中国科学: 信息科学 2018 年 第 48 卷 第 11 期: 1558–1574
SCIENTIA SINICA Informationis
社会媒体计算与自然语言处理专刊 • 论文

基于分层注意力网络的社交媒体谣言检测

廖祥文[1,2,3*], 黄知[1,2,3], 杨定达[1,2,3], 程学旗[4], 陈国龙[1,2,3]

1. 福州大学数学与计算机科学学院, 福州 350116
2. 福建省网络计算与智能信息处理重点实验室 (福州大学), 福州 350116
3. 数字福建金融大数据研究所, 福州 350116
4. 中国科学院网络数据科学与技术重点实验室, 北京 100086
* 通信作者. E-mail: liaoxw@fzu.edu.cn

收稿日期: 2018–05–25; 接受日期: 2018–09–14; 网络出版日期: 2018–11–14

国家自然科学基金 (批准号: 61772135, U1605251)、中国科学院网络数据科学与技术重点实验室开放基金课题 (批准号: CASNDST201708, CASNDST201606) 和北邮可信分布式计算与服务教育部重点实验室主任基金 (批准号: 2017KF01) 资助项目

摘要 在社交媒体谣言检测问题上, 现有的基于特征表示学习的研究工作大多数先把微博事件划分为若干个时间段, 再对每个时间段提取文本向量表示、全局用户特征等, 忽略了时间段内各微博间的时序信息, 且未利用到在传统机器学习方法中已取得较好效果的文本潜在信息和局部用户信息, 导致性能较低. 因此, 本文提出了一种基于分层注意力网络的社交媒体谣言检测方法. 该方法首先将微博事件按照时间段进行分割, 并输入带有注意力机制的双向 GRU 网络, 获取时间段内微博序列的隐层表示, 以刻画时间段内微博间的时序信息; 然后将每个时间段内的微博视为一个整体, 提取文本潜在特征和局部用户特征, 并与微博序列的隐层表示相连接, 以融入文本潜在信息和局部用户信息; 最后通过带有注意力机制的双向 GRU 网络, 得到时间段序列的隐层表示, 进而对微博事件进行分类. 实验采用了新浪微博数据集和 Twitter 数据集, 实验结果表明, 与目前最好的基准方法相比, 该方法在新浪微博数据集和 Twitter 数据集上正确率分别提高了 1.5% 和 1.4%, 很好地验证了该方法在社交媒体谣言检测问题上的有效性.

关键词 谣言检测, 分层注意力网络, 社交媒体, 时序信息, 深度学习

1 引言

谣言 (rumor) 是在人与人之间传播的, 含有公众关心的信息且真实性不能很快得到证明或是得不到证明的一种特殊陈述 [1]. 谣言往往是一些不正确的或是误导人的信息 [2], 又被称为误信息 (misinformation)、假消息 (fake news) 等 [3]. 近年来随着社交媒体用户使用率的持续增长, 谣言得以在社交媒体中快速而广泛的传播 [1,4], 社交媒体已成为谣言传播的重要媒介 [5~8]. 谣言检测 (rumor detection)

引用格式: 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. 中国科学: 信息科学, 2018, 48: 1558–1574, doi: 10.1360/N112018-00134
Liao X W, Huang Z, Yang D D, et al. Rumor detection in social media based on a hierarchical attention network (in Chinese). Sci Sin Inform, 2018, 48: 1558–1574, doi: 10.1360/N112018-00134

© 2018《中国科学》杂志社

旨在研究如何从事件相关的言论中检测事件的真实性 [9], 对于改善网络信息生态环境质量 [6]、维护社会稳定 [8] 具有重要意义.

大多数研究者将社交媒体谣言检测问题看作一个二分类问题, 现有的工作主要有以下两类方法. (1) 基于传统机器学习的方法: 首先通过运用特征工程手段, 提取事件相关信息中的情感极性 [10~12]、质疑更正信号 [13,14] 、用户影响力 [10]、地理位置信息 [12]、传播结构的相似性 [15] 等特征, 之后训练决策树 [11,13,14,16]、支持向量机 [10,12,15,17] 等分类器将事件分类为谣言和非谣言. 但是, 该类方法依赖于人工特征, 而人工特征定义需要耗费大量的人力、物力和时间, 所得到的特征向量的鲁棒性也不够强. (2) 基于特征表示学习的方法: 首先将微博事件分割成若干个时间段, 并将每个时间段内的微博视为一个整体, 提取由 tf-idf [7], doc2vec [8,18] 等计算得到的文本表示以及对 "事件 – 用户" 关系矩阵和 "用户 – 用户" 关系矩阵进行奇异值分解得到的全局用户特征 [18], 之后采用双层 GRU 网络模型 (GRU-2) [7]、卷积神经网络模型 (CAMI) [8]、混合深度模型 (CSI) [18] 等, 获得微博事件的隐层表示, 并利用该隐层表示对微博事件进行分类. 但是, 直接将时间段内的微博视为一个整体提取特征的方式, 导致了模型无法刻画时间段内各微博间的时序信息; 并且, 在用户量大的情况下, 全局用户特征的计算开销十分巨大. 此外, 质疑更正信号 [13] 等文本潜在特征和局部用户特征已在以往的研究工作中取得了较为显著的效果, 而该类方法没有利用到这些潜在特征, 导致性能较低.

针对上述问题, 本文提出一种基于分层注意力网络的社交媒体谣言检测方法. 一方面, 该方法通过两层带有注意力机制的双向 GRU 网络 (BiGRU-Attention 网络), 从微博层面和时间段层面分层学习时间段内微博序列的隐层表示和时间段序列的隐层表示; 另一方面, 该方法还对时间段提取了质疑更正信号、用户认证情况等多种文本潜在特征和局部用户特征, 进而将这些特征与时间段内微博序列的隐层表示进行特征融合. 该方法能够捕获到时间段内微博间的时序信息, 并且将文本潜在信息和局部用户信息融入到时间段表示中, 有利于提升模型的谣言检测性能. 本文采用 Ma 等 [7] 在 2016 年公开的新浪微博数据集和 Twitter 数据集进行实验. 结果表明, 与目前最好的基准方法相比, 该方法在新浪微博数据集上正确率提高了 1.5%, 在 Twitter 数据集上正确率提高了 1.4%, 很好地验证了该方法的有效性. 此外, 在早期谣言检测任务上, 该方法也表现出了优于基准方法的性能.

本文第 2 节介绍社交媒体谣言检测的相关研究工作; 第 3 节为模型方法, 介绍本文提出的基于分层注意力网络的社交媒体谣言检测方法; 第 4 节为实验及分析, 通过与基准方法进行对比实验, 验证本文方法在社交媒体谣言检测中的有效性; 第 5 节为结束语.

2 相关工作

目前国内外针对社交媒体谣言检测的研究工作主要可以分为两个大类, 一类是基于传统机器学习的方法, 另一类是基于特征表示学习的方法. 基于传统机器学习的方法大多集中在特征设计上, 该类方法采用特征工程手段, 从事件相关信息中提取人工特征, 并利用决策树、SVM 等分类器对事件进行分类. Castillo 等 [11] 在 2011 年提取了情感分数、包含 URL 的微博数、用户注册天数等特征, 并利用决策树算法来检测谣言; Yang 等 [12] 在 2012 年采用了微博中涉及的地理位置、发布微博的客户端, 以及文本符号的情感极性等特征, 进而利用 SVM 分类器检测谣言; 这两项研究工作分别在 Twitter 数据集和新浪微博数据集上提取了多种浅层特征, 但浅层特征依赖于人工设计, 难以得到较好的鲁棒性, 后续的研究工作在这些特征基础上做了进一步的拓展, 在内容特征、用户特征、传播特征这 3 个方面挖掘更多新的潜在特征.

(1) 在内容特征和用户特征方面, Qazvinian 等 [19] 在 2011 年采用 Bayes 分类器构建了积极模型

和消极模型, 提取了用户发布或转发信息的行为特征、语言模型提取的文本特征等多种潜在特征; Sun 等 [16] 在 2013 年针对谣言所存在的图文不符的特点, 提取了图片最早出现的时间和微博发布时间之间的时间差等特征; Zhang 等 [10] 在 2015 年采用了评论中所表达的观点、情感极性、用户影响力等多种潜在特征; Liang 等 [14] 在 2015 年提取了带有质疑文本内容的微博所占的比例、平均每天发布的微博数、平均每天关注的用户数等内容特征和用户行为特征; Zhao 等 [13] 在 2015 年统计了微博中带有质疑和更正信号的微博比例、带有 hashtag 的微博比例、带有 URL 的微博比例等特征; Hamidian 等 [20] 在 2016 年采用了用户对推文的信任度等特征.

(2) 在传播特征方面, Ma 等 [17] 在 2015 年针对以往工作忽略了与时间相关的传播特征的缺陷, 提出了采用动态时间序列模型检测谣言, 获取了部分现有的特征在信息传播的生命周期内随时间动态变化的特征; 之后 Ma 等 [15] 又在 2017 年提出通过评估不同类型的传播树结构的相似性来捕获微博事件的高阶特征; Liu 等 [21] 在 2017 年对谣言和非谣言之间的差异进行了系统分析, 提取了信息传播过程中传播树的深度和宽度等特征.

虽然随着特征设计的深入, 基于传统机器学习的方法在谣言检测性能上也得到了逐步提升, 但人工特征所存在的需要耗费人力、物力和时间来设计特征、鲁棒性不足等问题仍无法较好的解决.

近年来, 随着深度学习方法在计算机视觉 [22]、语音识别 [23] 等领域取得了显著效果, CNN [24], RNN [25], 以及 LSTM [26] 等也被用于对文本序列等进行特征表示学习, 解决句子分类 [27]、情感分类 [28]、文档分类 [29], 以及机器翻译 [30] 等问题. 因此, 这种基于特征表示学习的方法也开始被逐渐应用到谣言检测问题上. 该类方法先将事件按照时间段分割, 再将时间段内微博视为一个整体提取特征向量, 输入到神经网络模型中, 学习事件的隐层表示, 最后对事件进行分类. Ma 等 [7] 在 2016 年首次将深度学习模型应用在社交媒体谣言检测问题上, 利用 tf-idf 计算得到各个时间段的微博文本向量, 并输入双层的 GRU 网络学习事件的特征表示; Yu 等 [8] 在 2017 年利用 doc2vec 方法获取时间段内微博的文本向量, 将各时间段文本向量拼接成事件的特征矩阵, 并采用 CNN 学习事件的隐层表示; Ruchansky 等 [18] 在 2017 年提出了一种混合深度模型, 采用单层 LSTM 神经网络学习事件的隐层表示, 并对 “事件 – 用户” 关系矩阵和 “用户 – 用户” 关系矩阵进行奇异值分解得到用户的全局特征, 最后采用连接后的事件特征向量对事件分类; Ma 等 [9] 在 2018 年将谣言检测作为多任务学习的一个子任务, 利用基于 RNNs 的多任务学习模型, 得到事件的隐层表示, 最后将事件按照真实性分为 4 个类别. 虽然该类方法已取得了比传统机器学习方法更好的效果, 但未考虑文本潜在信息和局部用户信息, 且忽略了时间段内各微博间的时序信息, 对谣言检测效果造成了一定的影响.

3 模型方法

3.1 问题描述

社交媒体谣言检测任务旨在研究如何从社交媒体中检测谣言, 实质上是一个二分类问题, 该任务的形式化定义如下: 给定一个事件集合 $E = \{e_1, e_2, e_3, e_4, \ldots\}$ 和一个类别集合 $L = \{l_1, l_2\}$. 其中, e_i 代表一个微博事件, e_i 中包含有若干条微博 $m_{i,j}$ 及其时间戳 $t_{i,j}$, 即 $e_i = \{(m_{i,j}, t_{i,j})\}$, l_1, l_2 分别代表谣言和非谣言这两个类别标签. 社交媒体谣言检测任务是要学习一个分类模型 f, 将每个事件 e_i 映射成一个类别标签 l_j, 即 $f : e_i \to l_j$. 模型的输入是一个包含有若干条微博的事件, 输出是该事件对应的标签 (谣言/非谣言), 即输出该事件是否为谣言.

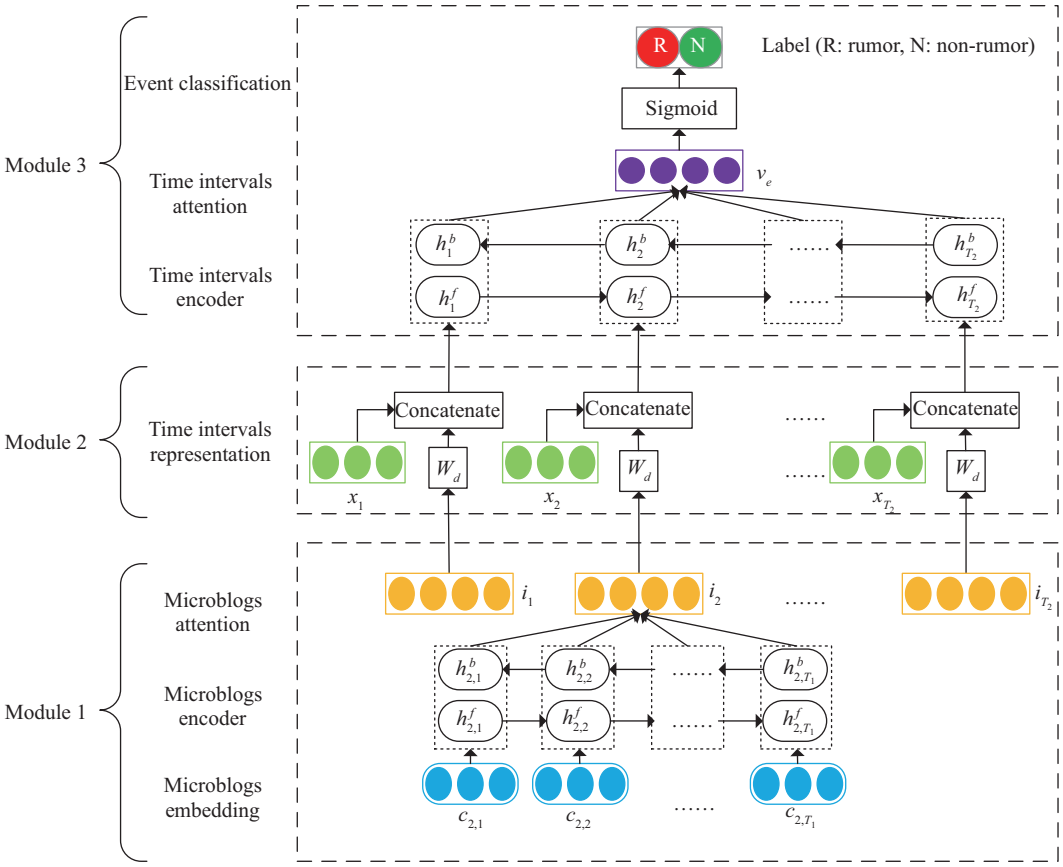

图 1 (网络版彩图) 基于分层注意力网络的社交媒体谣言检测模型

Figure 1 (Color online) The model of rumor detection in social media based on hierarchical attention networks

3.2 基于分层注意力网络的社交媒体谣言检测方法

为了捕获时间段内微博之间的时序信息, 本文采用了分层注意力网络 (hierarchical attention networks) 分别从微博层面和时间段层面学习微博序列和时间段序列的隐层表达. 此外, 本文在将时间段内微博视为一个整体提取特征时, 除了采用 doc2vec 方法 [31] 计算得到的微博文本表示外, 还结合了质疑更正信号等文本潜在特征和用户认证情况等局部用户特征, 以在时间段表示中融入文本潜在信息和局部用户信息. 本文提出的基于分层注意力网络的社交媒体谣言检测模型如图 1 所示.

对于给定的事件相关信息, 模型的输入包括两个部分: (1) 时间段内各微博的文本表示 $c_{t,k}$, 其中 t 表示第 t 个时间段, k 表示时间段内第 k 条微博; (2) 事件各个时间段的特征向量 x_t, 包括质疑更正信号、用户认证情况等特征. 模型的输出为事件的标签 (谣言/非谣言). 模型主要包括如下 3 个模块:

(1) 基于 BiGRU-Attention 网络的微博序列表示. 在该模块中, 首先将时间段内的每条微博的文本利用 doc2vec 方法计算得到微博文本表示, 并将其作为 BiGRU-Attention 网络在各个时刻的输入, 该模块的输出为所学习到的时间段内微博序列的表示;

(2) 基于特征融合的时间段表示. 在该模块中, 采用特征工程手段提取时间段内微博的相关特征构造时间段特征向量, 并与上一模块学习到的时间段内微博序列的表示相连接, 构成时间段的表示;

(3) 基于 BiGRU-Attention 网络的时间段序列表示. 在该模块中, 将各个时间段的表示作为输入,

通过 BiGRU-Attention 网络获得时间段序列的表示, 并作为事件的隐层表达, 进而对事件分类.

3.2.1 基于 BiGRU-Attention 网络的微博序列表示

基于 BiGRU-Attention 网络的微博序列表示是从微博层面学习各个时间段内微博序列的隐层表达. 该模块主要包含一个双向 GRU 层 (BiGRU) 和一个注意力层 (attention). 对于第 t 个时间段, 该模块的输入是时间段内各微博的文本向量 $c_{t,k}$, 其中 k 表示该时间段内的第 k 条微博, 输出是所学习到的时间段内微博序列的表示 i_t.

一方面, 在事件的各个时间段内, 该模块采用双向 GRU 层来捕获时间段内微博间的时序信息. 传统的单向循环神经网络只考虑了过去的微博信息, 而双向的 GRU 网络能够同时结合过去和未来的微博的文本表示生成当前时刻的输出. 另外, GRU 和 LSTM 一样可以解决长期依赖问题, 且 GRU 所需训练的参数比 LSTM 少. 该双向 GRU 层接收时间段内各微博的文本表示 $c_{t,k}$ 作为各个时刻的输入, 具体的计算公式如下:

$$h^f_{t,k} = \overrightarrow{\mathrm{GRU}}(c_{t,k}), \quad k \in [1, T_1], \tag{1}$$

$$h^b_{t,k} = \overleftarrow{\mathrm{GRU}}(c_{t,k}), \quad k \in [T_1, 1], \tag{2}$$

$$h_{t,k} = \left(h^f_{t,k}, h^b_{t,k}\right), \tag{3}$$

其中, $h^f_{t,k}$ 和 $h^b_{t,k}$ 分别为前向和后向的 GRU 网络在时刻 k 的输出, T_1 为时间步长, 即时间段内微博序列的长度, $(h^f_{t,k}, h^b_{t,k})$ 表示这两个向量的连接, $h_{t,k}$ 表示双向 GRU 层在时刻 k 的输出.

另一方面, 在双向 GRU 层之上, 该模块采用注意力层对双向 GRU 层在各个时刻的输出进行加权求和, 以获取微博序列的隐层表示. 对双向 GRU 网络引入注意力机制能给双向 GRU 网络输出的隐状态序列 $(h_{t,1}, h_{t,2}, \ldots, h_{t,T_1})$ 赋予不同的权重, 这样在学习微博序列的表示时模型能够有所侧重地利用时间段内的微博序列信息. 该注意力层将双向 GRU 网络在各个时刻的输出 $h_{t,k}$ 作为输入, 输出时间段内微博序列的表示 i_t, 具体公式如下:

$$u_{t,k} = \tanh(W_c h_{t,k} + b_c), \tag{4}$$

$$a_{t,k} = \frac{\exp(u^{\mathrm{T}}_{t,k} u_c)}{\sum_j \exp(u^{\mathrm{T}}_{t,j} u_c)}, \tag{5}$$

$$i_t = \sum_k a_{t,k} h_{t,k}, \tag{6}$$

其中, $u_{t,k}$ 为 $h_{t,k}$ 经过以 tanh 为激活函数的隐藏层后的所得到的 $h_{t,k}$ 的隐层表示, u_c 为随机初始化的权重, $a_{t,k}$ 为赋予双向 GRU 网络各个时刻隐含状态的权重, 是通过对隐含状态 $h_{t,k}$ 的隐层表示 $u_{t,k}$ 进行 softmax 标准化后得到, i_t 为加权求和后所得到的时间段内微博序列的表示.

3.2.2 基于特征融合的时间段表示

基于特征融合的时间段表示是将所学习到的微博序列表示与采用特征工程手段提取到的时间段特征进行连接, 得到时间段的表示. 该模块包含一个全连接层和一个融合层, 输入包括上一模块学习到的微博序列表示 i_t 以及通过特征工程方法提取的时间段特征向量 x_t, 输出为时间段的表示 d_t.

首先, 将时间段内的微博看作一个整体, 采用特征工程手段提取时间段的特征, 并将所提取到的特征进行连接, 得到时间段特征向量 x_t, 其中包括文本向量表示 (text vector)、带有质疑信号的微博比例

表 1 时间段特征

Table 1 The features of the time intervals

	Feature	Description
Content-based features	Text vector	Calculated by utilizing doc2vec
	Enquiries and corrections	% of microblogs with enquiries and corrections
	Length of microblogs	Average length of microblogs
User-based features	Personal description	% of users that provide personal description
	Verified users	% of verified users
	Users reputation	Followers/followees ratio
	Users activeness	Followees/followers ratio

表 2 质疑更正信号正则表达式

Table 2 The regular expression list of enquiries and corrections

Chinese	English
(?: 这 | 那 | 它) 是真的吗	is\s(?:that|this|it)\s true
什么 [?!][?1]*	wh[a]*t[?!][?1]*
真的?| 真的? | 求证 | 真的假的 | 真的吗 | 未经证实	real?|really?|unconfirmed
谣言 | 揭穿	rumor|debunk
(?: 那 | 这 | 它) 不是真的 | 假的	(?:that|this|it)\s is\s not\s true

(enquiries and corrections)、微博平均长度 (length of microblogs) 等文本潜在信息以及提供个人描述的用户比例 (personal description)、认证的用户比例 (verified users)、用户声望得分 (users reputation) 和用户积极性 (users activeness) 等局部用户信息. 具体特征如表 1 所示.

在表 1 中, 质疑更正信号是用户对微博事件真实性所表达的质疑或更正态度的相关文本内容, 如“真的吗”“真的假的” 等. Zhao 等 [13] 曾在 2015 年采用正则表达式匹配 Twitter 中的质疑更正信号. 在微博事件的各个时间段内, 本文同样采用了正则表达式来匹配文本中是否存在质疑更正信号, 进而采用统计时间段内带有质疑更正信号的微博比例. 采用的质疑更正信号正则表达式如表 2 所示.

然后, 将上一模块学习到的时间段内微博序列的表示 i_t 送入一个全连接层, 该层采用 tanh 函数作为激活函数, 从而得到时间段内微博序列的隐层表示 f_t:

$$f_t = \tanh(W_d i_t + b_d). \tag{7}$$

最后, 将提取到的时间段特征向量 x_t 与时间段内微博序列的隐层表示 f_t 送入融合层进行连接, 得到时间段的表示 d_t, d_t 中包含了时间段内微博序列的时序信息、文本潜在信息及局部用户信息:

$$d_t = (f_t, x_t). \tag{8}$$

3.2.3 基于 BiGRU-Attention 网络的时间段序列表示

基于 BiGRU-Attention 网络的时间段序列表示是从时间段层面学习时间段序列的隐层表达. 该模块也包含有一个双向 GRU 层和一个注意力层. 输入为时间段表示 d_t, 输出为时间段序列表示 v_e.

双向 GRU 层接收各时间段的特征向量 d_t 作为各个时刻的输入, 采用该双向 GRU 层能够使得模

型同时结合过去和未来的时间段信息, 生成当前时刻的隐含状态 h_t. 具体计算公式如下:

$$h_t^f = \overrightarrow{\mathrm{GRU}}(d_t), \quad t \in [1, T_2], \tag{9}$$

$$h_t^b = \overleftarrow{\mathrm{GRU}}(d_t), \quad t \in [T_2, 1], \tag{10}$$

$$h_t = \left(h_t^f, h_t^b\right), \tag{11}$$

其中, h_t^f 为前向的 GRU 网络在时刻 t 的输出值, h_t^b 为后向的 GRU 网络在时刻 t 的输出值, T_2 为时间步长, 即时间段的数量, (h_t^f, h_t^b) 为两个向量的连接. h_t 表示双向 GRU 网络在时刻 t 的输出.

同样, 本文在双向 GRU 层之上增加了一个注意力层, 使得模型在时间段序列的表示时能够有所侧重地利用到各个时间段的信息. 该注意力层接收双向 GRU 网络在各个时刻输出的隐含状态 h_t, 生成时间段序列的表示 v_e. 注意力层的相关计算公式如下:

$$u_t = \tanh(W_d h_t + b_d), \tag{12}$$

$$a_t = \frac{\exp(u_t^{\mathrm{T}} u_d)}{\sum_m \exp(u_m^{\mathrm{T}} u_d)}, \tag{13}$$

$$v_e = \sum_t a_t h_t, \tag{14}$$

其中, u_t 为 h_t 经过以 tanh 为激活函数的隐藏层后的所得到的 h_t 的隐层表示, u_d 为随机初始化的权重, a_t 为赋予双向 GRU 网络各个时刻隐含状态的权重, 是通过对隐含状态 h_t 的隐层表示 u_t 进行 softmax 标准化后得到的, v_e 为加权求和后所得到的时间段序列表示.

所得到时间段序列表示 v_e 可以看作是最终得到的微博事件的隐层表达, 进而将该隐层表达送入一个全连接层, 该层采用 sigmoid 作为激活函数, 输出该微博事件的预测标签 $\hat{L}_e$:

$$\hat{L}_e = \sigma(W_e^{\mathrm{T}} v_e + b_e). \tag{15}$$

3.3 模型求解

为了使模型输出的预测标签更接近于真实标签, 本文采用交叉熵来衡量真实标签和预测标签之间的相似程度. 真实标签和预测标签的交叉熵越小, 说明二者的相似度越高, 模型的谣言检测性能也越好. 因此, 在训练过程中, 最小化如式 (16) 所示的交叉熵损失函数, 其中 N 表示训练集中微博事件的总数, L_j 是第 j 个微博事件的真实标签, $\hat{L}_j$ 是第 j 个微博事件的预测标签, θ 为模型的参数集合.

$$\mathrm{Loss} = -\frac{1}{N}\sum_{j=1}^{N}[L_j \ln \hat{L}_j + (1 - L_j)\ln(1 - \hat{L}_j)] + \frac{\lambda}{2}\|\theta\|_2^2. \tag{16}$$

同时, 为了降低模型过拟合的程度, 本文对全连接层的权重 W_d 和 GRU 的权重施加 L2 正则化, 并对 GRU 运用 dropout 正则化方法. 此外, 为了加快模型的收敛速度, 本文采用 Adam 优化算法 [32] 来更新模型参数.

本文模型的训练算法步骤描述如算法 1 所示. 在每次迭代中, 模型将事件集合作为原始的输入数据, 通过分层注意力网络, 获得事件的隐层表示, 并对事件进行分类, 得到事件的预测标签.

表 3 原始的实验数据集统计表

Table 3 Statistics of the dataset

Statistic	Sina Weibo	Twitter
Events#	4664	992
Rumors#	2313	498
Non-rumors#	2351	494
Microblogs#	3805656	1101985
Users #	2746818	491229
Average time length/event (h)	2460.7	1582.6
Average # of posts/event	816	1111
Max # of posts/event	59318	62827
Min # of posts/event	10	10

算法 1 基于分层注意力网络的社交媒体谣言检测模型的训练算法

输入: 训练数据集事件集合 $E=\{e_1,e_2,\ldots\}$, 其中 $e_i=\{(m_{i,j},t_{i,j})\}_{j=1}^{n_i}$; 事件对应的真实标签集合 $Y^*=\{y_1^*,y_2^*,\ldots\}$.

初始化模型参数集合 θ, 最大迭代次数 MAXEPOCH, 当前迭代次数 epoch $\Leftarrow 1$;
while epoch $\leqslant$ MAXEPOCH **do**
  对各个事件 e_i, 计算对应的预测标签 L_i;
  根据式 (16) 计算损失值 Loss;
  根据 Loss 的值利用 Adam 优化算法更新参数集合 θ;
  epoch $\Leftarrow$ epoch $+1$;
end while

输出: 训练后得到的最优的模型参数集合 θ.

4 实验及分析

4.1 实验数据集

本文采用 Ma 等 [7] 在 2016 年公开的用于社交媒体谣言检测研究的数据集, 该数据集包含新浪微博和 Twitter 两个部分. 目前该数据集已经在文献 [7,8,15,18] 中使用, 是社交媒体谣言检测问题上的经典数据集. 表 3 给出原始的实验数据集统计表.

新浪微博数据集包含 4664 个事件及事件对应的标签, 每个事件包含有若干条微博, 每条微博的具体内容均已在原始数据集中完整提供. Twitter 数据集包含 992 个事件及事件对应的标签, 每个事件包含若干条 Tweet. 但数据集中只提供了每个事件相关的 Tweet ID, 具体的内容需自行利用 Tweet ID 通过 Twitter 官方开放的 API 接口获取. 由于随着时间的推移, 部分 Tweet 已被删除或是设置了访问权限, 这部分 Tweet 的内容无法通过 API 接口获取到, 无法获取到内容的 Tweet 约占总数的 10%. 此外, 由于 Twitter 数据集中 ID 为 "E354" 的事件的 Tweet 内容均无法获取到, 实验中不使用该事件的相关数据.

4.2 实验评价指标

为了评估本文所提出的模型的有效性并与以往的研究工作进行对比, 本文采用了与文献 [7,8,18] 相同的评价指标, 总共包括正确率、准确率、召回率和 $F1$ 值这 4 个评价指标.

(1) 正确率 (accuracy). 正确率为被模型正确分类的事件数量与事件总数的比值, 正确率越高, 模型的分类性能越好.

(2) 准确率 (precision). 准确率为被模型正确分类为谣言 (或非谣言) 的事件数量与被分类为谣言 (或非谣言) 的事件总数的比值.

(3) 召回率 (recall). 召回率为被模型正确分类为谣言 (或非谣言) 的事件数量与实际为谣言 (或非谣言) 的事件总数的比值.

(4) $F1$ 值 ($F1$-score). $F1$ 值为准确率和召回率加权调和平均, 是对准确率和召回率的综合考虑, 其计算公式为 $F1 = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$. $F1$ 值越高, 模型的分类性能越好.

4.3 实验环境与参数设置

本文的实验环境为, 处理器: Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10 GHz, 操作系统: Ubuntu 14.04.5 LTS , 内存: 32 GB RAM, GPU: Tesla K40m, 开发平台: Python 2.7.13.

实验的数据集划分情况为, 事件中 75% 作为训练集, 10% 作为验证集, 15% 作为测试集. 由于本文方法是在基准方法基础上进行, 为了较好地与对比实验进行比较, 在实验参数设置上, 本文沿用了文献 [18] 的参数设置, 具体的实验参数设置如下: 段落向量维度设置为 100, 学习率设置为 0.001, 正则化系数 λ 设置为 0.01, dropout 的概率设置为 0.2, 全连接层和 GRU 层的单元数设置为 100. 此外, 实验中采用了 Ruchansky 等在文献 [18] 中使用的微博事件时间段分割方法, 该方法将微博发布时间的时间粒度从 “秒” 转换为 “时”, 将同一小时内的微博归入同一个时间段.

4.4 实验对比模型

将本文方法与所选取的基准方法在相同的数据集上开展实验, 本文选取了以下几个基准方法:

(1) DTC 模型 [11]. 该模型提取了情感分数、包含 URL 的微博数、用户注册天数、发布的微博数、平均粉丝数、平均关注数等特征, 并采用 J48 决策树进行分类.

(2) SVM-TS 模型 [17]. 该模型利用动态时间序列来捕获现有的微博事件特征随时间的变化情况, 并将所捕获的时间相关特征和现有的微博事件特征相连接, 采用 SVM 分类器进行分类.

(3) DT-Rank 模型 [13]. 该模型先利用正则表达式匹配与质疑更正信号相关的文本, 筛选出带有质疑更正信号的微博, 然后利用微博信息之间的相似度对这些微博进行聚类, 并为各个微博簇提取一段文本描述. 之后利用该描述向微博簇中加入其他微博信息, 最后利用带有质疑更正信号的微博比例等统计特征对微博簇进行排序.

(4) GRU-2 模型 [7]. 该模型先利用算法对微博事件进行时间段分割, 然后采用 tf-idf 方法计算每个时间段的文本表示, 最后采用双层的 GRU 网络来学习各个微博事件的隐层表示, 并进一步实现对微博事件的分类.

(5) CAMI 模型 [8]. 该模型先将微博事件划分为等长的时间段, 并采用 doc2vec 方法获取每个时间段的文本表示, 构建微博事件的特征矩阵, 之后利用卷积神经网络学习微博事件的隐层表示. 进而进行微博事件的分类.

(6) CSI 模型 [18]. 该模型是一种混合深度模型, 分成 3 个模块, 第 1 个模块采用 LSTM 网络来学习微博事件的隐层表示, 第 2 个模块采用奇异值分解方法来获得全局用户特征, 第 3 个模块融合前两个模块的特征并进行微博事件的分类.

表 4 实验结果表 (R: 谣言, N: 非谣言)[a)]

Table 4 Rumor detection results (R: rumor, N: non-rumor)[a)]

Method	Class	Sina Weibo				Twitter			
		Accuracy	Precison	Recall	$F1$	Accuracy	Precison	Recall	$F1$
DT-Rank	R	0.732	0.738	0.715	0.726	0.614	0.618	0.604	0.611
	N		0.726	0.749	0.737		0.609	0.623	0.616
DTC	R	0.831	0.847	0.815	0.831	0.709	0.690	0.772	0.729
	N		0.815	0.847	0.830		0.733	0.643	0.685
SVM-TS	R	0.857	0.839	0.885	0.861	0.716	0.689	0.793	0.738
	N		0.878	0.830	0.857		0.754	0.639	0.692
GRU-2	R	0.910	0.876	0.956	0.914	0.723	0.712	0.743	0.727
	N		0.952	0.864	0.906		0.735	0.704	0.719
CAMI	R	0.933	0.921	0.945	0.933	0.752	0.722	**0.814**	0.765
	N		0.945	0.921	0.932		0.790	0.690	0.737
CSI	R	0.953	0.930	**0.976**	0.954	0.773	**0.806**	0.714	0.758
	N		**0.977**	0.931	0.953		0.746	**0.831**	**0.787**
HAN-FC	R	**0.968**	**0.966**	0.974	**0.970**	**0.787**	0.778	0.800	**0.789**
	N		0.971	**0.962**	**0.967**		**0.797**	0.775	0.786

a) Values in bold represent the best result in each category among all methods.

4.5 实验结果分析

4.5.1 本文方法与基准方法的谣言检测效果对比

为了验证本文方法的有效性, 将本文方法和现有的社交媒体谣言检测方法进行对比, 实验结果如表 4 所示, 表格从上到下分别是 DT-Rank, DTC, SVM-TS, GRU-2, CAMI, CSI 模型, 以及本文模型 HAN-FC (hierarchical attention network with features combination). 在表 4 中, 类别 (class) 分为 R 和 N, 分别代表谣言 (rumor) 和非谣言 (non-rumor). 在文献 [18] 中只采用了正确率和谣言类别的 $F1$ 值作为评价指标, 而本文的评价指标更为精细, 因此在表 4 中展示的是按照本文评价指标复现实验时所得到的实验结果.

从表 4 可以看出, 对于基准方法, SVM-TS 模型在新浪微博和 Twitter 数据集上分别达到了 85.7% 和 71.6% 的正确率, 在基于传统机器学习的方法中取得了最好的效果. 而在基于特征表示学习的方法中, CSI 模型在新浪微博和 Twitter 数据集上分别达到了 95.3% 和 77.3% 的正确率, 获得了基准方法中最好的效果. 说明基于特征表示学习的方法与基于传统机器学习的方法相比有着明显的优势. 而本文所提出的方法在新浪微博和 Twitter 数据集上正确率分别达到 96.8% 和 78.7%, 分别比 CSI 模型提高了 1.5% 和 1.4%, 分别比 SVM-TS 模型提高了 11.1% 和 7.1%, 取得了比基准方法更好的谣言检测效果. 另外, 从表 4 中还可以看出新浪微博数据集上的性能与 Twitter 数据集上有较大的差异, 这可能是由于 Twitter 数据集上有更多的噪声信息 [7] 且部分信息无法获取对时序信息完整性造成了一定影响.

此外, 从表 4 的实验结果还可以看出, 在新浪微博数据集上, 本文提出的方法在正确率和 $F1$ 值两个评价指标上均高于基准方法中的最优值. 与基准方法中最好的 CSI 模型相比, 针对谣言事件的 $F1$ 值提高了 1.6%. 在 Twitter 数据集上, 本文提出的方法在正确率上及针对谣言事件的 $F1$ 值上均高于基准方法中的最优值. 与基准方法中最优的 CSI 模型相比, 针对谣言事件的 $F1$ 值提高了 3.1%.

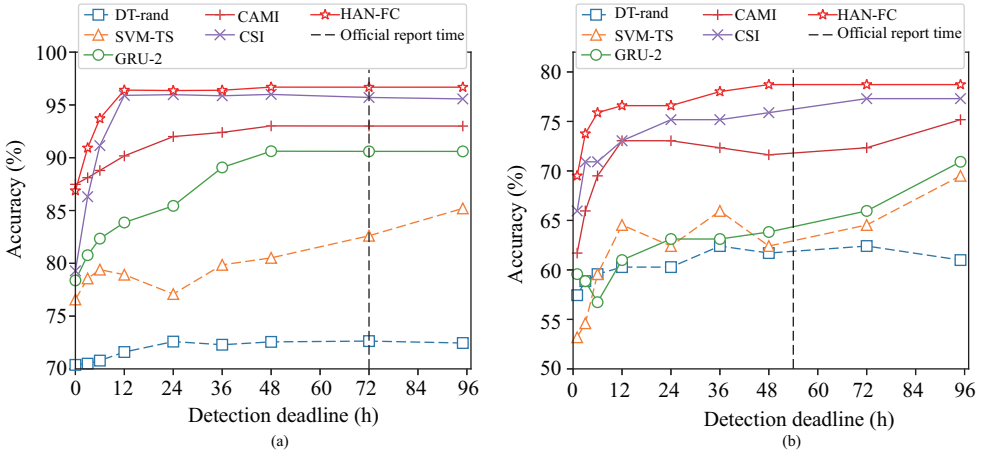

图 2　(网络版彩图) 早期谣言检测结果

Figure 2　(Color online) Results of rumor early detection. (a) Sina Weibo dataset; (b) Twitter dataset

综上所述, 与基准方法相比, 本文提出的方法能够在社交媒体谣言检测问题上表现出更好的效果, 由此验证了本文方法对社交媒体谣言检测问题的有效性.

4.5.2　本文方法与基准方法在早期谣言检测任务中的效果对比

早期谣言检测 (rumor early detection) 任务是在信息传播生命周期的早期阶段, 检测事件是否为谣言. 在该任务中, 只能将事件相关的第 1 条微博发布后若干个小时内的微博信息作为原始数据.

为了评估本文方法在早期谣言检测任务中的效果, 本文选取了从事件相关的第 1 条微博发出后的 9 个时间点 (分别取 1, 3, 6, 12, 24, 36, 48, 72, 96 小时). 在每个时间点, 可以确定一个从事件相关的第 1 条微博发布时间至该时间点的时间范围, 测试集上只采用这个时间范围内的相关微博信息作为输入, 并采用同一个已训练好的模型对测试集上的事件分类. 此外, 本文与文献 [7, 8, 17] 同样将 72 小时和 54 小时分别作为新浪微博和 Twitter 上的平均官方媒体辟谣时间 (official report time). 本文方法以及 DT-Rank, SVM-TS, GRU-2, CAMI, CSI 模型在各个时间点的谣言检测效果如图 2 所示.

从图 2 中可以观察到, 在 12 个小时内, 各个模型的谣言检测正确率总体上都呈现出不同梯度的上升状态. 说明在事件相关信息较少时, 模型均有较低的正确率, 谣言检测效果较差. 正确率上升的梯度越大, 说明模型对该时间范围内的微博信息的依赖性越强, 在模型接收到该时间范围内的微博信息后, 能够明显提高正确率, 模型可以表现出更好的效果.

从图 2 中还可以看出, 在经过 12~48 小时后, HAN-FC 等大多数基于特征表示学习的方法的正确率已基本趋于稳定. 说明基于特征表示学习的方法在早期谣言检测任务中能够更快发挥出模型的最高性能. 另外, 在达到 96 小时后, 各个模型的正确率与表 4 所展示的正确率较为接近. 说明达到 96 小时后, 模型已获得了较为丰富的微博信息来检测事件是否为谣言, 在新浪微博数据集和 Twitter 数据集上均能够表现出与获得所有微博信息时相近的谣言检测效果.

此外, 在图 2 中可以观察到, 在新浪微博数据集上, 本文模型以及 CSI 模型的正确率均在 12 小时附近达到稳定点, 而 CAMI, GRU-2, 以及 DT-Rank 模型在 24~48 小时范围内达到稳定点, SVM-TS 模型在 48 小时后正确率仍然在有所提升. 本文模型以及 CSI 模型在最短时间内达到了稳定点, 且比平均官方媒体辟谣时间早了约 60 个小时. 而在 Twitter 数据集上, 本文模型正确率达到稳定点的时间点推后至 48 小时附近, DT-Rank 模型等在 48 小时后正确率仍在提升. 说明相比基准方法, 本文模

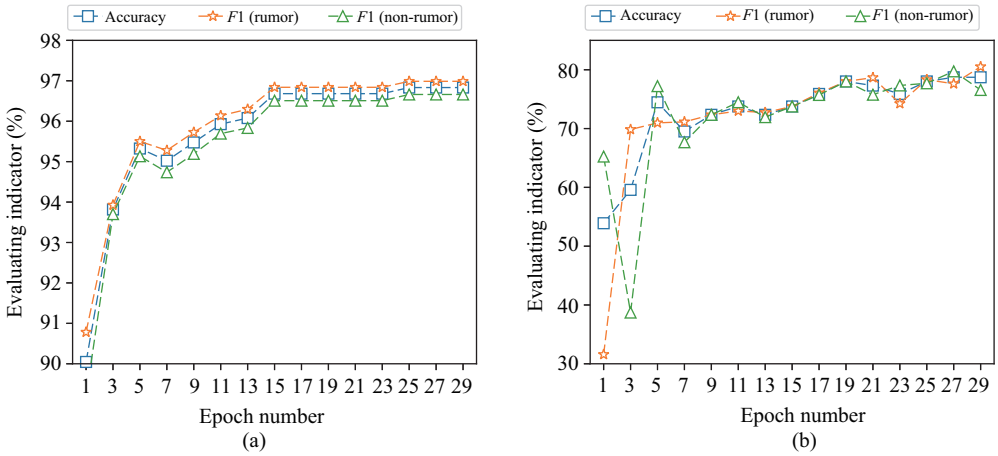

图 3 (网络版彩图) 迭代次数对实验结果的影响

Figure 3 (Color online) The effects of epoch num on experimental results. (a) Sina Weibo dataset; (b) Twitter dataset

型以及 CSI 模型所依赖的微博信息最少, 能够在最短时间内表现出模型的最高性能. 另外, 本文模型在各个时间点基本都能取得相比于 CSI 等模型更高的谣言检测正确率, 说明本文模型所提取到的时间段内微博间的时序信息及融入的文本潜在信息和局部用户信息有效地优化了事件中时间段的表示, 使得时间段的表示能够更为细致地刻画该时间段内的信息, 从而更早捕获到能够有效区分谣言与非谣言的潜在信息, 提升早期谣言检测的效果.

综上所述, 与基准方法相比, 本文方法能够在更短时间内发挥出模型的最高性能, 且能获得更好的谣言检测效果, 由此可以验证本文方法在早期谣言检测任务上的有效性.

4.5.3 模型训练的迭代次数对谣言检测效果的影响

为了分析模型训练的迭代次数对谣言检测效果的影响, 在实验中, 本文设置了 15 组不同的迭代次数. 对于每一组迭代次数, 分别在两个数据集上对模型进行训练和测试, 并根据测试集上正确率、谣言类别下的 $F1$ 值、非谣言类别下的 $F1$ 值这 3 个评价指标绘制了如图 3 所示的折线图.

从图 3 可以看出, 当迭代次数为 1 次时, 模型在 3 个评价指标上都相对处于较低的数值, 此时模型无法有效地学习微博事件的潜在表示, 谣言检测效果相对较低. 之后, 随着模型训练的迭代次数的增加, 3 个评价指标总体上表现出先逐步提升然后在最优值附近逐渐平稳的趋势. 在新浪微博数据集上, 随着迭代次数从 1 次逐步增加至 5 次, 3 个评价指标均以相近的梯度快速提升; 在迭代次数从 5 次逐步增加至 25 次的过程中, 3 个评价指标均上下波动并缓慢提升, 波动的幅度也随着迭代次数的增加逐渐减小; 在迭代次数增加至 25 次以上后, 3 个评价指标已基本达到最优值并趋于平稳. 在 Twitter 数据集上, 评价指标也表现出与新浪微博数据集上相似的变化趋势. 上述现象表明, 模型训练的迭代次数对谣言检测的效果有重要的影响. 随着迭代次数增加, 谣言检测的效果得到了逐步的提升, 并在一定迭代次数后趋于稳定, 此时模型表现出了最好的谣言检测性能.

4.5.4 本文所选取的特征对社交媒体谣言检测任务的有效性分析

为了分析本文所选取的特征对社交媒体谣言检测任务的有效性, 本文对所选取的带有质疑更正信号的微博比例、提供个人描述的用户比例、认证的用户比例、用户声望得分这 4 个特征在两个数据集

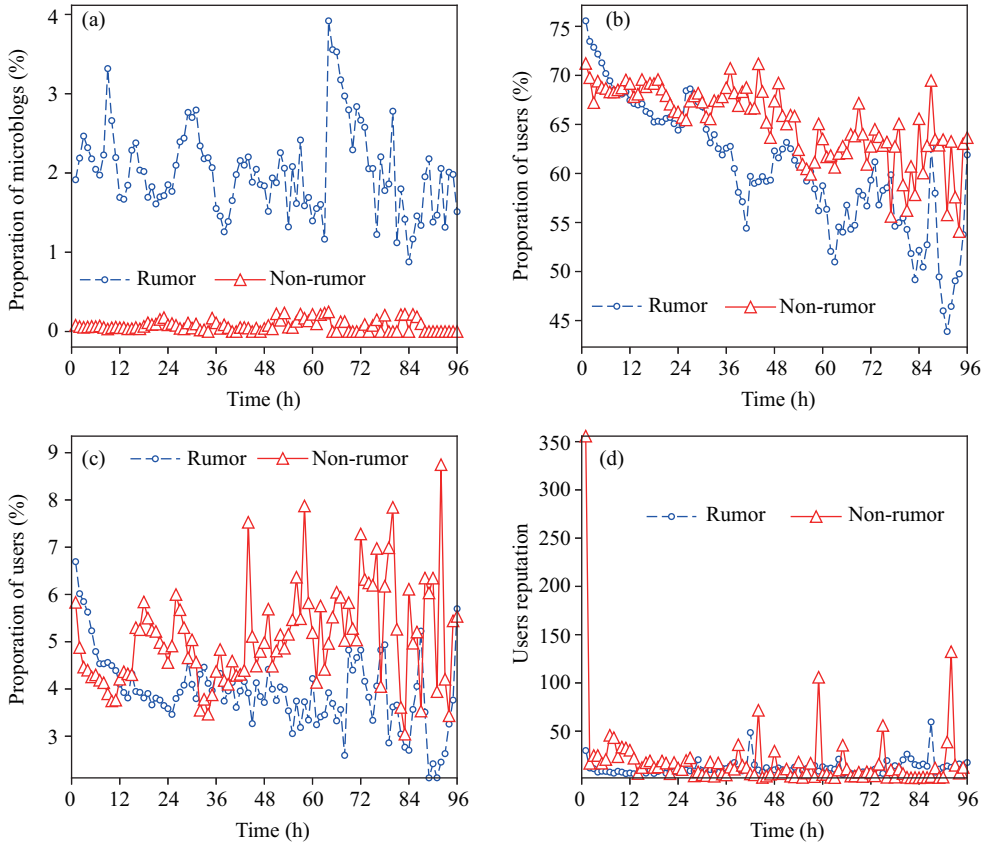

图 4　(网络版彩图) 新浪微博数据集上各个特征随时间的变化情况

Figure 4　(Color online) The features changing over time on Sina Weibo dataset.　(a) Enquiries and corrections; (b) personal description; (c) verified users; (d) users reputation

上对谣言和非谣言的区分能力进行了统计分析. 首先将数据集上的事件集合划分为谣言事件集合和非谣言事件集合, 然后在两个集合上对每个小时内的微博信息分别统计上述的 4 个特征, 最后根据统计结果绘制了 96 个小时内各个特征随时间变化曲线图, 如图 4 和 5 所示.

从图 4 中可以看出, 在新浪微博数据集上, 相比于非谣言事件, 谣言事件中带有质疑更正信号的微博比例在各个时间点总是维持在相对较高的水平. 这是由于随着时间的推移, 谣言事件在传播过程中受到了越来越多用户的质疑和更正, 这与非谣言事件有着明显的区别. 另外, 非谣言事件中用户提供个人描述的比例、用户认证的比例, 以及用户声望这 3 个特征在大部分时间点都高于谣言事件. 这是由于有提供个人描述信息或是进行了认证的用户通常是社交媒体中的活跃用户, 而用户声望 (平均粉丝数与平均关注数的比值) 较高的用户通常是一些明星、企业账号等较为活跃的用户, 用户声望较低的用户往往不活跃, 甚至可能是社交媒体中的水军, 这些用户相比于活跃用户更可能传播谣言. 综合上述分析, 这 4 个特征能够很好地捕获到新浪微博数据集上谣言和非谣言的差异, 对于谣言和非谣言具有很好的区分能力.

而从图 5 中可以观察到, 在 Twitter 数据集上, 谣言事件与非谣言事件在 4 个特征上的差异较为不明显, 这是由于 Twitter 数据集上的噪声更大 [7], 且实验数据集中未获取到的部分 Tweet 对 4 个特征的统计结果造成了一定影响. 但谣言事件中带有质疑更正信号的微博比例仍然在大部分时间点高于非谣言事件, 而非谣言事件中提供个人描述的用户比例、认证的用户比例、用户声望得分也在大部分

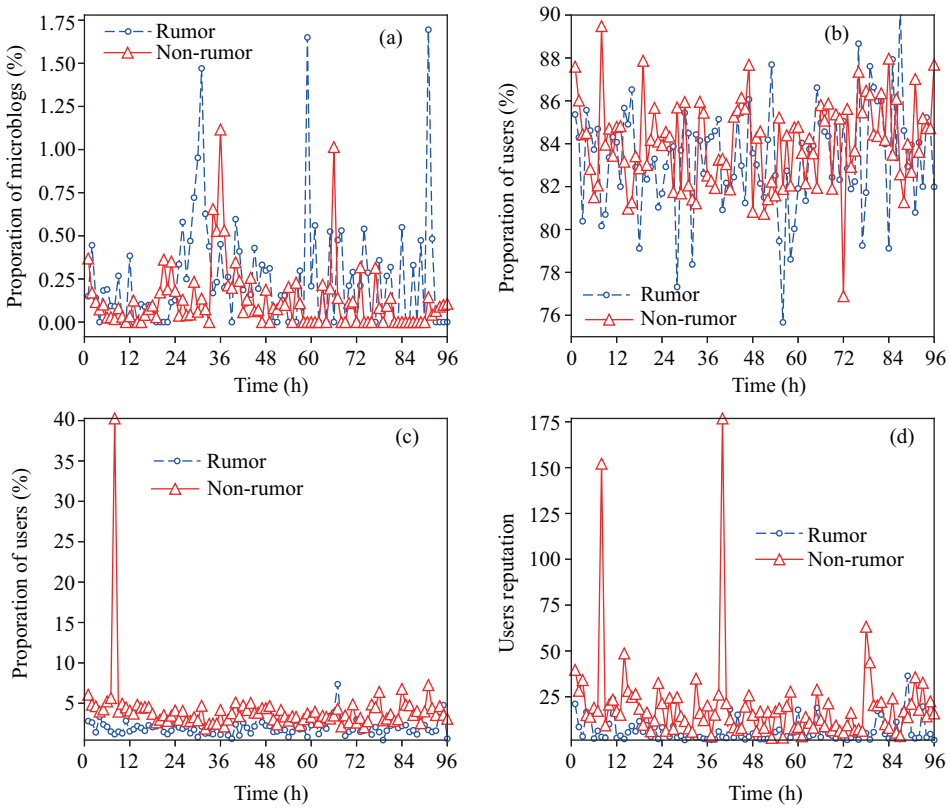

图 5 (网络版彩图) Twitter 数据集上各个特征随时间的变化情况

Figure 5 (Color online) The features changing over time on Twitter dataset. (a) Enquiries and corrections; (b) personal description; (c) verified users; (d) users reputation

时间点相比于谣言事件更高, 谣言事件和非谣言事件表现出与新浪微博数据集上类似的差异. 类似地, 微博平均长度和用户积极性这两个特征也同样能够在一定程度上捕获到这种差异, 在此不做更细致的分析. 因此, 这些特征在 Twitter 数据集上仍然能够较好地捕获到谣言和非谣言的差异, 对于谣言和非谣言具有较好的区分能力.

综上所述, 这些特征在两个数据集上都能够较好地捕获到谣言事件和非谣言事件在信息传播过程中的文本、用户等方面的差异, 较为有效地区分谣言事件和非谣言事件, 体现了这些特征在社交媒体谣言检测任务上的有效性.

5 结束语

本文提出了一种基于分层注意力网络的社交媒体谣言检测方法, 通过采用两层带有注意力机制的双向 GRU 网络从微博层面和时间段层面分别获取微博序列的隐层表示和时间段序列的隐层表示, 从而在事件的特征表示中融入了时间段内各微博间的时序信息. 此外, 还针对各个时间段提取了局部用户特征及文本潜在特征, 并将这些特征融入到时间段的表示中, 进一步捕获这些特征随时间变化的隐层表达. 在新浪微博和 Twitter 这两个公开数据集上, 与目前最好的基准方法相比, 本文方法取得了更高的正确率, 验证了本文方法在社交媒体谣言检测问题上的有效性. 同时, 在早期谣言检测任务中, 本

文方法也表现出了更好的性能. 在进一步的研究中, 计划结合强化学习探究在模型训练过程中自动对事件进行时间段划分的方法, 以避免人为划分可能带来的信息丢失问题, 进一步提升谣言检测的效果.

参考文献

1 Liu Z Y, Zhang L, Tu C C, et al. Statistical and semantic analysis of rumors in Chinese social media. Sci Sin Inform, 2015, 45: 1536–1546 [刘知远, 张乐, 涂存超, 等. 中文社交媒体谣言统计语义分析. 中国科学: 信息科学, 2015, 45: 1536–1546]

2 Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science, 2018, 359: 1146–1151

3 Waldrop M M. News feature: the genuine problem of fake news. Proc Natl Acad Sci USA, 2017, 114: 12631–12634

4 Tan Z H, Shi Y C, Shi N X, et al. Rumor propagation analysis model inspired by gravity theory for online social networks. J Comput Res Dev, 2017, 54: 2586–2599 [谭振华, 时迎成, 石楠翔, 等. 基于引力学的在线社交网络空间谣言传播分析模型. 计算机研究与发展, 2017, 54: 2586–2599]

5 Liu Y H, Jin X L, Shen H W, et al. A survey on rumor identification over social media. Chinese J Comput, 2018, 41: 1536–1558 [刘雅辉, 靳小龙, 沈华伟, 等. 社交媒体中的谣言识别研究综述. 计算机学报, 2018, 41: 1536–1558]

6 Chen Y F, Li Z Y, Liang X, et al. Review on rumor detection of online social networks. Chinese J Comput, 2018, 41: 1648–1677 [陈燕方, 李志宇, 梁循, 等. 在线社会网络谣言检测综述. 计算机学报, 2018, 41: 1648–1677]

7 Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, 2016. 3818–3824

8 Yu F, Liu Q, Wu S, et al. A convolutional approach for misinformation identification. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, 2017. 3901–3907

9 Ma J, Gao W, Wong K F. Detect rumor and stance jointly by neural multi-task learning. In: Proceedings of the Web Conference Companion, Lyon, 2018. 585–593

10 Zhang Q, Zhang S Y, Dong J, et al. Automatic detection of rumor on social network. In: Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing, Nanchang, 2015. 113–122

11 Castillo C, Mendoza M, Poblete B. Information credibility on twitter. In: Proceedings of International Conference on World Wide Web, Hyderabad, 2011. 675–684

12 Yang F, Liu Y, Yu X H, et al. Automatic detection of rumor on Sina Weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, Beijing, 2012. 13–20

13 Zhao Z, Resnick P, Mei Q Z. Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide Web, Florence, 2015. 1395–1405

14 Liang G, He W B, Xu C, et al. Rumor Identification in Microblogging systems based on users' behavior. IEEE Trans Comput Soc Syst, 2015, 2: 99–108

15 Ma J, Gao W, Wong K F. Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017. 708–717

16 Sun S Y, Liu H Y, He J, et al. Detecting event rumors on Sina Weibo automatically. In: Proceedings of the Web Technologies and Applications, Sydney, 2013. 120–131

17 Ma J, Gao W, Wei Z Y, et al. Detect rumors using time series of social context information on microblogging websites. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, 2015. 1751–1754

18 Ruchansky N, Seo S, Liu Y. CSI: a hybrid deep model for fake news detection. In: Proceedings of ACM on Conference on Information and Knowledge Management, Singapore, 2017. 797–806

19 Qazvinian V, Rosengren E, Radev D R, et al. Rumor has it: identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, 2011. 1589–1599

20 Hamidian S, Diab M. Rumor identification and belief investigation on Twitter. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, San Diego, 2016. 3–8

21 Liu Y H, Jin X L, Shen H W, et al. Do rumors diffuse differently from non-rumors? a systematically empirical analysis in Sina Weibo for rumor identification. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Jeju, 2017. 407–420

22 Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature

detectors. Comput Sci, 2012, 3: 212–223
23 Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Vancouver, 2013. 6645–6649
24 LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput, 1989, 1: 541–551
25 Elman J L. Finding structure in time. Cogn Sci, 1990, 14: 179–211
26 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780
27 Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, 2014. 1746–1751
28 Chen H M, Sun M S, Tu C C, et al. Neural sentiment classification with user and product attention. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Austin, 2016. 1650–1659
29 Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologiesm, San Diego, 2016. 1480–1489
30 Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, 2014. 1724–1734
31 Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the International Conference on Machine Learning, Beijing, 2014. 1188–1196
32 Kingma D P, Ba J. Adam: a method for stochastic optimization. 2014. ArXiv:1412.6980

Rumor detection in social media based on a hierarchical attention network

Xiangwen LIAO[1,2,3*], Zhi HUANG[1,2,3], Dingda YANG[1,2,3], Xueqi CHENG[4] & Guolong CHEN[1,2,3]

1. *College of Mathematics and Computer Science, Fuzhou University, Fuzhou* 350116*, China*;
2. *Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University), Fuzhou* 350116*, China*;
3. *Digital Fujian Institute of Financial Big Data, Fuzhou* 350116*, China*;
4. *Key Laboratory of Web Data Science and Technology, Chinese Academy of Sciences, Beijing* 100086*, China*
* Corresponding author. E-mail: liaoxw@fzu.edu.cn

Abstract For rumor detection in social media, the majority of feature-representation-based studies capture textual features or global users' features according to a partitioned sequence of microblogs. However, these studies ignore the time-series information across the microblogs in one time interval. Moreover, the latent textual information and the local users' information, which have been proven effective according to the traditional machine learning methods, are overlooked in capturing the time intervals' features, resulting in low performance. Therefore, we propose a rumor detection method in social media based on a hierarchical attention network. First, microblogs are partitioned into several time intervals. Then, the variation of information across the microblogs changing over time is learned using a bidirectional gated recurrent unit neural network with an attention mechanism. After that, the variation of features across the microblogs is combined with hand-crafted features to incorporate latent textual information and local users' information into the time intervals' features. Finally, we capture features' variation across the time intervals by using the bidirectional gated recurrent unit neural network, with the attention mechanism, and classify microblog events. Experimental results over two public datasets, Sina Weibo and Twitter, show that the proposed method outperforms (in terms of the accuracy) state-of-the-art methods by 1.5% and 1.4% over the two datasets, respectively, and it is effective for rumor detection.

Keywords rumor detection, hierarchical attention network, social media, time series information, deep learning

Xiangwen LIAO born in 1980, Ph.D. degree, Associate Professor, CCF senior member. His research interests include information retrieval, opinion mining, sentiment analysis, and natural language processing.