检测[J]. 中国科学:信息科学**,2018,48(11):1558-1574.

作者简介:

王繁(1996-),**男,云南曲靖人,硕士生,CCF 会员**(G4216G),**研究方向为自然语言处理和深度学习。E-mail**:1127933736@qq.com

WANG Fan, born in 1996, MS candidate, CCF member(G4216G), his research interests include natural language processing and deep learning.