文章编号:1672-3961(2022)04-0089-10 DOI:10.6040/j.issn.1672-3961.0.2021.296

基于混合深度模型的虚假信息早期检测

黄皓[1],周丽华[1*],黄亚群[1],姜懿庭[2]
(1.云南大学信息学院, 云南 昆明 650000;2.云南师范大学信息学院, 云南 昆明 650000)

摘要:针对一种信息特征进行检测方法在信息传播早期阶段提取的特征信息往往不充分,导致传播早期阶段检测准确率较低的问题,提出一个新颖的混合深度模型 EGSI,模型由 EXTRACT、GRU、SCORE 和 INTERATE 4 个模块组成。EXTRACT 通过卷积神经网络提取信息的传播路径特征,GRU 通过门控循环单元捕获信息的文本特征和反馈特征,SCORE 基于用户行为挖掘用户特征,INRERATE 整合以上特征并预测出信息事件类标。EGSI 通过整合信息最基本的 4 种特征(文本、用户、反馈、传播路径),从而可以在信息传播的早期阶段充分提取可用特征信息,进而较准确地检测出虚假信息。真实数据集的试验结果表明,模型在信息传播 60 min 内的准确率达到 95.9%。相比基准方法,EGSI 模型在检测虚假信息的准确率和时效性之间取得了较好的平衡。

关键词:虚假信息早期检测;混合深度模型;神经网络;时序分析;信息特征
中图分类号:TP181 **文献标志码**:A
引用格式:黄皓,周丽华,黄亚群,等. 基于混合深度模型的虚假信息早期检测[J].山东大学学报(工学版),2022,52(4):89-98.
HUANG Hao, ZHOU Lihua, HUANG Yaqun, et al. Early detection of fake news based on hybrid deep model[J]. Journal of Shandong University (Engineering Science), 2022, 52(4):89-98.

Early detection of fake news based on hybrid deep model

HUANG Hao[1], ZHOU Lihua[1*], HUANG Yaqun[1], JIANG Yiting[2]
(1. School of Information, Yunnan University, Kunming 650000, Yunnan, China;2. School of Information, Yunnan Normal University, Kunming 650000, Yunnan, China)

Abstract: The feature information extracted in the early stage of information dissemination was often insufficient in the detection method based on a kind of information feature, which led to the problem of low detection accuracy in the early stage of dissemination, a novel hybrid depth model EGSI was proposed, the model was composed of four modules: EXTRACT, GRU, SCORE and INTERATE. EXTRACT extracted the characteristics of the propagation path of information through convolutional neural networks, GRU captured the text features and feedback features of the information through gated recurrent units, SCORE mined user′s features based on user behavior, and INRERATE integrated the above features and predicted the information′s event category. EGSI integrated the four most basic characteristics of information (text, user, feedback, and communication path), which could fully extract the available feature information in the early stage of information dissemination, and then more accurately detect false information. Experimental results on real data sets showed that the model could detect false information with an accuracy of 95.9% within 60 min of information dissemination. Compared with the benchmark method, the EGSI model had achieved a better balance between the accuracy and timeliness of detecting false information.

Keywords: early detection of fake news; hybrid deep model; neural network; time series analysis; information characteristics

收稿日期:2021-06-07;**网络首发时间**:2022-06-27 17:14:01
网络首发地址:https://kns.cnki.net/kcms/detail/37.1391.T.20220624.1749.002.html
基金项目:国家自然科学基金项目(61762090,62062066,61966036 和 61662086);云南省大学创新计划(IRTSTYN);国家社会科学基金项目(18XZZ005)
第一作者简介:黄皓(1994—),男,四川成都人,硕士研究生,主要研究方向为数据挖掘、信息扩散。E-mail:840670997@qq.com
*通信作者简介:周丽华(1968—),女,云南昆明人,教授,博士,CCF 会员,主要研究方向为数据挖掘、社会网络分析、人工智能。
E-mail: lhzhou@ynu.edu.cn

0 引言

随着社交媒体网站和社交软件的日益普及，用户生成的消息可以快速地到传送给大众，网络社交媒体已成为虚假信息传播的理想场所。同时虚假信息的大规模扩散可能造成巨大社会危害和经济损失，甚至引起人们恐慌。比如，在2021年1月9日新冠疫情期间，快手某主播为了提高自身人气，在直播过程过散布“红寺堡疫情变严重了，红寺堡月底就封城”的谣言，引起人们的讨论和担心，并出现市民连夜出走和抢购物资的情况。因此，在信息传播过程中尽可能早的检测出虚假信息，具有较大的社会需求和较好的社会利益。

传统的检测算法大多基于某一种特定的特征，比如文本、用户、情感、传播网络等检测虚假信息，但这些方法都有各自的局限性，往往不能充分利用信息事件蕴含的信息，从而有效的检测虚假信息。比如：文献[1-4]中基于文本特征的算法，在处理Twitter和Weibo这类主流社交网站上信息文本简短的情况，算法通常无法提取到足够的可用文本特征信息进行检测；对可能包含图片、音频、视频或几者混合的信息事件不能进行准确检测。文献[5]采用基于源用户特征的算法，主要根据信息事件发布者的信息特征设计检测算法，但是这种方法忽略了信息发布者和信息传播者之间的联系，没有利用信息传播者包含的特征信息。最近的研究中，文献[6-8]和文献[9-10]分别探索了通过用户评论提取的时态语言特征和通过传播网络提取的时态结构特征进行信息检测的算法，虽然比起传统算法检测效果更好，但在检测虚假信息的时效性上有巨大限制。时态语言特征和时态结构特征在信息传播的早期阶段往往无法充分提取，因为网络用户往往都是直接转发源信息事件且没有任何评论，并且在信息传播早期阶段，用户转发评论很容易被操纵，因为发布虚假信息的组织可能对虚假信息进行转发评论，从而促进虚假信息的传播，并且阻碍检测，这导致算法在信息传播的早期阶段检测准确率不高。

针对以上问题，本研究提出一种混合深度模型EGSI（EXTRACT、GRU、SCORE、INTERATE），整合信息的文本、用户、反馈、传播路径4种基本特征，从而使得模型可以在信息传播的早期提取到充分的可用信息特征进行虚假信息检测。同时在深度神经网络上建立模型，可以自动提取特征而不需要人工选择提取。在真实数据集的试验结果表明，模型可以在信息传播的早期较为准确地检测出虚假信息，在准确率和时效性两者之间取得了较好的平衡。

1 相关工作

近年来，社交媒体上的虚假信息已经引起人们的极大关注，现有研究的主要方向旨在开发基于机器学习的分类器，以基于各种信息特征设计方法自动检测在社交媒体环境中传播的信息事件的真假。

基于文本特征对虚假信息进行检测是最常用的方法，其中较为新颖的方法是从信息事件和其相关用户转发评论组成的序列中提取的时态语言特征对信息进行检测。文献[7]利用循环神经网络（recurrent neural network，RNN）提取信息事件和转发评论之间随时间变化的隐藏联系，即时态语言特征，从而检测虚假信息。作为对文献[7]方法的延伸，文献[11]将软注意力机制纳入递归神经网络，由此集中不同的时态语言特征，从而进行检测。文献[12]提出基于多级卷积神经网络的模型，提取了局部卷积特征和全局语义特征并结合计算敏感词权重来检测信息。

用户特征也常用于研究检测虚假信息，用户特征相比文本特征更难被操纵，因为虚假信息传播者可以为他们传播的虚假新闻提供虚假评论，从而导致文本特征无用。文献[13]提出基于可信度评分的新方法，该方法可以给信息发布者和用户的可信度评分，然后检测虚假信息。文献[14]认为虚假信息的主要传播工具是自动与用户互动的虚假软件机器人，根据拥有机器人朋友的数量设置轻信用户和可信用户从而检测虚假新闻。文献[15]提出基于用户特征和图嵌入的混合方法，从Twitter用户的关注者和关注者社交图谱中提取特征从而检测虚假新闻。文献[16]通过在用户的社交网络中发现潜在的用户连接，并重建用户网络，有效识别出倾向于传播假新闻的用户来检测虚假信息。

基于社交网络中信息的传播路径来提取特征也被广泛研究。文献[8]提出一种基于图内核的SVM分类器，该分类器通过评估传播树结构之间的相似性来捕获高阶模式以区分不同类型的虚假信息。文献[9]利用流行病学模型来描述Twitter中由真实信息和虚假信息构建的信息级联，从而学习信息事件的扩散方式和过程来检测信息。文献[17]利用新闻故事的对话片段之间的隐式链接来

预测其真实性。文献[18]利用卷积神经网络(convolutional neural networks，CNN)和门控循环单元(gated recurrent unit，GRU)构建出神经网络模型PPC(propagation path classification)，可以将信息的传播路径构建固定长度多变量序列，然后通过CNN和GRU神经网络分别挖掘传播路径的局部和全局特征，能在信息传播早期检测出虚假信息，但检测虚假信息的准确率不高。

反馈特征在现有研究中也逐渐得到关注。文献[19]提出了一种带有用户响应生成器(two-level convolutional neural networks-user response generator，TCNN-URG)的新型两级卷积神经网络，其中TCNN捕获信息文本中的语义信息，URG从历史用户响应中了解用户对信息文本的生成模型，该模型可用于生成对新文章的响应，以协助虚假信息检测。文献[20]提出新闻报道的反馈时间模型在理解内容本身的属性上面起着重要作用。还有一种流行方法是通过在社交图谱上研究新闻报道传播后收到的反馈响应来检测虚假信息[21-23]。

除此之外，一种较为新颖的方法是结合不同类型的特征来检测虚假信息。文献[24]结合了用户、语言、结构和时间特征，以在不同的时间窗口内检测假新闻。文献[25]提出能够提取文本、反馈、用户3种特征的混合深度模型CSI(capture score integrate)，其中C模块利用长短时记忆神经网络挖掘信息事件和其相关转发评论的时间序言特性，S模块通过对用户参与信息事件的数目构建的矩阵进行奇异值分解，得到用户特征，I模块对前两个模块的输出向量进行整合，输入到全连接层，能较为准确地预测出信息事件类标，但在信息传播的早期检测中准确率较低。本研究提出了一种结合卷积神经网络CNN、门控循环单元GRU和对基于用户行为对用户评分的3种设计方式构建的混合深度模型，自动提取了文本、用户、反馈、传播路径4种信息特征，可以在信息传播的早期阶段提取到充分的特征信息，进而在传播的早期较为准确的检测虚假信息。

2　EGSI模型构建及分析

2.1　EGSI模型介绍

模型由EXTRACT、GRU、SCORE和INTERATE 4个模块组成，模型可以将用户、反馈和文本、传播路径的特征信息转换为低维向量并整合起来以最终检测虚假信息。模型结构如图1所示。

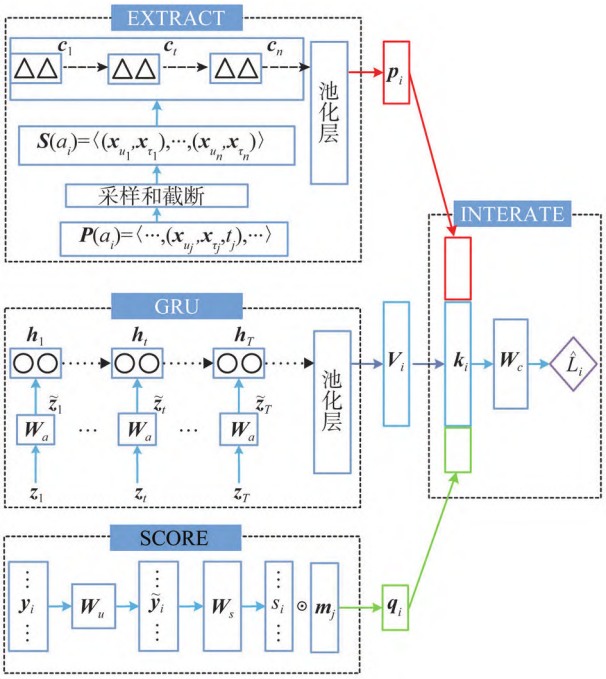


图1　EGSI模型结构

Fig.1　EGSI model structure

2.2　EXTRACT模块

2.2.1　CNN输入向量设置

此模块中，使 $A=\{a_1,a_2,\cdots,a_{|A|}\}$ 表示信息事件的集合，$U=\{u_1,u_2,\cdots,u_{|U|}\}$ 表示社交媒体中用户的集合。用 $\boldsymbol{x}_{u_j}\in\mathbf{R}^d$ 表示用户 $u_j\in U$ 的用户特征向量，$\boldsymbol{x}_{\tau_j}\in\mathbf{R}^d$ 表示用户 u_j 发出信息的文本特征向量。信息发布的时刻设置为0，因此，当 $t>0$ 时表示信息事件的转发评论已经发出。模块首先将给定信息事件 a_i 的传播路径构建为可变长度多变量时间序列 $\boldsymbol{P}(a_i)=\langle\cdots,(\boldsymbol{x}_{u_j},\boldsymbol{x}_{\tau_j},\boldsymbol{t}_j),\cdots\rangle$，其中元组 $(\boldsymbol{x}_{u_j},\boldsymbol{x}_{\tau_j},t_j)$ 表示用户 u_j 在 t_j 时刻对信息事件 a_j 的发表或者回复的内容 τ_j。获得 $\boldsymbol{P}(a_i)$ 后，需要将其转换为长度为 n 的固定长度多变量序列 $\boldsymbol{S}(a_i)=\langle(\boldsymbol{x}_{u_1},\boldsymbol{x}_{\tau_1}),\cdots,(\boldsymbol{x}_{u_n},\boldsymbol{x}_{\tau_n})\rangle$ 后才能作为卷积神经网络的输入。如果 $\boldsymbol{P}(a_i)$ 中的元组数量超过 n，则截短 $\boldsymbol{P}(a_i)$，使得 $\boldsymbol{S}(a_i)$ 组数为 n，如果 $\boldsymbol{P}(a_i)$ 的元组数量少于 n，则通过随机过采样使得 $\boldsymbol{S}(a_i)$ 的组数为 n。

2.2.2　CNN神经网络

使用卷积神经网络学习 $\boldsymbol{S}(a_i)$ 的向量表示。构建高度为 h 的1-D卷积神经网络学习每个给定信息事件 a_i 下提取到的向量表示 $\langle(\boldsymbol{x}_{u_1},\boldsymbol{x}_{\tau_1}),\cdots,(\boldsymbol{x}_{u_n},\boldsymbol{x}_{\tau_n})\rangle$，生成特征向量

$$\boldsymbol{c}_t=\tanh(\boldsymbol{W}_f*\langle(\boldsymbol{x}_{u_1},\boldsymbol{x}_{\tau_1}),\cdots,(\boldsymbol{x}_{u_n},\boldsymbol{x}_{\tau_n})\rangle+\boldsymbol{b}_f),\tag{1}$$

式中：$\boldsymbol{c}_t\in\mathbf{R}$，$(\boldsymbol{x}_{u_1},\boldsymbol{x}_{\tau_1})$ 表示源推文的用户特征和文本特征，$(\boldsymbol{x}_{u_n},\boldsymbol{x}_{\tau_n})$ 表示第 n 个转发评论用户的用户

特征和文本特征。因此，$\langle(\boldsymbol{x}_{u_1},\boldsymbol{x}_{\tau_1}),\cdots,(\boldsymbol{x}_{u_n},\boldsymbol{x}_{\tau_n})\rangle$ 表示源推文发布后所有的用户和文本构建的传播路径信息。然后通过包含 K 个过滤器的卷积层进行相同卷积运算以生成$\boldsymbol{c}_t\in\mathbf{R}$。最后利用池化层生成传播路径特征 $\boldsymbol{p}_i$。

2.3 GRU 模块

2.3.1 信息时间段划分方法

通过将每条推文和转发评论输入到 Doc2vec 得到特征向量，然后作为神经网络的输入。单独将事件和转发评论嵌入为维度相同的向量，这样处理不能提取到推文和转发评论之间的时间联系，同时由于转发评论的文本简短，这样提取的信息往往不够充分。因此，通常研究中，会对信息事件的传播周期进行等间隔的划分，但是这种方式会导致在后期时间段内利用信息急剧减少甚至没有可利用信息提供给模型的情况。本研究提出一种动态划分方法，将信息事件传播周期划分为非等间隔的固定大小时间分片，通过更改分片和分片之间的间隔大小来划分传播周期，对处于同一时间分片内的信息文本统一放入到 doc2vec 得到神经网络的输入 $\boldsymbol{z}_\tau$。对于信息事件 a_i，假设其信息传播周期为$[0,T]$，0 表示微博发布时刻，T 表示评论终止时刻，固定时间跨度为 B，n 表示时间分片数量，表达式为

$$n=T/B。\tag{2}$$

根据式(2)可以首先得到 n 个时间长度相同的时间分片，然后对 n 个时间分片分别进行检测，若某个分片内没有用户转发评论该推文，则分片为空，舍弃此分片，因此时间分片之间间隔的时间跨度不一致，最后得到 $m(m\leqslant n)$个时间分片，然后依次将非空时间分片的内容通过 doc2vec 转换为向量，最后得到事件 a_i 的特征向量矩阵 $\boldsymbol{z}_t$。a_i 时间分片的第 t 个分区处，其特征向量具有以下形式：η 是第 t 个非空分片下推文转发评论的数量；Δt 是当前分片和上一个非空分片的时间差；$\boldsymbol{z}_u$ 是 t 分区下转发回复事件 a_i 的所有用户 u_i 的用户特征的平均值；$\boldsymbol{z}_\tau$ 是 t 分区下文本内容的特征；最后 m 个分片构建出向量矩阵 $\boldsymbol{z}_t$。

2.3.2 GRU 输入向量设置

此模块中，通过用户转发回复的频率和分布来捕获用户和事件之间的语言时间特征。换言之，不仅需要提取出与事件 a_i 互动的用户数量，还要提取出事件 a_i 下用户互动的时间间隔。通过从事件发布和用户转发回复的互动中，提取信息的文本特征和反馈特征。使用 GRU 神经网络来对信息进行特征提取，GRU 已被证明在捕获数据中的时间模式和整合不同信息源方面有着不错的性能。GRU 的一个关键组成部分是选择用作每个输入单元的张量，输入张量 $\boldsymbol{z}_t$ 和张量矩阵 $\boldsymbol{z}_i$ 表达式为

$$\begin{cases}\boldsymbol{z}_t=(\eta,\Delta t,\boldsymbol{z}_u,\boldsymbol{z}_\tau)\\ \boldsymbol{z}_i=[\boldsymbol{z}_t,\cdots,\boldsymbol{z}_t,\cdots,\boldsymbol{z}_m]\end{cases},\tag{3}$$

式中：$\boldsymbol{z}_t$ 表示事件 a_i 的第 t 个时间分片提取到的信息特征；η 表示事件转发评论的数量；Δt 表示用户之间转发回复的时间差，η 和 Δt 可以表示出事件收到的反馈的频率和时间分布；$\boldsymbol{z}_u$ 和 $\boldsymbol{z}_\tau$ 分别是第 t 个时间分片内，所有参与转发回复的用户集合中提取的用户特征和所有用户转发评论的文本特征；$\boldsymbol{z}_i$ 表示给定事件 a_i 下所有非空时间分片提取的特征向量所构建的特征向量矩阵。

2.3.3 GRU 神经网络

由于提取到的特征信息表现形式并不统一，所以并不能直接作为 GRU 神经网络的输入，需要在原始特征 $\boldsymbol{z}_t$ 和 GRU 神经网络的输入特征 $\tilde{\boldsymbol{z}}_t$ 中间添加了一个嵌入层。这个嵌入层也是一个全连接层，表达式为

$$\tilde{\boldsymbol{z}}_t=\tanh(\boldsymbol{W}_a\boldsymbol{z}_t+\boldsymbol{b}_a),\tag{4}$$

式中：$\boldsymbol{W}_a$ 是应用于 t 时刻的初始特征 $\boldsymbol{z}_t$ 的权重矩阵，$\boldsymbol{b}_a$ 是一个偏置向量。$\boldsymbol{W}_a$、$\boldsymbol{b}_a$ 也都应用于所有的 $\boldsymbol{z}_t$，通过嵌入层得到 $\tilde{\boldsymbol{z}}_t$，GRU 神经网络表达式为

$$\begin{cases}\boldsymbol{g}_t=\sigma(\boldsymbol{U}_z\tilde{\boldsymbol{z}}_t+\boldsymbol{W}_z\boldsymbol{h}_{t-1})\\ \boldsymbol{r}_t=\sigma(\boldsymbol{U}_r\tilde{\boldsymbol{z}}_t+\boldsymbol{W}_r\boldsymbol{h}_{t-1})\\ \tilde{\boldsymbol{h}}_t=\tanh(\boldsymbol{U}_h\tilde{\boldsymbol{z}}_t+\boldsymbol{h}_{t-1}\boldsymbol{W}_h\boldsymbol{r}_t)\\ \boldsymbol{h}_t=(1-\boldsymbol{g}_t)\boldsymbol{h}_{t-1}+\boldsymbol{g}_t\tilde{\boldsymbol{h}}_t\end{cases},\tag{5}$$

式中：将 $\tilde{\boldsymbol{z}}_t$ 和 $\boldsymbol{h}_{t-1}$ 作为输入，$\boldsymbol{h}_{t-1}$ 是上一时刻的隐含状态，最后输出 $\boldsymbol{h}_t$。$\boldsymbol{U}_z$、$\boldsymbol{U}_r$、$\boldsymbol{U}_h\in\mathbf{R}^{m\times d}$，$\boldsymbol{W}_z$、$\boldsymbol{W}_r$、$\boldsymbol{W}_h\in\mathbf{R}^{m\times m}$都是权重矩阵，其中 d 是文本向量的维度，m 是 GRU 单元输出的维度。然后用均值池保留主要特征同时降低维数，防止过拟合，提高模型的泛化能力，通过公式 $\boldsymbol{v}_j=\dfrac{1}{n}\sum_{t=1}^{n}\boldsymbol{h}_t$ 来减少 GRU 单元输出的向量序列$\langle\boldsymbol{h}_1,\cdots,\boldsymbol{h}_n\rangle$，最后输出表示给定推文 a_i 和其转发评论的时间序列模型的低维向量表示 $\boldsymbol{v}_i$，即文本特征和反馈特征的整合，之后将 $\boldsymbol{v}_i$ 送到 INTERNATE用于信息检测。

2.4 SCORE 模块

此模块中，提取转发回复用户的行为特征，首先根据设置的阈值 Threshold 从所有用户构成的集合中提出出现次数最多的对应数量用户，并与所有微博事件构建出用户事件关联矩阵，通过这种方式

降低矩阵的复杂度,以便提取到用户主要信息。然后对矩阵使用奇异值分解(singular value decomposition, SVD),得到 $\boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^{\mathrm{T}}$,其中 $\boldsymbol{U}'$表示左奇异矩阵,$\boldsymbol{V}'$表示右奇异矩阵,$\boldsymbol{\Sigma}'$表示奇异值,通过 $\boldsymbol{U}'\boldsymbol{\Sigma}'$得到每个用户的用户特征的低维向量示 $\boldsymbol{y}_i$。接着使用全连接层来提取每个用户的向量表示形式,表达式为

$$\tilde{\boldsymbol{y}}_i=\tanh(\boldsymbol{W}_u\,\boldsymbol{y}_i+\boldsymbol{b}_u),\tag{6}$$

式中:$\boldsymbol{W}_u$ 是权重矩阵,$\boldsymbol{b}_u$ 是偏置矩阵。带有参数 λ 的 L2-正则约束使用在 $\boldsymbol{W}_u$ 上,并通过使用权重矢量 $\boldsymbol{W}_s$ 为每个用户生成标量分数 s_i,表达式为

$$s_i=\sigma(\boldsymbol{W}_s^{\mathrm{T}}\cdot\tilde{\boldsymbol{y}}_i+\boldsymbol{b}_s),\tag{7}$$

式中:$\boldsymbol{b}_s$ 是全连接层的偏置向量,σ 是 sigmoid 函数。为了与其他两个模块的输出信息更好整合,对 s_i 施加掩码 $\boldsymbol{m}_j$,最终得到特征向量 $\boldsymbol{q}_i$。

2.5　INTERATE 模块

EXTRACT、GRU 和 SCORE 3 个模块输出了模型所需要的 4 个特征,为了整合这些信息特征,提出了第 4 个模块 INTERATE,在该模块中,整合传播路径特征 $\boldsymbol{p}_i$,信息文本特征和反馈特征 $\boldsymbol{v}_i$,用户评分 $\boldsymbol{q}_i$ 生成结果向量 $\boldsymbol{k}_i$,输入到全连接层中为每则微博生成预测标签 $\tilde{L}_i$,表达式为

$$\tilde{L}_i=\sigma(\boldsymbol{W}_c^{\mathrm{T}}\boldsymbol{k}_i+\boldsymbol{b}_c)。\tag{8}$$

最后整合步骤使得所有模块整合起来生成预测。通过与 EXTRACT、GRU 和 SCORE 模块联合训练 EGSI,使得该模型可以同时学习用户信息、路径信息和推文信息。同时,EGSI 模型提取出不同的信息特征,并将这些特征整合起来用损失函数进行处理以进行最终预测,得到事件标签。其中 EGSI 模型的损失函数为

$$J=\frac{1}{n}\sum_{j=1}^{N}[L_j\lg\tilde{L}_J+(1-L_j)\lg(1-\tilde{L}_j)]+\frac{\lambda}{2}\|\boldsymbol{W}_u\|_2^2,\tag{9}$$

式中 L_j 是微博的真实标签。为了防止 EGSI 模型在训练中过拟合,在训练过程中将随机丢弃 $\boldsymbol{W}_a$、$\boldsymbol{W}_s$ 中的神经单元。在这些约束下,EXTRACT、GRU 和 SCORE 以及 INTERATE 中的参数通过反向传播进行联合训练。

3　试验分析

3.1　数据集

本研究使用从新浪微博收集的数据集进行试验,这数据集由文献[7]提供,数据集包含了 4 种特征信息:文本特征、反馈特征、用户特征以及传播路径特征。数据集统计信息如表 1 所示。在训练模型过程中,将数据集划分为训练集、验证集、测试集 3 个部分。为了试验结果对比公平,使用相同的数据集设置,将数据集的 80%设置为训练集,将 5%的数据集设置为验证集,15%的数据集设置为测试集。

表 1　数据集统计
Table 1　DatasetStatics

微博事件数量	用户数量	转发回复数量	虚假微博数量
4664	2 819 338	3 752 459	2313
真实微博数量	事件最大评论数量	事件最小评论数量	微博平均回复
2351	59 138	10	1808

数据集样本示例如表 2 所示,在数据集中每条源微博和其相关的一系列用户转发评论被看作为一个事件,事件编号表示每个源微博的唯一标识,用户编号表示每个用户的唯一标识,传播路径由发布的源微博和其发布后一系列相关的用户转发评论构建,时间戳表示微博或者转发评论的时间,文本内容表示微博或者转发评论的文本,每条源微博都有二分类标签 L,比如“0”和“1”,“1”表示事件为真,“0”表示事件为假。

表 2　数据集示例
Table 2　Dataset example

序号	源微博信息	用户 1	用户 2	……
1	事件编号:3566276356194468 文本:【10 万元可全家移民瑞典?】新华网:欧洲移民又推新渠道,10 万瑞典克朗(约 10 万元 RMB)成立一家公司即可获全家居留许可。无经营范围限制,无员工数量要求,无须收入证明…。 用户编号:1762675215 时间戳:1365750004 L: 1	文本内容:给真? 我要克,将就可以留个学嘛 用户编号: 1972272995 时间戳:1365750076	文本内容: 转发微博 用户编号:1888930763 时间戳:1365750095	……

表2(续)

序号	源微博信息	用户1	用户2	……
2	事件编号:3911246166755503 文本内容:【平价三文鱼时代到了？转基因三文鱼获FDA批准】经过5年全面严格的审查后,三文鱼成为首个登上餐桌的转基因动物。这种三文鱼18个月就能长成成体,传统三文鱼则需要三年…。 用户编号:1850988623 时间戳:1447997216 L: 0	文本内容:吃! 用户编号:3502638475 时间戳:1447997237	文本内容:这个好啊 用户编号:1748891461 时间戳:1447997245	……
……	……	……	……	……

3.2 模型设置

EGSI模型的参数如表3所示,其中卷积层中过滤器个数和过滤器高度分别是100和10,有两层池化层,分别是最大池化层和均值池化层,池化大小为3,E-Dropout应用于全连接层上的隐藏层,防止过拟合。GRU模块中GRU神经网络只有一层循环单元,模型使用交叉验证使得损失函数中的正则化参数λ =0.01,模型学习率设置为0.001,并使用Nadam优化器,模型训练周期设置为30。其他对比试验,相同的深度学习神经网络使用相同的超参数。单层长短期记忆网络(onelevellong short-term memory, LSTM-1)[7]和双层门控循环单元(two level gated recurrent unit, GRU-2)[7]都使用AdaGrad算法作为优化算法,随机森林分类器(random forest classifer, RFC)[24]、CSI[25]和PPC[18]使用Adam算法作为优化算法。

表3 超参数设置

Table 3 Hyperparameter settings

CNN过滤器	CNN过滤器高度	CNN最大池化层大小	CNN均值池化层大小
100	10	3	3
CNN丢失率	CNN输出维度	GRU丢失率	GRU输出维度
0.5	10	0.2	100

3.3 基准方法

使用随机森林分类器RFC[24]、LSTM-1[7]、,GRU-2[7]、PPC[18]和CSI[25]5种基准算法与EGSI进行试验对比,其中RFC是提取用户、语言和结构特征的集成学习算法,LSTM-1和GRU-2是提取语言时间模型特征的单一神经网络深度学习算法,PPC是提取传播路径的神经网络组合模型算法,CSI是提取用户、反馈和文本特征的混合深度学习模型的算法,每种算法简介如下。

(1)RFC。一种装袋集成算法,该算法在训练过程中根据数据集的用户、语言和结构特征来产生多棵决策树,构建出基于随机森林的分类器,每棵决策树会预测输出,然后采用投票机制选择类别众数作为预测结果。

(2)LSTM-1。该网络是循环神经网络RNN的变种,能够学习长期的依赖关系,能够更好提取推文和评论的时间语言模型,对信息进行检测。

(3)GRU-2。循环神经网络LSTM的变种,具有两层神经网络结构的门控循环单元网络,能够提取推文和评论的时间语言模型,对信息进行检测。

(4)PPC。首先将信息的传播路径转化为固定长度多变量序列,然后输入到由卷积神经网络CNN和门控循环单元网络GRU两者结合构建出的传播路径分类器中,分别提取出信息传播路径的局部变化和全局变化,能在信息传播的早期进行检测。

(5)CSI。利用LSTM构建深度学习模块提取信息的文本和反馈特征,然后又根据对用户行为评分构建非深度学习模块来提取信息的用户特征,结合两个模块组成模型对信息进行检测。

除了和基准试验对比,也对EGSI模型进行消融试验,以此评估不同的模块对EGSI模型的影响,分别减少INTERATE模块以外的1块模块后,重新构建模型,得到EGI、GSI、ESI 3个简化模型。EGI模型通过卷积神经网络CNN和门控循环单元GRU,整合了传播路径特征、文本特征、反馈特征。GSI模型通过门控循环单元GRU和对用户行为评分,整合了文本、反馈、用户3个特征。ESI模型通过卷积神经网络CNN和对用户行为评分,整合了传播路径特征和用户征。

本研究所提模型EGSI整合了文本特征、用户特征、反馈特征和传播路径特征,相比基准试验CSI模型,EGSI模型加入了传播路径特征。通过对比各个试验提取特征的数量种类不同,分析在信息传播的早期阶段,不同的特征对模型检测虚假信息效果和时效的影响,也进一步说明EGSI模型中不同模块的作用。

3.4　评估指标

使用 EGSI 模型对虚假信息检测,对于真实信息和虚假信息进行二分类。为了评估检测的效果选择 4 个指标进行分析:准确率 A、精确率 P、召回率 R 和 F_1。A 越高,代表模型对虚假信息检测越准确;P 越高,代表模型对负样本的区分程度越高;R 越高,代表模型对正样本的识别程度越好;F_1 结合了精确率和召回率,F_1 越高,代表模型越稳定。

3.5　结果和分析

3.5.1　检测效果比较

为了与其他论文中的试验模型公平比较,将检测的终止时间设置为 3 个月,即只保存推文发布后 3 个月内转发用户的信息和评论,这样可以让对比试验能提取足够信息进行检测。表 4 展示了检测终止时间为 3 个月时,EGSI 模型和基准试验模型以及 EGSI 简化模型各自在数据集上的检测效果。EGSI 模型性能最优,准确率达到 95.9%,数据集中检测虚假推文和真实的 F_1 指标也分别达到 95.8%和 96%。

表 4　数据集上模型结果比较

Table 4　Comparison of model results on the data set

方法	类别	A	P	R	F_1
LSTM-1	Fake	0.891	0.916	0.898	0.888
	True		0.877	0.899	0.907
GRU-2	Fake	0.910	0.986	0.956	0.914
	True		0.864	0.864	0.906
RFC	Fake	0.930	0.944	0.941	0.932
	True		0.916	0.920	0.928
PPC	Fake	0.922	0.896	0.962	0.923
	True		0.949	0.889	0.918
CSI	Fake	0.953	0.959	0.950	0.951
	True		0.946	0.957	0.955
GSI	Fake	0.947	0.964	0.930	0.947
	True		0.926	0.961	0.944
EGI	Fake	0.937	0.918	0.955	0.926
	True		0.948	0.906	0.936
ESI	Fake	0.913	0.961	0.958	0.915
	True		0.866	0.873	0.911
EGSI	**Fake**	**0.959**	**0.974**	**0.945**	**0.960**
	True		**0.943**	**0.973**	**0.958**

在基准试验之中,LSTM-1 和 GRU-2 检测准确率分别只有 89.1%和 91%,相比其他试验模型,二者检测虚假信息效果较差,两者都是提取信息文本之间的时间语言特征,但 GRU-2 通过循环层堆叠的方式,构建两层门控循环单元,增大了网络容量,以此增强了神经网络的拟合能力,可以更有效地提取信息特征,相比之下 LSTM-1 只有一层长短期记忆层,因此,LSTM-1 效果较 GRU-2 较差。

PPC 模型检测虚假信息的准确率达到 92.2%,它首先将传播路径构建为固定长变量时间序列,然后通过 CNN 和 GRU 两个神经网络,分别提取表示传播路径的序列的局部变化和全局变化,然后整合二者提取的信息输入到构建的传播路径分类器中进行检测,传播路径相比信息文本和用户评论更难被操纵,因而提取的特征,具有更少的无用或者错误信息,所以相较 LSTM-1 和 GRU-2,具有更好的检测效果。

CSI 模型效果仅次于 EGSI 模型,准确率达到 95.3%。因为它对文本进行时间分片后进行文本转向量,更好地提取了推文和转发回复之间的时间联系,然后通过 LSTM 提取到文本特征和反馈特征,和从用户行为中提取用户特征,整合了信息 3 个最基本特征进行检测,当检测的终止时间较长时,CSI 模型能够提取足够的有用信息,从而能有较好的检测效果。

相较基准试验 CSI 模型和 EGSI 模型,添加了 EXTRACT 模块,模块使用卷积神经网络 CNN 提取用户的传播路径特征,相较 CSI 的 C 模块使用 LSTM 神经网络提取文本特征和卷积特征,CGSI 模型在 GRU 模块中使用 GRU 神经网络提取文本特征和卷积特征,当检测终止时间设置为 60 min 时,转发回复的用户较少,GRU 相较 LSTM 神经网络在处理规模较小的数据集时一般性能较好。

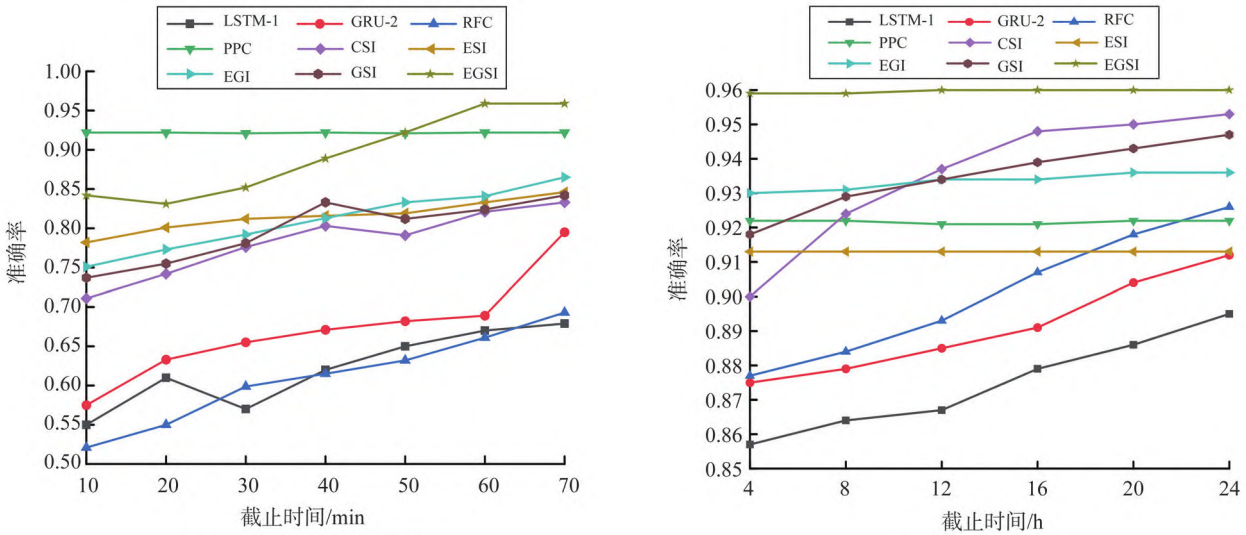
分析 EGSI 模型的 3 个简化模型,由图 2 可知,ESI 检测效果能更早稳定并且达到最优,但是准确率最低,只有 91.3%。GSI 在简化模型中效果最好,准确率达到 94.7%,但它在信息传播的早期阶段,检测效果较差。GSI 和 EGI 的检测效果也优于大部分基准试验。由此可以假设具有 EXTRACT 模块的模型检测效果能较快的趋于稳定,可以在信息传播的早期进行检测,使得模型检测的时间效率明显提升,而具有 GRU 模块的模型准确率相比较高,检测效果更好。

3.5.2　检测时效比较

本研究目标是在确保准确率较高的前提下,想要尽可能早地检测出虚假信息,因此对检测虚假信息的时效性进行分析。将检测的终止时间设置为 24 h。图 2 展示了各个试验模型检测虚假信息准确率和终止时间的关系,由图 2(a)可知,PPC 检测效果在信息传播 10 min 后进行检测,效果已经稳定,并且达到最佳,虽然它检测信息的时效性很高,但是它的准确率只有92.2%。至于其他基准试验模型检测信息的时效性都差于 PCC 和 EGSI 模型,因为在信息传播的早期阶段,用户转发评论的数量较

少，使得基准试验没有提取到足够的有用信息。图 3 展示了 EGSI 模型在信息传播早期，准确率与终止时间的关系。由图 2（b）和图 3 可以看出 EGSI 模型在信息传播 60 min 后进行检测，检测效果稳定并且达到最佳。虽然 PPC 在时效性上优于 EGSI 模型，但是准确率比起 EGSI 模型低了 3.8%，检测效果上有明显差距。并且 PPC 模型的准确率为 92.2%，对于社交平台庞大的信息发布数量而言，依旧会有数量巨大的虚假信息无法检测出，所以检测虚假信息模型需要在保证准确率的情况下提高时效性才更有意义。

结合图 2 和图 3，比较 EGSI 模型和 CSI 模型，两者有着相近的结构，但 EGSI 模型比起 CSI 模型多整合了传播路径特征，CSI 模型需要 3 个月的检测终止时间才能达到最佳检测效果，EGSI 模型只需要推文发布 60 min 就可以进行检测，准确度能达到 95.9%，检测效果略优于 CSI 模型，且检测信息的时效性明显提升。一般而言，信息往往需要数个小时的时间才会广泛传播，因而在 60 min 时检测出虚假信息，EGSI 模型更具有实际意义。



(a) 70 min检测结果对比图　　(b) 24 h检测结果对比图

图 2　检测结果对比图

Fig.2　Comparison of test results

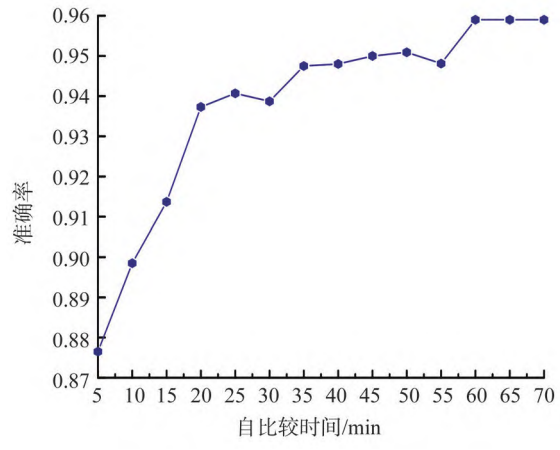


图 3　EGSI 检测信息的时效比较图

Fig.3　Comparison of EGSI detection efficiency

3.6　模型超参数分析

EGSI 模型使用卷积神经网络 CNN 和门控循环单元 GRU 来提取信息的特征，2 种神经网络的组合可能导致训练中出现过拟合现象，为此将验证模型的 4 个超参数（$D_{\text{Dropout-E}}$、$D_{\text{Dropout-G}}$、正则化参数 λ 和优化器的学习率 β）对检测效果的影响，如图 4 所示。

丢失率 D_{Dropout} 是一种在训练过程中随机删除神经元的概率，$D_{\text{Dropout-G}}$、$D_{\text{Dropout-E}}$ 分别表示 GRU 模块和 EXTRACT 模块使用的丢失率。模型通过损失函数加上 $L2$ 范数进行权值衰减来抑制过拟合，λ 则是 $L2$ 范数的超参数，λ 越大，则对大的权重施加的惩罚就越重。优化器的学习率 β 影响神经网络优化的速度，用来减小神经网络的损失。在这些超参数默认通常取值范围内进行测试，图 4（a）表示 $\beta=0.001$，$D_{\text{Dropout-G}}=0.2$，$D_{\text{Dropout-E}}=0.5$ 条件下 λ 变化示意图；图 4（b）表示 $\lambda=0.01$，$D_{\text{Dropout-G}}=0.2$，$D_{\text{Dropout-E}}=0.5$ 条件下，β 的变化对准确率的影响；图 4（c）表示 $\beta=0.001$，$\lambda=0.01$，$D_{\text{Dropout-E}}=0.5$ 条件下，$D_{\text{Dropout-G}}$ 变化对准确率的影响；图 4（d）表示 $\beta=0.001$，$\lambda=0.01$，$D_{\text{Dropout-G}}=0.2$ 条件下，$D_{\text{Dropout-E}}$ 对准确率的影响，可以看出 4 个超参数对模型检测准确率的影响明显。因此，EGSI 模型超参数对于检测虚假信息也具有较大影响。

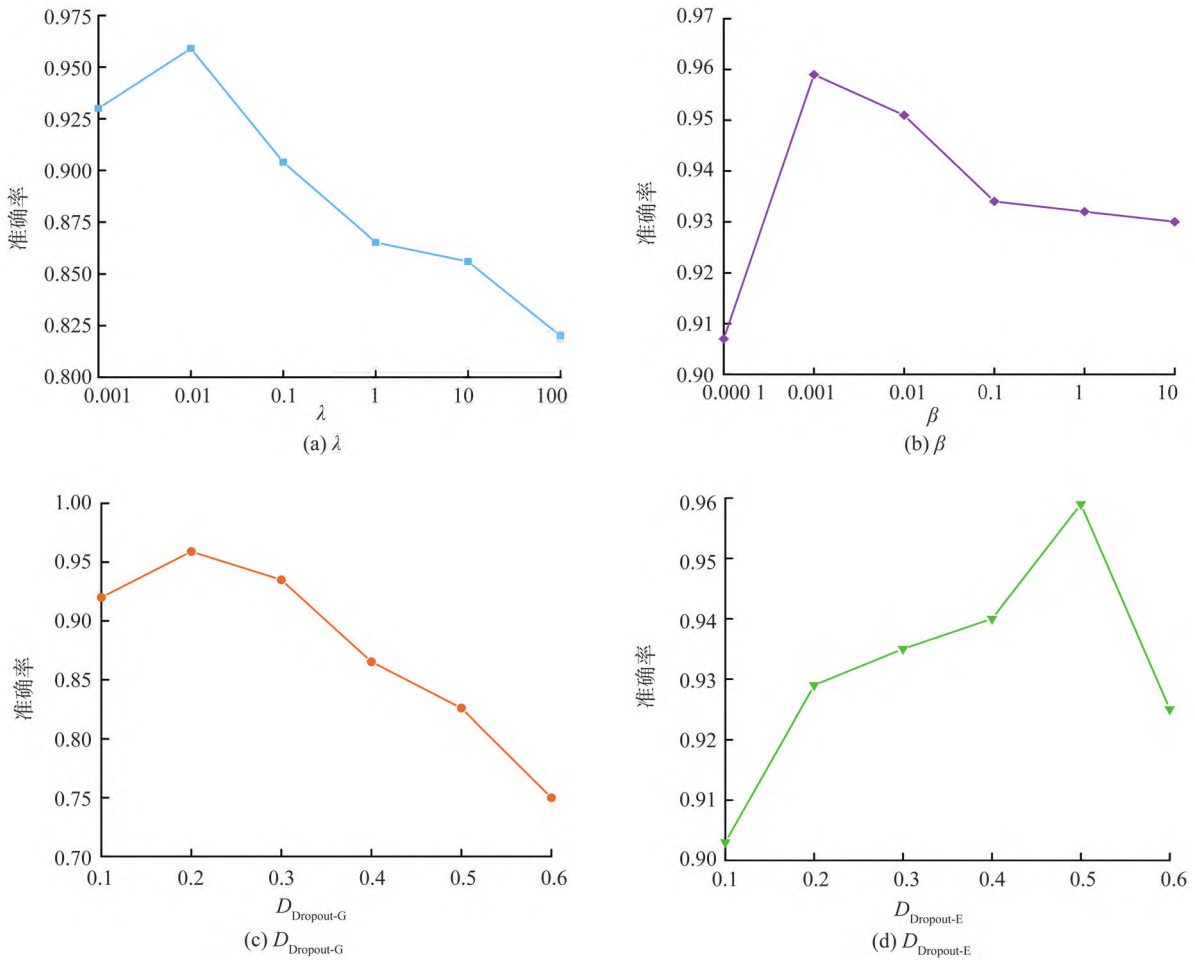


图 4　λ、β、$D_{\text{Dropout-G}}$和 $D_{\text{Dropout-E}}$超参数设置影响图

Fig.4　Hyperparameter settings impact diagram of λ, β, $D_{\text{Dropout-G}}$ and $D_{\text{Dropout-E}}$

3.7　训练集比例分析

为了研究 EGSI 模型检测虚假信息与事件样本数量的关系,通过改变训练集大小进行训练,观察准确率变化。将数据集的 5%作为验证集,15%作为测试集。分别取数据集的 10%、20%、40%、60%和 80%作为训练集,由图 5 可知,在训练集为数据集 10%的时候,各个模型检测效果差异最大,EGSI 模型效果依旧最好,因此 EGSI 模型更加轻巧,可以使用更少的训练集更容易训练。同时可以假设即使在训练集规模较小时,EGSI 模型由于提取信息的 4 个特征,可以提取到较多的特征信息进行检测,因此检测效果较好。随着训练集规模增大,各个模型能提取的特征信息也随之增多,因而检测效果差异也随之减小,也可以验证假设成立。同时 EGSI 模型不管训练集大小如何变化,检测效果依旧优于其他模型,并当数据集的 80%设置为训练集时,效果达到最优,差异也随之减小,也可以验证假设成立。同时 EGSI 模型不管训练集大小如何变化,检测效果依旧优于其他模型,并当数据集的 80%设置为训练集时,效果达到最优。

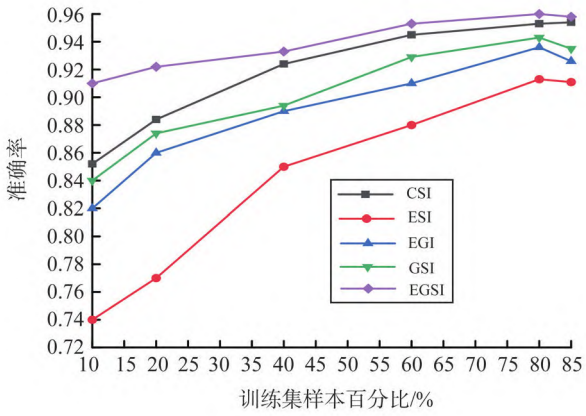


图 5　准确性与训练集比例关系图

Fig.5　Relationship between accuracy and training set ratio

4　结语

本试验研究了信息传播的早期检测问题,提出了一个由 4 个模块组成的混合深度模型,EXTRACT 通过卷积神经网络提取信息的传播路径

特征,GRU通过门控循环单元捕获信息的文本特征和反馈特征,SCORE基于用户行为挖掘用户特征,INRERATE整合以上特征并预测出信息事件类标。在微博数据集上的试验结果表明,本研究提出的模型在准确率和时效性上取得较好的平衡,可以在信息传播的早期较为准确地检测出虚假信息。同时计划将来在模型中融入半监督学习,以处理社交媒体上大量未标记的新闻故事。

参考文献:

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

[2] QAZVINIAN V, ROSENGREN E, RADEV D R, et al. Rumor has it: identifying misinformation in microblogs [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK: EMNLP, 2011: 1589-1599.

[3] GUPTA A, KUMARAGURU P, CASTILLO C, et al. TweetCred: real-time credibility assessment of content on twitter[C]//Proceedings of the International Conference on Social Informatics. Barcelona, Spain: SocInfo, 2014: 228-243.

[4] POPAT K. Assessing the credibility of claims on the Web [C]//Proceedings of the 26th International Conference on World Wide Web. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017: 735-739.

[5] YANG F, YU X, LIU Y, et al. Automatic detection of rumor on Sina Weibo [C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. Sydney, Australia: ACM, 2015: 1-7.

[6] ZHAO Z, RESNICK P, MEI Q. Enquiring minds: early detection of rumors in social media from enquiry posts [C]//Proceedings of the 24th International Conference on World Wide Web. Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015: 1395-1405.

[7] 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. New York, USA: AAAI Press, 2016: 3818-3824.

[8] MA J, GAO W, WONG K F. Detect rumors in microblog posts using propagation structure via kernel learning[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: IJCAI, 2017: 708-717.

[9] JIN F, DOUGHERTY E R, SARAF P, et al. Epidemiological modeling of news and rumors on Twitter[C]// Proceedings of the 7th Workshop on Social Network Mining and Analysis. Chicago, USA: ACM, 2013: 1-9.

[10] WU K, YANG S, ZHU K Q. False rumors detection on Sina Weibo by propagation structures[C]//Proceedings of the 31st IEEE International Conference on Data Engineering. Seoul, Korea: ICDEW, 2015: 651-662.

[11] CHEN T, LI X, YIN H, et al. Call attention to rumors: deep attention based recurrent neural networks for early rumor detection[C]// Proceedings of Trends and Applications in Knowledge Discovery and Data Mining. Melbourne, Australia: LNCS, 2018: 40-52.

[12] LI Q, HU Q, LU Y, et al. Multi-level word features based on CNN for fake news detection in cultural communication[J]. Personal and Ubiquitous Computing, 2020, 24(2): 259-272.

[13] CHOWDHURY R, SRINIVASAN S, GETOOR L. Joint estimation of user and publisher credibility for fake news detection[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York, America: CIKM, 2020: 1993-1996.

[14] BALESTRUCCI A, NICOLA R D. Credulous users and fake news: a real case study on the propagation in Twitter[C]//Proceedings of 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS). Bari, Italy: IEEE, 2020: 1-8.

[15] HAMDI T, SLIMI H, BOUNHAS I, et al. A hybrid approach for fake news detection in Twitter based on user features and graphembedding[C]//Proceedings of Distributed Computing and Internet Technology. Bhubaneswar, India: ICDCIT, 2020: 266-280

[16] JIANG S, CHEN X, ZHANG L, et al. User-characteristic enhanced model for fake news detection in social media[C]//Proceedings of Natural Language Processing and Chinese Computing. Dunhuang, China: NLPCC, 2019: 634-646.

[17] SAMPSON J, MORSTATTER F, WU L, et al. Leveraging the implicit structure within social media for emergent rumor detection[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. Turin, Italy: CIKM, 2016: 2377-2382.

[18] LIU Y, WU Y F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Louisiana, USA: AAAI, 2018: 354-361.

(下转第109页)

computer Systems, 2019, 40(8): 1595-1560.
[22] LUO S, MIAO D Q, ZHANG Z F, et al. A neighborhood rough set model with nominal metric embedding [J]. Information Sciences, 2020, 520:373-388.
[23] 盛魁,王伟,卞显福,等.混合数据的邻域区分度增量式属性约简算法[J].电子学报,2020,48(4):682-696.
SHENG Kui, WANG Wei, BIAN Xianfu, et al. Neighborhood discrimination incremental attribute reduction algorithm for mixed data [J]. Journal of Electronics, 2020, 48(4): 682-696.
[24] YANG X, CHEN H, LI T, et al. Neighborhood rough sets with distance metric learning for feature selection [J]. Knowledge-Based Systems, 2021, 224: 107076.
[25] 范雪莉,冯海泓,原猛. 基于互信息的主成分分析特征选择算法[J]. 控制与决策,2013, 28(6): 915-919.
FAN Xueli, FENG Haihong, YUAN Meng. Principal component analysis feature selection algorithm based on mutual information[J]. Control and Decision, 2013, 28(6): 915-919.
[26] 梁海龙,谢珺,续欣莹. 新的基于区分对象集的邻域粗糙集属性约简算法[J].计算机应用,2015,35(8): 2366-2370.
LIANG Hailong, XIE Jun, XU Xinying. A new attribute reduction algorithm for neighborhood rough sets based on distinguishing object sets [J]. Computer Applications, 2015, 35(8): 2366-2370.
[27] ZHANG Y Z, WANG Y Q. Research on classification model based on neighborhood rough set and evidence theory[J]. Journal of Physics: Conference Series, 2021, 1746(1):012018.
[28] HU M, TSANG E C C, GUO Y T, et al. A novel approach to attribute reduction based on weighted neighborhood rough sets[J]. Knowledge-Based Systems, 2021, 220: 106908.

(编辑:李骏)

(上接第 98 页)

[19] QIAN F, GONG C, SHARMA K, et al. Neuraluser response generator: fake news detection with collective user intelligence [C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Washington, USA: IJCAI, 2018: 3834-3840.
[20] CASTILLO C, EL-HADDAD M, PFEFFER J, et al. Characterizing the life cycle of online news stories using social media reactions [C]//Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. Baltimore, Maryland: CSCW, 2014: 211-223.
[21] FRIGGERI A, ADAMIC L A, ECKLES D, et al. Rumor cascades [C]//Proceedings of International AAAI Conference on Web and Social Media. Ann Arbor, USA: ICWSM, 2014: 1-13.
[22] KUMAR S, WEST R, LESKOVEC J. Disinformation on the web: impact, characteristics, and detection of wikipediahoaxes[C]//Proceedings of the 25th International Conference on World Wide Web. Quebec, Canada: International World Wide Web Conferences Steering Committee, 2016: 591-602.
[23] STARBIRD K, MADDOCK J, ORAND M, et al. Rumors, false flags, and digital vigilantes: misinformation on Twitter after the 2013 Boston Marathon Bombing [C]//Proceedings of the IConference 2014. Illinois, America: ISchool, 2014: 654-662.
[24] KWON S, CHA M, JUNG K. Rumor detection over varying time windows[J]. PLOS ONE, 2017, 12(1): 1-19.
[25] RUCHANSKY N, SEO S, LIU Y. CSI: ahybrid deep model for fake news detection[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: CIKM, 2017: 797-806.

(编辑:李骏)