小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI: 10.20009/j.cnki.21-1106/TP.2021-0277
2022 年 12 月 第 12 期
Vol. 43 No. 12 2022

融合多元用户特征和内容特征的微博谣言实时检测模型

黄学坚[1,2],王根生[1,2,3],罗远胜[2],闵 潞[1],吴小芳[2],李志鹏[2]

[1](江西财经大学 人文学院,南昌 330013)
[2](江西财经大学 计算机实践教学中心,南昌 330013)
[3](江西财经大学 国际经贸学院,南昌 330013)
E-mail: wgs74@126.com

摘 要: 针对目前基于单文本语义特征深度学习的微博谣言实时检测模型泛化能力不足的问题,提出一种融合多元用户特征和内容特征的实时检测模型. 首先,在传统用户基本特征和内容统计特征的基础上,利用用户的历史行为数据,挖掘用户理性值和用户专业度两个深层次特征; 然后,基于词向量和带有注意力机制的双向 GRU 神经网络构建文本语义特征学习模型; 最后,采用分层特征级联和全连接的方式进行特征融合,把融合特征输入分类模型进行训练. 实验结果表明,该模型的检测准确率达到了 91.74%,相比其他只关注文本语义特征的深度学习实时检测模型具有更好的识别效果,相比于其他改进型的实时检测模型 F1-Measure 值也提高了 2.19%.

关 键 词: 微博谣言; 实时检测; 特征融合; 深层特征; 深度学习
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2022)12-2518-10

Weibo Rumors Real-time Detection Model Based on Fusion of Multi User Features and Content Features

HUANG Xue-jian[1,2], WANG Gen-sheng[1,2,3], LUO Yuan-sheng[2], MIN Lu[1], WU Xiao-fang[2], LI Zhi-peng[2]
[1](School of Humanities, Jiangxi University of Finance and Economic, Nanchang 330013, China)
[2](Computer Practice Teaching Center, Jiangxi University of Finance and Economics, Nanchang 330013, China)
[3](School of International Trade and Economics, Jiangxi University of Finance and Economic, Nanchang 330013, China)

Abstract: In order to solve the problem of low generalization ability of Weibo rumor real-time detection model based on deep learning of text semantic features, a real-time detection model fusion of multi user features and content features is proposed. First of all, in addition to the traditional user basic features and content statistical features, two implicit features of user rationality and user professionalism are mined based on user's historical behavior data; Then, a text semantic feature learning model is constructed based on word vector and bidirectional GRU neural network with attention mechanism; Finally, the hierarchical cascade and full connection are used for feature fusion, and the fused features are input into the classification model for training. The experimental results show that the accuracy of the model is 91.74%, which is better than other deep learning real-time detection models based on text semantic features. Compared with other improved real-time detection models, F1-Measure value also increased by 2.19%.

Key words: Weibo rumors; real-time detection; feature fusion; implied features; deep learning

1 引 言

随着 Web2.0 的快速发展以及移动互联网的普及,以微博为代表的社交媒体成为人们获取和分享信息的一个重要平台. 根据中国互联网络信息中心(CNNIC) 2020 年发布的第 45 次《中国互联网络发展状况统计报告》显示,截止 2020 年 3 月我国网民规模达到 9.04 亿,互联网普及率达 64.5%,网络新闻用户规模达到 7.31 亿,占网民整体 80.9%,微博使用率达到 42.5%. 微博等社交媒体方便人们信息分享的同时,也为谣言的传播提供了新的温床. 根据微博辟谣官方发布的数据显示,2019 年微博站方有效处理不实信息 77742 条. 2020 年初,关于新冠肺炎的谣言频出,如 "钟南山建议盐水漱口防病毒"、"板蓝根 + 熏醋可预防新型肺炎"、"国家不再对新冠肺炎病人免费治疗"等谣言给疫情防疫工作带了很多困扰. 社交媒体中,谣言具有传播速度快、影响范围广、监测难度大、危害程度深等特点,谣言的产生和传播不仅妨碍了人们对社交媒体的有效利用,而且可能造成民众的误解、引发负面情绪、扰乱社会秩序、甚至影响社会稳定和国家安全[1].

为了控制谣言的传播,微博官方建立了微博社区管理中心,受理网民对不实信息的举报; 中央网信办建立了中国互联

收稿日期: 2021-04-02 收修改稿日期: 2021-04-22 基金项目: 国家自然科学基金项目(72061015,61562031) 资助; 江西省教育厅科技项目(GJJ200539) 资助. 作者简介: 黄学坚,男,1990 年生,硕士,讲师,CCF 会员,研究方向为数据挖掘、文本分析; 王根生,男,1974 年生,博士,副教授,CCF 会员,研究方向为网络舆情、数据挖掘; 罗远胜,男,1976 年生,博士,高级工程师,CCF 会员,研究方向为自然语言处理; 闵 潞,女,1982 年生,硕士,助理研究员,研究方向为网络舆情; 吴小芳,女,1990 年生,硕士,工程师,研究方向为数据挖掘; 李志鹏,男,1988 年生,硕士,实验师,研究方向为数据挖掘.

网联合辟谣平台,为广大群众提供谣言举报、查证等功能; 腾讯推出了专业事实查证平台“较真”,对各类假新闻、谣言等进行查证. 这些措施对控制谣言的传播、降低谣言的危害起到了一定作用. 然而,这些网站或系统在识别谣言时主要依赖于民众举报和人工验证的方式,不但需要耗费大量的人力和财力,而且在谣言识别上存在较大的时间滞后问题. 因此,研究社交网络谣言自动识别模型,在谣言散布后尽可能短的时间内识别出谣言,对遏制谣言的传播具有重要的意义.

2　相关研究

谣言的传播生命周期粗略可划分为潜伏期、爆发期和消亡期[2],所以按检测时间属性划分,可分为滞后检测、早期检测和实时检测 3 类[3]. 滞后检测主要针对谣言爆发期后的历史数据集进行检测,因历史数据包含了丰富的谣言传播特征(如评论、转发、点赞等数据),从而可用的检测特征相对较多,是初期谣言检测研究的主要方向. 例如,2011 年 Castillo 等人首次针对 Twitter 上的谣言数据进行整理,提出基于用户特征、内容特征、主题特征和传播特征的决策树谣言分类模型[4]; 2012 年 Yang 等人首次针对新浪微博上的谣言数据进行分析,在已有的特征基础上提出了发布信息客户端类型和事件发生位置两个新特征,并使用 SVM 分类器构建谣言识别模型[5]; 贺刚等人认为 Yang 提出的浅层文本特征不能有效的区分谣言和非谣言,因此提出了符号特征、链接特征、关键词分布特征和时间差等 4 个新特征[6]; 曾子明等人以 2016 年雾霾谣言为例,在用户特征和传播特征的基础上利用 LDA 主题模型挖掘微博文本的主题分布特征,并利用随机森林算法进行谣言识别模型的训练[7]. 这一类的研究主要基于“人工特征工程 + 传统统计机器学习”的方式构建谣言检测模型,由于依赖的数据丰富,实验过程中也能获得较好的检测效果,但实际应用中不能尽早地识别谣言,缺少实用价值. 所以,如何构建谣言及早检测模型成为近年来的研究热点.

早期检测主要针对尚处于潜伏期内的谣言进行识别,一般基于谣言早期传播过程中的时序数据进行建模. 例如,Wu 等人基于谣言的转发数据,构建谣言传播树形结构,使用基于图核函数的 SVM 分类器进行谣言检测[8]; Ma 等人基于谣言传播过程中的转发时序数据,构建基于循环神经网络(Recurrent Neural Network,RNN)的谣言检测模型[9]; Chen 等人提出注意力机制与循环神经网络相结合的谣言检测模型,使模型更加关注于转发时序数据中具有谣言特征的部分[10]; 谢柏林等人提出一种基于把关人行为的微博虚假信息检测方法,利用隐半马尔可夫模型刻画信息转发者和评论者对真实信息的把关行为,基于此来识别微博上流行的虚假信息[11]; 刘知远等人利用谣言传播过程中的怀疑和反驳信息,提出可信检测点的概念,基于深度神经网络构建谣言早期检测模型[12]; 廖祥文等人把谣言传播过程中的转发信息按时间段进行分割,输入带有注意力机制的双向门控循环神经网络(Gate Recurrent Unit,GRU)中,构建基于分层注意力网络的谣言检测模型[13]; 李力钊等人利用 Doc2vec 把谣言传播过程中的评论向量化,通过卷积神经网络(Convolutional Neural Networks,CNN)中的卷积层学习特征表示,将特征序列输入 GRU 中进行谣言检测[14]. 这一类的研究从基于人工特征工程的统计机器学习方法逐渐发展到基于语义特征的深度学习方法,大部分检测模型只要利用 10%–30% 的历史传播数据就可以获得 80% 以上的检测准确率,检测延时在 12–24 小时左右,相比于滞后检测具有较高的实用价值.

实时检测即信息一经发布就立即开始检测其是否为谣言信息,不依赖于任何的传播动态信息,也称为谣言的冷启动检测问题,是当今谣言检测研究的难点问题[3],部分学者也进行了相关探索研究. 例如,Ajao 等人提出卷积神经网络和长短时记忆网络(Long-Short Term Memory,LSTM)相结合的模型,自动提取 Twitter 中虚假谣言的语义特征,实现谣言的实时检测[15]; 李奥等人提出一种生成对抗网络模型用于谣言检测,通过对抗网络生成器和判别器的相互促进作用,强化谣言文本特征的学习[16]. 这一类的方法把谣言识别视为单文本分类问题,检测方法重点关注于谣言文本的语义信息[17]. 然而,用户散布的谣言可能不具有典型的谣言文本特征,单纯的文本分类有时并不能取得很好的效果. 针对这个问题,马鸣等人将待检测样本和官方谣言库中的样本进行相似度计算,将其值和传统的用户统计特征、内容统计特征进行融合,输入 SVM 分类模型进行谣言检测[18]; 尹鹏博等人结合用户属性和微博文本,提出基于卷积神经网络和长短期记忆网络的谣言检测模型[19]. 通过研究分析发现,在单文本内容特征的基础上融入更多的辅助特征是提升实时谣言检测效果的重要手段. 所以,本文基于已有的研究基础,提出融合多元用户特征和内容特征的谣言实时检测模型,通过结合多元异构信息以弥补单一文本信息的不足,提高谣言实时检测的准确率.

3　检测模型构建

通常在社交媒体谣言检测中用到的特征主要有用户特征、内容特征和传播特征[20],而实时检测的谣言处于刚散布阶段,还不存在谣言的传播信息,所以只能从用户和内容信息挖掘出识别特征. 在用户特征挖掘上,通过谣言用户和非谣言用户的属性差异选择用户基本特征,并利用用户的历史行为数据,挖掘用户理性值和用户专业度两个深层次特征; 在内容特征挖掘上,构建双向 GRU 神经网络 + 注意力机制的文本语义特征学习模型,并统计符号、表情、URL 等内容统计特征. 检测模型如图 1 所示.

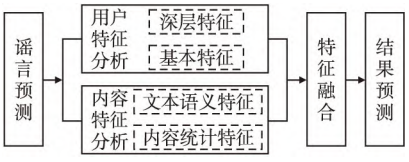

图 1　检测模型

Fig. 1　Detection model

在图 1 中,采用分层特征级联和全连接的方式进行特征融合,把融合特征输入分类模型进行训练,预测分类结果.

3.1　用户特征分析

3.1.1　用户基本特征

刘雅辉等人通过对谣言用户和非谣言用户的基本特征进

行分析发现五点差异: 谣言用户一般不会利用真实照片作为头像、话题型用户名的用户更具可信度、女性相比男性散布谣言的概率更大、用户所在地的差异影响谣言的识别能力、非认证用户比认证用户散布谣言的可能性更大[20]; Morris 等人研究发现用户的社交关系网络特征可以作为谣言用户的判断依据,粉丝数远小于关注数的用户更可能散布谣言信息[21]; Castillo 等人研究发现用户注册的时间越长、发布的信息越多其谣言识别能力越强,散布谣言的概率越小[4]. 所以本文基于这些研究结果,构建用户基本特征选项,并对相关特征表示进行数字化处理,具体特征选项如表 1 所示.

表 1 用户基本特性选项

Table 1 User basic feature options

编号	特征描述	特征表示
*FU*1	头像是否是真实照片	0: 否,1: 是
*FU*2	用户名是否是话题型	0: 否,1: 是
*FU*3	用户性别	0: 女,1: 男
*FU*4	用户所在地	数字
*FU*5	是否是认证用户	0: 否,1: 是
*FU*6	粉丝数	数值
*FU*7	关注数	数值
*FU*8	注册时长(多少个月)	数值
*FU*9	发布的历史微博数	数值
*FU*10	VIP 等级	数值

除了这些基本特征外,如何挖掘用户深层特征对提高谣言检测的准确率具有重要作用[22],所以本文基于用户的历史行为数据,提出用户理性值和专业度两个深层特征.

3.1.2 用户深层特征

1) 用户理性值

"流言止于智者",智者是具有独立思考判断的理性人,他们不轻易相信谣言,也不轻易传播谣言. 通过对数据分析发现,理性用户发布的微博一般不会带有强烈的个人情感,微博行文客观公正,不容易引起广泛关注,用户评论态度也比较中立; 而非理性用户发布的微博一般喜欢附加个人情感,微博行文主观臆断,更容易激起人们的关注,用户评论也更有争议性,即评论情感倾向程度明显[23,24]. 所以文本提出,通过对用户的历史微博文本情感和评论情感进行分析,计算用户的理性值,计算方法如公式(1) 所示:

$$Rat_u = \frac{1}{n}\sum_{i=1}^{n}(Senti_i + Argue_i) \tag{1}$$

其中,Rat_u 表示用户 u 的理性值,n 表示用户 u 发布的历史微博数,$Senti_i$ 和 $Argue_i$ 分别表示微博 i 的情感度和争议度,其计算分别如公式(2) 、公式(3) 所示:

$$Senti_i = \sum_{w_t \in SentiDic}(|SentiDegree_{w_t} * AdvDegree_{w_t}|) \tag{2}$$

$$Argue_i = \log|neg + pos| * \frac{pos^2 + neg^2}{pos * neg} \tag{3}$$

在公式(2) 中,w_t 表示微博文本分词后的词语,$SentiDic$ 表示包含情感值的情感词库,$SentiDegree_{w_t}$ 表示 w_t 的情感值,$AdvDegree_{w_t}$ 表示修饰 w_t 的程度副词的程度值,微博的情感度计算不区分情感极性,取绝对值进行相加. 在公式(3) 中,neg 和 pos 分别表示评论中正面情感评论数和负面情感评论数,中立情况不考虑,情感评论数越多微博的争议度越大; 情感评论数相同的情况下,正负情感评论数越不均衡,微博的争议度越大.

2) 用户专业度

用户经常发布某一主题的微博,说明用户对该主题有一定的认识,发布该主题下的微博更具可信度. 文本基于 LDA(Latent Dirichlet Allocation) 主题模型[25],挖掘微博的潜在主题,提出基于主题相似度的用户专业度计算,计算方法如公式(4) 所示:

$$Pro_{u,w} = \sum_{i=1}^{n}\cos(\vec{\theta}_i, \vec{\theta}_w) \tag{4}$$

其中,$Pro_{u,w}$ 表示用户 u 对待检测微博 w 的专业度,n 表示用户 u 发布的历史微博数,$\vec{\theta}_i = (p_1, p_2, \cdots, p_k)$ 表示历史微博 i 的主题概率分布向量,$\vec{\theta}_w = (p_1, p_2, \cdots, p_k)$ 表示待检测微博 w 的主题概率分布向量,k 为设定的主题数,p_k 为主题 k 的概率. 用户发布的历史微博 i 和待检测微博 w 的主题概率分布越相似,$\cos(\vec{\theta}_i, \vec{\theta}_w)$ 余弦值越大,用户对微博 w 的专业度越高.

3.2 内容特征分析

3.2.1 文本语义特征

为了挖掘文本深层语义特征,文本构建双向 GRU 神经网络 + 注意力机制的语义特征学习模型. 双向 GRU 使序列某点的输出不仅依赖与之前的信息,还依赖与未来的信息,融合上下文内容生产输出,符合人类理解文本的方式. 注意力机制让模型更加关注于文本中具有谣言模式的部分. 具体语义特征学习模型如图 2 所示.

1) 语义特征学习模型算法

语义特征学习模型的算法如下:

Step 1. 将文本进行分词和预处理(去除停用词、表情、符号等) ,利用词向量(Word2vec) 进行词表示,构建文本词序列向量表示 $W = [w_1, w_2, \cdots, w_t]$,$w_i$ 为分词预处理后的第 i 个词的向量表示,$w_i = [v_1, v_2, \cdots, v_k]$,$k$ 为词向量的维度.

Step 2. 依次把文本词序列向量作为双向 GRU 网络的输入.

Step 3. 把双向 GRU 状态输入到全连接层,计算输出结果 y_t,计算过程如公式(5) 所示:

$$y_t = \sigma(W_g[\overrightarrow{h}_t, \overleftarrow{h}_t] + b_y) \tag{5}$$

其中,y_t 表示序列中第 t 个节点的输出,σ 表示激活函数,W_g 表示全连接层参数矩阵,b_y 表示偏置项参数,$\overrightarrow{h}_t$ 表示前向传播 GRU 中 t 时刻的状态,$\overleftarrow{h}_t$ 表示后向传播 GRU 中 t 时刻的状态.

Step 4. 利用注意力机制为每个节点的输出赋予不同的权重值,计算最终文本语义特征 FC_s,计算过程如公式(6) 所示:

$$FC\_s = \sum_{i=1}^{t}\alpha_t y_t \tag{6}$$

其中,α_t 表示 y_t 的权重,其计算过程如公式(7) ~ 公式(8) 所示:

$$u_t = \tanh(W_w y_t + b_w) \tag{7}$$

$$\alpha_t = \frac{\exp(u_t^T u_w)}{\sum_{i=1}^{t}\exp(u_t^T u_w)} \tag{8}$$

其中,W_w 表示神经网络连接参数,b_w 表示偏置项参数,u_w 表示随机初始化权重.

2) GRU 单元结构

图 2 模型中的 GRU 单元结构如图 3 所示. GRU 单元结构中 h_t 计算过程如公式(9) ~公式(12)所示.

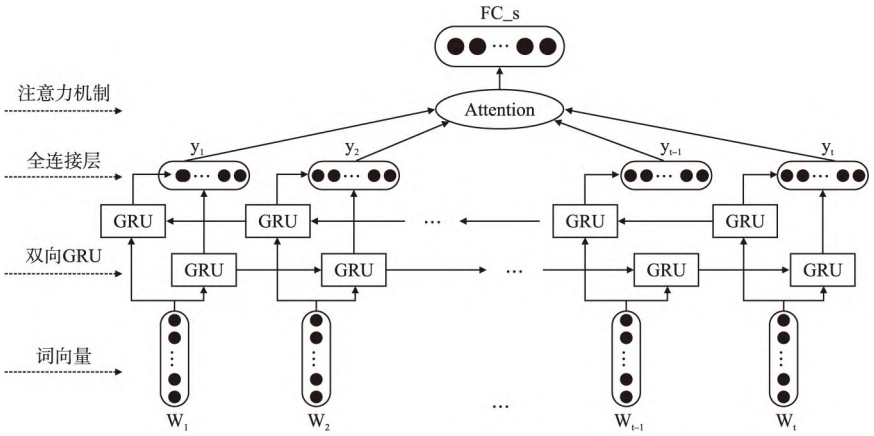

图 2　语义特征学习模型

Fig. 2　Semantic feature learning model

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{9}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{10}$$

$$\overline{h_t} = \tanh(W_{\overline{h}} \cdot [r_t * h_{t-1}, x_t] + b_a) \tag{11}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \overline{h_t} \tag{12}$$

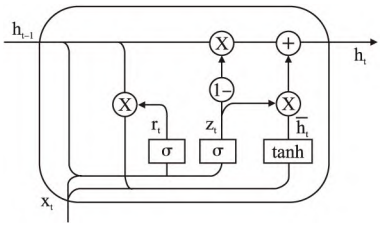

图 3　GRU 单元结构

Fig. 3　GRU unit structure

3.2.2　内容统计特征

在挖掘文本语义特征时,通常忽略了文本内容中的符号、表情、URL等信息,而这些信息对谣言的识别也具有一定的辅助作用. 例如,一些谣言为了骗取用户流量通常会在文本中加入 URL 链接. 本文为了弥补单一文本语义特征的不足,提出融合相关内容统计特征来丰富谣言检测内容特征. 具体内容统计特征如表 2 所示.

表 2　内容统计特征

Table 2　Statistical characteristics of content

编号	特征描述	特征表示
*FC*1	是否有#标识的话题	0: 无 1: 有
*FC*2	包含的 URL 数量	数值
*FC*3	包含的表情符号数量	数值
*FC*4	包含的@ 符号数量	数值
*FC*5	包含的? 号数量	数值
*FC*6	包含的! 号数量	数值
*FC*7	是否附有图片或视频	0: 无,1: 有

3.3　特征融合

用户特征和内容统计特征都是单值类型,文本语义特征为多维向量类型,为了使他们保持一致,文本采用分层级联 + 全连接的方式进行特征融合,具体融合模型如图 4 所示.

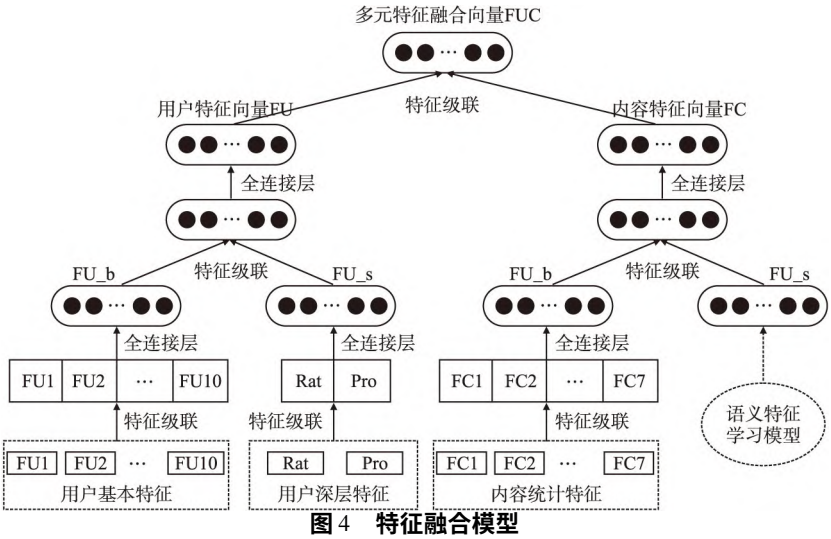

图 4　特征融合模型

Fig. 4　Feature fusion model

首先,针对用户基本特征、用户深层特征(理性值、专业度)、内容统计特征分别采用特征级联 + 全连接的方式得出用户基本特征向量 FU_b、用户深层特征向量 FU_s、内容统计特征向量 FC_b;然后,继续采用特征级联 + 全连接的方式,把 FU_b 和 FU_s 进行融合得出用户特征向量 FU,把 FC_b 和文本语义特征向量 FC_s 进行融合得出内容特征向量 FC;最后,把用户特征向量 FU 和内容特征向量 FC 进行级联,得出最终多元融合特征向量 FUC. 整个计算过程如公式(13) ~公式(18)所示:

$$FU\_b = f(W(FU1 \oplus \cdots \oplus FU10) + b) \tag{13}$$

$$FU\_s = f(W(Rat \oplus Pro) + b) \tag{14}$$

$$FC\_b = f(W(FC1 \oplus \cdots \oplus FC7) + b) \tag{15}$$

$$FU = f(W(FU\_b \oplus FU\_s) + b) \tag{16}$$

$$FC = f(W(FC\_b \oplus FC\_s) + b) \tag{17}$$

$$FUC = FU \oplus FC \tag{18}$$

其中,$\oplus$表示级联操作,即向量的拼接;W 和 b 分别表示全连接层参数矩阵和偏置项;f 表示全连接层的激活函数.

3.4 结果预测

连接文本语义特征学习模型和特征融合模型,把融合特征 FUC 输入全连接层,最后通过 $softmax$ 分类器进行谣言检测,分类预测结果计算如公式(19)所示:

$$p = softmax(W \cdot FUC + b) \tag{19}$$

模型训练过程中基于最小化交叉熵损失函数对整个模型的参数进行优化,损失函数计算如公式(20)所示:

$$L = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log p_i + (1 - y_i)\log(1 - p_i)) + \frac{\lambda}{2}\|W\|_2^2 \tag{20}$$

其中,y_i 表示样本 i 的真实标签值,p_i 表示模型的预测值,$\frac{\lambda}{2}\|W\|_2^2$ 为 L_2 正则化.

4 实 验

4.1 实验数据

4.1.1 数据样本

本实验使用的数据由 3 部分组成:1)标注了类别的谣言和非谣言微博文本数据;2)谣言和非谣言微博的用户基本信息;3)用户发布该微博之前的历史微博文本数据和对应的评论数据. 目前微博谣言公开数据集有 Ma 等人公开的数据集[9]和清华大学自然语言处理与社会人文计算实验室公布的中文谣言数据集[24]等,但这些数据集都不包含用户的历史微博数据,不符合本实验的要求,所以本实验数据是通过编写网络爬虫对微博社区管理中心和微博站点进行收集所得. 微博站点选择 weibo.cn,该站点相比 weibo.com 站点的页面结构更加简单,并且页面没有使用 ajax 异步加载数据和相关加密技术. 为了避免站点对单个 IP 和账号的访问次数限制,爬虫在 Scrapy 框架的基础上利用了 IP 代理和账号池技术. 微博社区管理中心公示了自 2012 年至今的近 4 万条不实信息,不实信息的公示结果如图 5 所示.

图 5 微博社区管理中心不实信息公示

Fig. 5 False information publicity of weibo community management center

1)谣言样本

随着时间的推移,微博用户的属性可能会发生较大的变化,所以本实验只爬取微博社区管理中心近两年被证实的不实信息来构成谣言样本. 爬虫采用广度优先的策略:

第 1 步. 爬取文本内容不少于 30 个字符的谣言微博信息. 如果文本内容过短,文本缺乏语义信息,所以本文根据一个完整句子的大概长度 30 设置过滤条件.

第 2 步. 爬取谣言散布用户的基本信息,如果用户已注销

则删除对应的谣言样本.

第 3 步. 爬取用户散布该谣言之前最近的 200 条历史微博和对应的评论数据.

在爬取过程中,如果多条谣言属于同一用户,则只保留最新的一条. 最后,获得 3756 条谣言数据.

2) 非谣言样本

根据统计分析发现 88.9% 的微博谣言会在一周内被举报[24],所以本实验中非谣言样本是通过爬取发布时间超过一周且未被举报为不实信息的热门微博(评论数 + 转发数 + 点赞数大于 100),这类微博包含的广告类、个人动态类和转发类的信息较少,微博内容更符合谣言检测任务. 非谣言样本爬取数量、内容过滤条件、历史微博选择方式和谣言样本保持一致. 具体实验样本统计信息如表 3 所示.

表 3　实验样本统计信息

Table 3　Statistics of the experimental dataset

统　计	数量
用户数	7512
谣言样本数	3756
非谣言样本数	3756
谣言用户历史微博样本数	548376
非谣言用户历史微博样本数	661056
历史微博评论总数	13454931

4.1.2　数据预处理

对数据的预处理主要包括特征统计、归一化、过滤、分词、去停用词、文本向量表示:

1) 特征统计. 用户基本特征 *FC*3 至 *FU*10 可以直接根据原始数据进行表示,而 *FU*1(用户头像是否是真实照片)和 *FU*2(用户名是否话题型)两个特征需要人工判断; 内容统计特征 *FC*1 至 *FC*7 可以根据字符统计、正则表达式进行统计判断得到.

2) 归一化处理. 为了提高模型的训练速度,对所有统计数值特征转化到 [0,1] 区间表示.

3) 文本内容过滤. 微博内容包含大量的符号、表情、URL 等,在进行分词前需要对这些特殊字符进行过滤,本文通过正则表达式进行筛选.

4) 分词. 利用 HanLP 分词工具对所有的微博(谣言微博、非谣言微博和用户的历史微博)进行分词.

5) 去停用词. 利用哈工大停用词表,对分词结果中的停用词进行删除.

6) 文本向量表示. 利用由北京师范大学和人民大学研究者开源的中文词向量库(Chinese-Word-Vectors)[25]进行词向量表示,再由词向量构成微博文本向量表示. 该词向量库的向量维度为 300,有针对微博特定领域训练的词向量,相比其他全领域的词向量库更适合微博谣言识别任务.

4.2　评价指标

文本使用准确率(Accuracy)、谣言查准率($Precision_r$)、非谣言查准率($Precision_n$)、谣言查全率($Recall_r$)、非谣言查全率($Recall_n$)、谣言 F1-Measure($F1_r$)、非谣言 F1-Measure($F1_n$)作为模型检测评价指标,其计算分别如公式(21) ~ 公式(27)所示:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

$$Precision_r = \frac{TP}{TP + FP} \tag{22}$$

$$Precision_n = \frac{TN}{TN + FN} \tag{23}$$

$$Recall_r = \frac{TP}{TP + FN} \tag{24}$$

$$Recall_n = \frac{TN}{TN + FP} \tag{25}$$

$$F1_r = \frac{2 \times Precision_r \times Recall_r}{Precision_r + Recall_r} \tag{26}$$

$$F1_n = \frac{2 \times Precision_n \times Recall_n}{Precision_n + Recall_n} \tag{27}$$

其中,TP 表示实际为谣言,预测也为谣言的样本数; TN 表示实际为非谣言,预测也非谣言的样本数; FP 表示实际为非谣言,预测为谣言的样本数; FN 表示实际为谣言,预测为非谣言的样本数.

4.3　实验步骤和参数设置

4.3.1　实验步骤

本文模型的实验步骤如表 4 所示.

表 4　本文模型的实验步骤

Table 4　Experimental steps of this model

步骤	内　　容
Step 1.	计算每个用户的理性值 *Rat*:
	i.　基于 BosonNLP 情感词典(该情感词典是基于微博、新闻、论坛等语料构建的,对微博这种非规范文本有较高的覆盖率)和知网发布的程度级别词语表,根据公式(2)计算用户历史微博的情感度 *Senti*;
	ii.　基于情感分析工具 SnowNLP 计算用户历史微博评论的情感极性值,把情感极性值大于 0.6 的归类为正面情感评论,情感极性值小于 0.3 的归类为负面情感评论,根据公式(3)计算用户历史微博的争议度 *Argue*;
	iii.　根据公式(1)计算用户的理性值 *Rat*.
Step 2.	计算用户专业度 Pro:
	iv.　基于 *LDA* 主题模型计算微博的主题概率分布向量;
	v.　根据公式(4)计算用户对待检测微博的专业度 Pro.
Step 3.	构建文本语义特征学习模型:
	vi.　根据 3.2.1 介绍的内容,基于双向 GRU 神经网络 + 注意力机制构建文本语义特征学习模型.
Step 4.	构建多元特征融合模型:
	vii.　根据 3.3 介绍的内容,构建用户基本特征、用户深层特征、内容统计特征和文本语义特征多元融合的模型.

续表

步骤	内 容
Step 5.	**构建预测结果输出模型:**
	viii. **根据** 3.4 **介绍的内容,构建一个由全连接层** + softmax **分类器组成的结果输出模型**.
Step 6.	**模型训练:**
	ix. **从预处理后的样本数据集中,随机选择** 80% **的样本作为训练集进行模型训练**.
Step 7.	**模型测试:**
	x. **利用剩下的** 20% **样本作为测试集,对训练后的模型进行测试**.

4.3.2 参数设置

在实验过程中,需要对相关模型参数和训练参数进行设置,本文根据先验知识和实验探索的方式对相关参数进行设置: 根据微博官方首页的分类,设置 LDA 主题模型的主题数;

表 5 主要参数设置

Table 5 Main parameter setting

参数类别	参数名称	参数值
LDA 模型参数	主题数 K	50
	文档-主题分布参数 α	0.02
	主题-词分布参数 η	0.02
语义特征学习模型参数	GRU 输入序列单元数 t	100
	GRU 单元 hidden_size	256
	输出 y_t 的全连接层尺寸	256
	实现 Attention 的全连接层尺寸	512
多元特征融合模型参数	计算 FU_b 的全连接层尺寸	64
	计算 FU_s 的全连接层尺寸	16
	计算 FU_b 的全连接层尺寸	64
	计算 FU 的全连接层尺寸	128
	计算 FC 的全连接层尺寸	256
输出模型参数	全连接层尺寸	512
训练参数	学习率 *learning_rate*	0.01
	迭代次数 *epoch_num*	1000
	批量训练的 *bath_size*	64

根据数据预处理后的微博最大词数量设置文本语义特征学习模型中 GRU 输入序列的单元数; 根据信息熵原理,特征蕴含的信息量越多,则其特征表示向量维度越大; 通过多次实验探索,选择相对合适的模型训练参数,如学习率、迭代轮次等. 本实验设定的主要参数如表 5 所示.

4.4 实验结果分析

4.4.1 不同特征融合的实验结果对比

1) 实验结果

为了验证融合多元用户特征和内容特征检测模型的有效性,进行以下 5 组不同特征融合的模型的实验对比:

①*FC_c*: 只利用文本语义特性;

②*FC_c* + *FC_b*: 融合本文语义特征和内容统计特征;

③*FC_c* + *FC_b* + *FU_b*: 融合本文语义特征、内容统计特征和用户基本特征;

④*FC_c* + *FC_b* + *FU_b* + *Rat*: 融合本文语义特征、内容统计特征、用户基本特征和用户理性值特征;

⑤*FC_c* + *FC_b* + *FU_b* + *Rat* + *Pro*: 融合本文语义特征、内容统计特征、用户基本特征、用户理性值特征和用户专业度特征.

5 组模型除融合特征不同,其他模型结构和参数保持一致,其实验结果对比如表 6 所示.

2) 结果分析

通过实验对比结果发现,随着融合的特征越多,模型的准确率、查准率、查全率和 F1-Measure 也越来越高,融合了 *FC_c* + *FC_b* + *FU_b* + *Rat* + *Pro* 模型的准确率达到 91.74%,比其他 *FC_c*、*FC_c* + *FC_b*、*FC_c* + *FC_b* + *FU_b*、*FC_c* + *FC_b* + *FU_b* + *Rat* 4 类融合模型的准确率分别高出 4.39%, 3.60%、1.26%、0.53%. 实验结果表明:

①本文融合多元用户特征和内容特征的有效,以及提出用户理性值和专业度两个深层次特征的有效;

②当在纯内容特征中加入用户特征时模型的准确率提高最多,说明用户特征对基于内容特征的谣言识别具有很好的补充作用;

表 6 不同特征融合的实验结果对比

Table 6 Comparison of experimental results of different feature fusion

融合特征	准确率	类别	查准率	查全率	F1-Measure
FC_c	87.35%	谣言	86.10%	89.08%	87.57%
		非谣言	88.69%	85.62%	87.13%
FC_c + *FC_b*	88.15%	谣言	86.87%	89.88%	88.35%
		非谣言	89.52%	86.42%	87.94%
FC_c + *FC_b* + *FU_b*	90.48%	谣言	90.64%	90.28%	90.46%
		非谣言	90.32%	90.68%	90.50%
FC_c + *FC_b* + *FU_b* + *Rat*	91.21%	谣言	91.54%	90.81%	91.18%
		非谣言	90.89%	91.61%	91.25%
FC_c + *FC_b* + *FU_b* + *Rat* + *Pro*	91.74%	谣言	92.19%	91.21%	91.70%
		非谣言	91.30%	92.28%	91.79%

③*FC_c* 和 *FC_c* + *FC_b* 两种只基于内容特征的模型对谣言类别的识别精度高于非谣言类别,而其他 3 类融合了用户特征和内容特征的模型对非谣言类别的识别精度略高于谣言类,说明内容特征更有利于对谣言的识别,而用户特征有利

于对非谣言的识别.

4.4.2　不同语义特征学习模型的实验结果对比

1) 实验结果

为了验证双向 GRU 神经网络 + 注意力机制的语义特征学习模型的有效性,进行以下 4 组不同模型的实验对比:

①GRU: 基于单向 GRU 神经网络的语义特征学习模型;

②GRU + Attention: 基于单向 GRU 神经网络 + 注意力机制的语义特征学习模型;

③BiGRU: 基于双向 GRU 神经网络的语义特征学习模型;

④BiGRU + Attention: 基于双向 GRU 神经网络 + 注意力机制的语义特征学习模型.

4 组模型除语义特征学习模型不同,其他模型结构和参数保持一致,其实验结果对比如表 7 所示.

表 7　不同语义特征学习模型的实验结果对比

Table 7　Comparison of experimental results of different semantic feature learning models

学习模型	准确率	类别	查准率	查全率	F1-Measure
GRU	90.28%	谣言	91.27%	89.08%	90.16%
		非谣言	89.34%	91.48%	90.39%
GRU + Attention	91.01%	谣言	91.85%	90.01%	90.92%
		非谣言	90.21%	92.01%	91.10%
BiGRU	90.81%	谣言	91.70%	89.75%	90.71%
		非谣言	89.96%	91.88%	90.91%
BiGRU + Attention	91.74%	谣言	92.19%	91.21%	91.70%
		非谣言	91.30%	92.28%	91.79%

2) 结果分析

通过实验对比结果发现:

①BiGRU 模型的准确率高于 GRU 模型,说明双向 GRU 神经网络的语义特征学习能力强于单向 GRU 神经网络;

②加入了注意力机制的 GRU + Attention 和 BiGRU + Attention 两个模型的准确率分别高于未加入注意力机制的 GRU 模型和 BiGRU 模型,说明注意力机制提升了语义特性学习能力,验证了本文 BiGRU + Attention 模型的有效性;

③双向 GRU 使序列某点的输出不仅依赖与之前的文本信息,还依赖与之后的文本信息,注意力机制让模型更加关注于文本中具有谣言模式的部分. 所以通过两者结合,挖掘的语义特征对谣言分类更加有效.

4.4.3　不同实时检测模型的实验对比

1) 实验结果

4.4.1 节和 4.4.2 节的实验对比都是模型本身的纵向对比. 所以,为了进一步验证文本模型,利用本文数据集对文献[15]的谣言实时检测模型(LSTM-CNN)、文献[16]的谣言实时检测模型(TG-BiA)、文献[18]的谣言实时检测模型(T-SVM)和文献[19]的谣言实时检测模型(C-LSTM)进行复现,和文本模型分别对比它们在训练集和测试集中谣言的查准率、查全率、F1-Measure 和方差(训练集 F1-Measure 减去测试集 F1-Measure),对比结果如表 8 所示.

表 8　不同实时检测模型的实验结果对比

Table 8　Comparison of experimental results of different real-time detection models

模型	数据集	查准率	查全率	F1-Measure	方差
LSTM-CNN	训练集	90.15%	91.00%	90.57%	3.47%
	测试集	86.90%	87.30%	87.10%	
TG-BiA	训练集	91.19%	92.27%	91.73%	3.72%
	测试集	87.76%	88.27%	88.01%	
T-SVM	训练集	85.41%	85.70%	85.55%	0.54%
	测试集	84.82%	85.20%	85.01%	
C-LSTM	训练集	91.58%	90.67%	91.12%	1.61%
	测试集	89.93%	89.10%	89.51%	
文本模型	训练集	93.32%	92.45%	92.88%	1.18%
	测试集	92.19%	91.21%	91.70%	

2) 结果分析

通过实验对比结果发现:

①LSTM-CNN 和 TG-BiA 两种只关注谣言文本语义信息的检测模型在训练数据集上能实现较高的 F1-Measure 值,分别达到了 90.57% 和 91.73%,而在测试数据集上分别只有 87.10% 和 88.01%,分别下降了 3.47% 和 3.72%,这可能是因为深度学习算法具有很强的特征学习能力,易学到和训练数据集高度相关的特征,导致模型的泛化能力不足;

②T-SVM 在基于文本内容统计特征的基础上,融合了用户统计特征和历史谣言的相似度特征,模型在训练集和测试集上的 F1-Measure 值分别为 85.55% 和 85.01%,准确率最低,方差也最小,这可能是因为基于人工特征工程 + 统计机器学习的算法模型学习能力低于数据驱动的深度学习算法模型,但统计特征具有全局性,所以模型的方差结果最小;

③C-LSTM 在文本语义特征的基础上融合了用户属性特征,是只关注于文本语义信息模型的改进,测试集上的 F1-Measure 值达到了 89.51%,相比 LSTM-CNN、TG-BiA、T-SVM 3 种模型分别提高了 2.41%、1.50%、4.5%,方差也相

对较小;

④本文提出的模型在文本语义特征的基础上融合了用户基本特征、用户深层特征和内容统计特征,进一步拓展了谣言实时检测的特征空间,在测试集上 F1-Measure 值达到了 91.70%,相比于改进型 C-LSTM 模型高出了 2.19%,在融合了更多的全局性特征后方差也相对更小.

4.4.4 谣言检查时效性分析

本文预测模型依赖于用户的基本信息和历史发文信息,对刚发布的微博进行预测时,如果数据库中已经存储了该发文用户的基本信息、理性值和专业度,那么待检测微博经过预处理后可以直接放入模型进行预测,以 8 核 CPU、32G 内存的计算机为例,整个计算过程毫秒级时间内就可完成. 如果数据库中还没有存储该用户的信息,则先需要进行数据收集和计算,将结果保存数据库. 以 200 条历史微博,单条微博 5 页评论计算,总需要抓取 1001 个页面,以单账号、单 IP、单机器的 Scrapy 为例,数据抓取过程大概在 3 分钟左右,数据预处理、理性值和专业度的计算时间相对可以忽略不计,整个预测过程在 3 分钟左右就能完成,保证了谣言检测的实时性.

5 总 结

针对谣言实时检测问题,本文提出融合多元用户特征和内容特征的检测模型:在传统用户基本特征的基础上,基于用户的历史行为数据,挖掘用户理性值和用户专业度两个深层次特征;在利用双向 GRU 神经网络 + 注意力机制学习文本语义特征的基础上,融合符号、表情等内容统计特征. 基于知识驱动和数据驱动相融合的思路,拓展了谣言实时检测的特征空间,弥补了单一文本数据学习的不足,并通过多种类型的实验结果对比,验证了本文模型的有效性.

虽然,本文模型提高了谣言实时检测的精准度,但也存在以下不足和可以改进的地方:1) 相关统计特征依赖于手工操作,降低了模型的灵活性;2) 无法对新用户进行理性值和专业度的计算,因为新用户没有历史行为数据;3) 用户特征需要更新,因为用户的历史数据和相关属性会随时间发生变化;4) 模型对超短文本类的微博检测效果不好,这类微博很难获取到语义特征和内容统计特征信息;5) 没有对微博中附加的图片和视频进行多模态特征联合分析,而目前一部分的谣言是通过图片和视频进行传播的. 这些不足都是下一步需要继续研究的方向.

References:

[1] Gao Yu-jun, Liang Gang, Jiang Fang-ting, et al. Social network rumor detection: a survey [J]. Acta Electronica Sinica, 2020, 48(7): 1421-1435.

[2] Zhang N, Huang H, Su B, et al. Dynamic 8-state ICSAR rumor propagation model considering official rumor refutation [J]. Physica a Statistical Mechanics & Its Applications, 2014, 415(C): 333-346.

[3] Chen Yan-fang, Li Zhi-yu, Liang Xun, et al. Review on rumor detection of online social networks [J]. Chinese Journal of Computers, 2018, 41(7): 1648-1677.

[4] Castillo C, Mendoza M, Poblete B. Information credibility on twitter [C] //20th International Conference on World Wide Web, ACM, 2011: 675-684.

[5] Yang F, Liu Y, Yu X, et al. Automatic detection of rumor on Sina Weibo [C] //ACM SIGKDD Workshop on Mining Data Semantics, ACM, 2012: 1-7.

[6] He Gang, Lv Xue-qiang, Li Zhuo, et al. Automatic rumor identification on microblog [J]. Library and Information Service, 2013, 57(23): 114-120.

[7] Zeng Zi-ming, Wang Jing. Research on microblog rumor identification based on LDA and random forest [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(1): 89-96.

[8] Wu K, Yang S, Zhu K Q. False rumors detection on Sina Weibo by propagation structures [C] //IEEE 31th International Conference on Data Engineering, IEEE, 2015: 651-662.

[9] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks [C] //25th International Joint Conference on Artificial Intelligence, AAAI Press, 2016: 3818-3824.

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

[11] Xie Bo-lin, Jiang Sheng-yi, Zhou Yong-mei, et al. Misinformation detection based on gatekeepers' behaviors in microblog [J]. Chinese Journal of Computers, 2016, 39(4): 730-744.

[12] Liu Zhi-yuan, Song Chang-he, Yang Cheng. Early detection of rumors in social media [J]. Global Media Journal, 2018, 5(4): 65-80.

[13] 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.

[14] 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): 102-106 + 115.

[15] Ajao O, Bhowmik D, Zargari S. Fake news identification on twitter with hybrid CNN and RNN Models [C] //9th International Conference on Social Media and Society, ACM, 2018: 226-230.

[16] Li Ao, Dan Zhi-ping, Dong Fang-min, et al. An improved generative adversarial network for rumor detection [J]. Journal of Chinese Information Processing, 2020, 34(9): 78-88.

[17] Pan De-yu, Song Yu-rong, Song Bo. New microblog rumor detection model based on attention mechanism [J]. Journal of Chinese Computer Systems, 2021, 42(2): 348-353.

[18] Ma Ming, Liu Yun, Liu Di-jun, et al. Rumor detection in microblogs based on topic and prevention model [J]. Transactions of Beijing Institute of Technology, 2020, 40(3): 310-315.

[19] Yin Peng-bo, Pan Wei-min, Peng Cheng, et al. Research on early detection of weibo rumors based on user characteristics analysis [J]. Journal of Intelligence, 2020, 39(7): 81-86.

[20] Liu Ya-hui, Jin Xiao-long, Shen Hua-wei, et al. A survey on rumor identification over social media [J]. Chinese Journal of Computers, 2018, 41(7): 1536-1558.

[21] Morris M R, Counts S, Roseway A, et al. Tweeting is believing?: understanding microblog credibility perceptions [C] //ACM 2012 Conference on Computer Supported Cooperative Work, Seattle,

USA,2012:441-450.
[22] Guo H,Cao J,Zhang Y,et al. Rumor detection with hierarchical social attention network [C] //27th ACM International Conference on Information and Knowledge Management, ACM, 2018: 943-951.
[23] Liu Li-ping,Liu Meng,Li Shao-ping. Sina microblog disputed-level analysis based on sentiment analysis [J]. Computer Engineering & Science,2016,38(10):2158-2164.
[24] Hu X,Tang J,Gao H,et al. Social spammer detection with sentiment information [C] //IEEE International Conference on Data Mining,Shenzhen,China,2014:180-189.
[25] Blei D M,Ng A Y,Jordan M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research,2003,3(3):993-1022.
[26] Liu Zhi-yuan,Zhang Le,Tu Cun-chao,et al. Statistical and semantic analysis of rumors in Chinese social media [J]. Scientia Sinica (Informationis),2015,45(12):1536-1546.
[27] Li S,Zhao Z,Hu R,et al. Analogical reasoning on chinese morphological and semantic relations [C] //56th Annual Meeting of the Association for Computational Linguistics,ACL,2018:138-143.

附中文参考文献:

[1] 高玉君,梁　刚,蒋方婷,等. 社会网络谣言检测综述 [J]. 电子学报,2020,48(7):1421-1435.
[3] 陈燕方,李志宇,梁　循,等. 在线社会网络谣言检测综述 [J]. 计算机学报,2018,41(7):1648-1677.
[6] 贺　刚,吕学强,李　卓,等. 微博谣言识别研究 [J]. 图书情报工作,2013,57(23):114-120.
[7] 曾子明,王　婧. 基于 LDA 和随机森林的微博谣言识别研究——以 2016 年雾霾谣言为例 [J]. 情报学报,2019,38(1):89-96.
[11] 谢柏林,蒋盛益,周咏梅,等. 基于把关人行为的微博虚假信息及早检测方法 [J]. 计算机学报,2016,39(4):730-744.
[12] 刘知远,宋长河,杨　成. 社交媒体平台谣言的早期自动检测 [J]. 全球传媒学刊,2018,5(4):65-80.
[13] 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.
[14] 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):102-106+115.
[16] 李　奥,但志平,董方敏,等. 基于改进生成对抗网络的谣言检测方法 [J]. 中文信息学报,2020,34(9):78-88.
[17] 潘德宇,宋玉蓉,宋　波. 一种新的考虑注意力机制的微博谣言检测模型 [J]. 小型微型计算机系统,2021,42(2):348-353.
[18] 马　鸣,刘　云,刘地军,等. 基于主题和预防模型的微博谣言检测 [J]. 北京理工大学学报,2020,40(3):310-315.
[19] 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.
[20] 刘雅辉,靳小龙,沈华伟,等. 社交媒体中的谣言识别研究综述 [J]. 计算机学报,2018,41(7):1536-1558.
[23] 刘莉平,刘　梦,李绍鹏. 基于情感分析的新浪微博争议度分析 [J]. 计算机工程与科学,2016,38(10):2158-2164.
[26] 刘知远,张　乐,涂存超,等. 中文社交媒体谣言统计语义分析 [J]. 中国科学: 信息科学,2015,45(12):1536-1546.

本刊检索与收录

国内

中文核心期刊
中国学术期刊文摘(中英文版)收录
中国科学引文数据库(CSCD)来源期刊
中国科技论文统计源期刊
中国期刊全文数据库(CJFD)收录期刊
中国科技期刊精品数据库收录期刊
中国学术期刊综合评价数据库(CAJCED)收录期刊
中国核心期刊(遴选)数据库收录期刊
中文科技期刊数据库收录期刊

国际

英国《科学文摘》(INSPEC)
荷兰《文摘与引文数据库》(SCOPUS)
俄罗斯《文摘杂志》(AJ,VINITI)
美国《剑桥科学文摘(自然科学)》CSA(NS); Cambridge Scientific Abstracts(Natural Science)
美国《剑桥科学文摘》CSA(T); Cambridge Scientific Abstracts(Technology)
美国《乌利希期刊指南》UPD(Ulrich's Periodicals Directory)
日本《日本科学技术振兴机构中国文献数据库》(JST,China)
波兰《哥白尼索引》(IC, Index of Copurnicus)