基于分层语义特征学习模型的微博谣言事件检测*

黄学坚[1,2]　马廷淮[1]　王根生[3]

[1](南京信息工程大学软件学院 南京 210044)

[2](江西财经大学虚拟现实(VR)现代产业学院 南昌 330013)

[3](江西财经大学人文学院 南昌 330013)

摘要:【目的】提高微博谣言事件检测的准确率和时效性。【方法】提出一种基于分层语义特征学习模型的微博谣言事件检测方法。首先,基于BERT预训练模型抽取事件中单条文本信息的语义特征;其次,基于时间域对事件传播数据进行动态划分,利用卷积神经网络挖掘各时间域中的文本集合的语义相关性特征;然后,把各时间域内的语义相关性特征输入深层双向门控循环神经网络,学习事件传播过程中的深层语义时序特征;最后,融合Attention机制使模型更加关注于语义时序特征中具有谣言特征的部分。【结果】在Weibo公开数据集上的实验结果表明,该模型的检测准确率达到95.39%,检测时延在12h以内。【局限】模型需要一定数量的转发评论信息,事件热度不够时检测效果不突出。【结论】分层语义特征学习模型实现了从局部语义到全局语义的学习过程,提升了微博谣言事件检测的性能。

关键词: 谣言检测　深度学习　语义特征　时序数据　分层语义

分类号: TP393　G250

DOI: 10.11925/infotech.2096-3467.2022.0613

引用本文: 黄学坚, 马廷淮, 王根生. 基于分层语义特征学习模型的微博谣言事件检测[J]. 数据分析与知识发现, 2023, 7(5): 81-91.(Huang Xuejian, Ma Tinghuai, Wang Gensheng. Detecting Weibo Rumors Based on Hierarchical Semantic Feature Learning Model[J]. Data Analysis and Knowledge Discovery, 2023, 7(5): 81-91.)

1　引　言

随着移动互联网的快速发展,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。微博方便人们信息交流的同时,也为谣言的传播提供了新渠道。社交媒体中,谣言具有传播速度快、影响范围广、危害程度深等特点,谣言的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]。为了遏制谣言的传播,微博官方建立了微博社区管理中心,为广大用户提供谣言举报、查证等功能。这种依赖于用户举报和人工验证的方式,不仅耗费大量的人力物力,而且可能存在严重的时间滞后问题。因此,研究微博谣言自动检测方法,在谣言传播的潜伏期内进行有效识别,具有重要的意义。

早期的谣言检测主要基于人工特征工程的统计机器学习方法,研究重点在于谣言特征的选择,由先验知识指导设计的手工特征缺乏全面性和灵活性,难以表示谣言的深层语义特征。随着深度学习技术的发展,研究者开始利用深度学习模型自动学习谣言的深层语义特征,谣言检测逐渐迎来数据

通讯作者(Corresponding author):马廷淮(Ma Tinghuai),ORCID:0000-0003-2320-1692,E-mail:thma@nuist.edu.cn。

*本文系国家重点研发计划(项目编号:2021YFE0104400)、国家自然科学基金项目(项目编号:72061015)和江西省教育厅科技项目(项目编号:GJJ200539)的研究成果之一。

The work is supported by the National Key R&D Program of China (Grant No.2021YFE014400), the National Natural Science Foundation of China (Grant No. 72061015), the Science and Technology Project of Jiangxi Provincial Department of Education(Grant No. GJJ200539).

驱动时代。目前，基于语义特征的深度学习谣言检测方法重点关注博文内容，把谣言检测视为单文本分类问题。然而，有时谣言散布者故意模仿真实信息的写作特点和风格，单纯基于文本语义特征的方法并不能有效识别谣言。微博传播过程中的转发和评论信息，包含其他用户对该微博的态度，如怀疑、反对、赞成等，这些评论态度相互影响和关联，对谣言的判断起到重要作用。所以，本文在单文本语义特征的基础上融合转发和评论信息，提出一种基于分层语义特征学习模型的微博谣言事件检测方法，依次学习单文本语义特征、文本集合语义相关性特征、事件传播语义时序特征，弥补单文本语义特征方法的不足。实验结果证明了模型的有效性。

2 相关研究

目前，国内外对谣言自动检测的研究主要分为基于人工特征工程的统计机器学习方法和基于语义特征的深度学习方法。

（1）基于人工特征工程的统计机器学习方法的研究重点在于谣言特征的选择。例如，Castillo等[2]对Twitter的谣言数据进行统计分析，基于最佳优先属性选择策略，从内容、用户、传播和主题4个属性空间中选择了15个谣言分类特征；Yang等[3]在Castillo等研究的基础上引入事件发生位置和发布信息客户端类型两个新特征；贺刚等[4]在Yang等的基础之上，又提出了符号特征、关键词分布特征、链接特征和时间差等新特征。在静态统计特征的基础上，有些研究者又提出了动态特征、情感特征和谣言相似度特征等。例如，Ma等[5]针对传统统计特征忽视了特征随时间的变化，提出了一种动态时间序列结构模型，捕获多种上下文特征随时间的变化；祖坤琳等[6]在传统统计特征的基础上，引入微博的评论情感倾向特征；马鸣等[7]将待检测样本与官方谣言子集中的微博进行相似度计算，将其值与统计特征进行融合。在分类方法选择上，主要有决策树、支持向量机（Support Vector Machine，SVM）和随机森林等。例如，Wu等[8]和Ma等[9]基于谣言的传播特征，建立谣言传播树形结构，使用基于图模式的SVM分类器进行谣言检测；曾子明等[10]在传播特征和用户特征的基础上，采用隐含狄利克雷分布（Latent Dirichlet Allocation，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构建谣言识别模型。

（2）随着深度学习技术的快速发展，研究者逐渐利用深度学习模型自动学习谣言的深层语义特征表示并进行分类。例如，Yu等[11]利用卷积神经网络（Convolutional Neural Network，CNN）挖掘谣言文本深层次特征；Ajao等[12]提出卷积神经网络和长短期记忆网络（Long-Short Term Memory，LSTM）相结合的模型，检测Twitter上发布的虚假信息；李奥等[13]和Ma等[14]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。这一类的深度学习方法重点关注谣言文本的语义特征，把谣言检测视为单文本分类问题。然而，有些谣言可能不具有典型的谣言文本语义特征，单纯基于文本分类的方法有时并不能获得很好的效果。针对这个问题，一些研究者在单文本语义特征的基础上融合传播结构特征、用户特征和语言学特征等[15]。例如，Tu等[16]和Ke等[17]提出了融合语义信息和传播结构的谣言检测模型；Ma等[18]从语言学、用户和传播结构收集统计特征，并把这些统计特征和文本语义特征联合生成特征图谱进行谣言检测；尹鹏博等[19]结合用户属性和微博文本，分别利用CNN和LSTM挖掘用户特征和文本语义特征，根据用户特征和语义特征的融合进行谣言分类。

（3）研究评述

基于人工特征工程的统计机器学习方法十分依赖于特征的选取，由先验知识指导设计的手工特征缺乏全面性和灵活性，难以表示谣言的深层语义特征。目前，基于语义特征的深度学习方法成为研究热点，但已有方法关注于博文内容，把谣言检测视为单文本分类问题，在某些场景下并不能获得很好的效果。微博传播过程中的转发和评论信息，包含其他用户对该微博的态度，这些评论态度相互影响和关联，对谣言的判断起到重要作用[20-22]。但是，转发和评论信息也面临数据量过多的问题，并且有些信息对谣言的判别没有贡献，所以，本文提出一种基于

分层语义特征学习模型（BERT-CNN-GRU-Attention，BCGA）的微博谣言事件检测方法。首先，基于BERT（Bidirectional Encoder Representation from Transformers）学习单条转发文本语义特征，得到单个用户对该事件的态度；其次，针对转发评论数量过多导致门控循环神经网络（Gated Recurrent Unit，GRU）长距离遗忘的问题，通过时间窗口对评论进行分组，利用CNN学习组内用户评论的语义相关特征，得到该时段内用户的总体态度；然后，通过GRU学习不同时段内总体态度的语义时序特征，得到用户总体态度的变化情况；最后，通过Attention使模型更加关注于语义时序特征中具有谣言特征的部分，整个模型实现了从微观层面分析到宏观层面分析的谣言判别。

3　BCGA模型构建

假定一个事件由一条原始微博及其传播过程中的转发数据和评论数据组成，把原始微博、转发数据和评论数据统称为事件信息。一个事件集合 $E=\{e_1, e_2, \cdots, e_i\}$，其中 $e_i=\{(m_0, t_0), (m_1, t_1), \cdots, (m_j, t_j)\}$ 表示一个微博事件，事件 e_i 中包含若干信息 m_j 及对应的时间戳 t_j，起始时间点 t_0 的信息 m_0 为原始微博信息，t_0 时间点以后的信息为转发或评论信息，转发和评论信息不做区分。谣言检测本质上是一个二分类问题，即学习一个分类模型 $f: e_i \rightarrow l_j$，将每个事件 e_i 映射为一个类别标签 $l_j \in \{0, 1\}$，模型输入为 e_i 的信息序列 $(m_0, m_1, \cdots, m_j)$，输出为事件 e_i 是否为谣言。本文提出的BCGA模型如图1所示，模型主要包括五大模块。

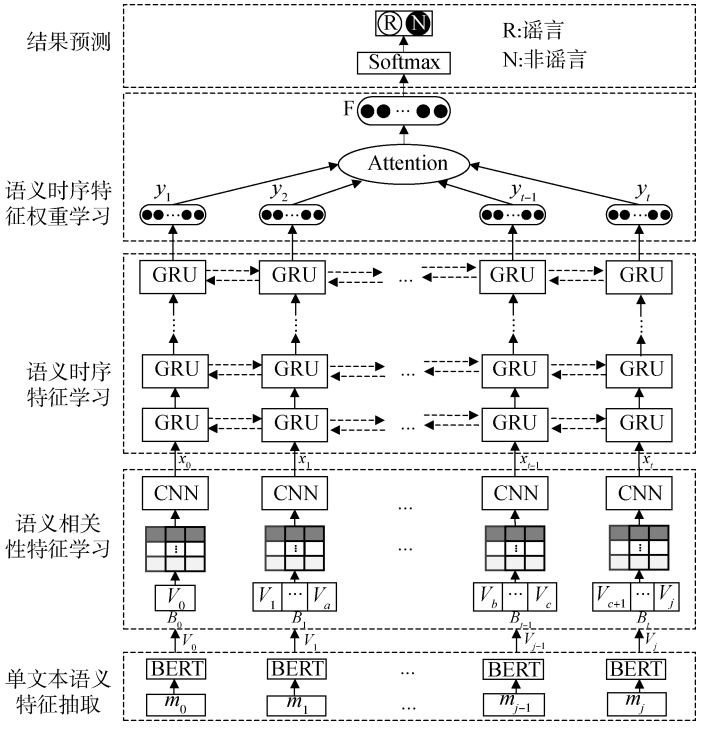

图1　BCGA模型

Fig.1　BCGA Model

（1）单文本语义特征抽取。基于BERT预训练模型，抽取事件 e_i 中所有文本信息 m_j 的语义特征 $\boldsymbol{v}_j$。

（2）语义相关性特征学习。首先，基于时间域对事件信息进行动态划分 $e_i=\{B_0, B_1, \cdots, B_t\}$，$B_t=\{\boldsymbol{v}_j, \boldsymbol{v}_{j+1}, \cdots, \boldsymbol{v}_{j+n}\}$ 为第 t 个时间块中的文本语义特征集合；然后，把信息块 B_t 表示成 $n \times k$ 的二维矩阵，n 为信息块 B_t 中包含的文本个数，k 为文本语义特征向量 $\boldsymbol{v}_j$ 的维度；最后，把信息块的二维矩阵表示输入

CNN中，学习信息块B_t中文本的语义相关性特征x_t。

（3）语义时序特征学习。把每个信息块B_t的语义相关性特征x_t依次输入深层双向GRU网络中，学习语义信息的时序特征y_t。

（4）语义时序特征权重学习。利用Attention机制对GRU网络的最后一层各时刻的输出y_t赋予不同的权重，得到最终分层语义特征F。

（5）结果预测。把特征F输入Softmax分类器中，判断事件是否为谣言。

3.1 单文本语义特征抽取

2018年，谷歌发布了BERT[23]模型，BERT采用了Transformer的双向Encoder结构，舍去了循环神经网络（Recurrent Neural Network，RNN）结构，完全基于Self-Attention对一段文本进行建模。为了获取比词更高级别的句子级别的语义特征，BERT的训练联合了MLM（Masked Language Model）和NSP（Next Sentence Prediction）两类任务。BERT相比于Word2Vec模型，一方面，考虑了上下文语境，解决了一词多义的问题；另一方面，通过分层学习得到不同层次的语义特征，为下游任务提供了丰厚的特征选择。目前，很多第三方平台开源了BERT的预训练模型，基于BERT预训练模型下游任务可以进行微调（改变BERT预训练模型参数）或特征抽取（不改变BERT预训练模型参数，只是把预训练模型的输出作为特征输入下游任务）。本文基于BERT预训练模型，抽取事件$e_i=\{(m_0,t_0),(m_1,t_1),\cdots,(m_j,t_j)\}$中所有文本信息$m_j$的语义特征向量$\boldsymbol{v}_j=\{d_1,d_2,\cdots,d_k\}$，作为谣言检测任务的特征输入。

3.2 语义相关性特征学习

（1）事件信息分块

一个微博事件中包含的转发和评论信息可能数以万计，虽然GRU能在一定程度上解决长距离依赖问题，但如果将数以万计的信息依次输入，GRU也会因梯度消失导致长距离遗忘问题。所以，本文对微博事件信息进行分块处理$e_i=\{B_0,B_1,\cdots,B_t\}$，其中$B_t=\{\boldsymbol{v}_i,\boldsymbol{v}_{i+1},\cdots,\boldsymbol{v}_{i+n}\}$，以信息块$B_t$作为GRU的输入单位。最简单的分块方法是根据数量均等划分，每个块包含的信息数量相同，但这种划分方法容易导致块内信息表意相差较大，影响后续的语义相关性学习。所以，本文提出基于时间域的动态划分方法，具体步骤如下。

①单独以原始微博m_0作为第一个块$B_0=\{\boldsymbol{v}_0\}$。

②计算事件中最后一个信息(m_j,t_j)和第二个信息(m_1,t_1)的时间差$TD=t_j-t_1$。

③根据设定的待划分块数N，计算时间间隔$TI=\dfrac{TD}{N-1}$。

④根据时间间隔TI划分事件信息，统计空块数i（该时间段内没有事件信息）和非空块数j。

⑤如果非空块数j小于待划分块数N，并且大于上一轮迭代中的非空块数，则缩短时间间隔$TI=\dfrac{j}{N}\times TI$，回到步骤④重新划分；否则，返回划分结果。

（2）信息块语义相关性特征学习

得到微博事件信息块$e_i=\{B_0,B_1,\cdots,B_t\}$后，把信息块$B_t=\{\boldsymbol{v}_i,\boldsymbol{v}_{i+1},\cdots,\boldsymbol{v}_{i+n}\}$表示成$n\times k$的二维矩阵，$n$为信息块$B_t$中包含的文本信息个数，$k$为文本语义特征向量$\boldsymbol{v}_j$的维度，把二维矩阵输入CNN，经卷积层和池化层处理后得到信息块语义相关性特征x_t。

不同于传统的二维卷积，本文使用的是一维卷积，即只设置卷积核的高度h，其宽度w固定等于输入矩阵的宽度。这是因为信息块中的每条信息是独立的，不存在上下文关系，需要挖掘的是不同信息间的关系，而不是单条文本内的局部上下文关系。由于每个信息块中包含的信息数量不同，导致输入CNN的矩阵高度不一致，为了保证最终输出数据x_t维度一致，本文在池化层使用1-max-pooling对特征图进行提取，即每个特征图只保留最大值特征，使最终语义相关性特征x_t的维数等于卷积核的个数，与输入矩阵的大小无关。

3.3 语义时序特征学习

微博事件中的信息是相互关联和影响的，存在相关时序特征。GRU是一个有效的时序特性学习模型，可以很好地保持时序数据中长短距离的依赖关系，并且相对LSTM具有更少的模型参数；双向GRU使序列某点的输出同时结合过去和未来的信息；深层GRU相对浅层GRU能够挖掘更深层次的语义时序特征。所以，本文利用深层双向GRU

(Deep Bidirectional GRU, DBiGRU)学习事件传播中的语义时序特征。DBiGRU第l层中t时刻的输出y_t^l的计算如公式(1)所示[24]。

$$y_t^l = \sigma(\boldsymbol{W}_g[\overrightarrow{h_t^l}, \overleftarrow{h_t^l}] + b_y) \tag{1}$$

其中,σ为激活函数;$\boldsymbol{W}_g$为神经网络连接参数;b_y为偏置项参数,$\overrightarrow{h_t^l}$为第l层中前向传播GRU中t时刻的状态h_t;$\overleftarrow{h_t^l}$为第l层中后向传播GRU中t时刻的状态h_t。GRU中t时刻的状态h_t计算过程如公式(2)-公式(5)所示[25]。

$$r_t = \sigma(\boldsymbol{W}_r[h_{t-1}, x_t] + b_r) \tag{2}$$

$$z_t = \sigma(\boldsymbol{W}_z[h_{t-1}, x_t] + b_z) \tag{3}$$

$$\overline{h_t} = \tanh(\boldsymbol{W}_{\overline{h}}[r_t \times h_{t-1}, x_t] + b_{\overline{h}}) \tag{4}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \overline{h_t} \tag{5}$$

其中,$\boldsymbol{W}_r$、$\boldsymbol{W}_z$和$\boldsymbol{W}_{\overline{h}}$分别表示神经网络连接参数;$b_r$、$b_z$和$b_{\overline{h}}$表示偏置项参数;$h_{t-1}$表示前一时刻GRU的状态;$x_t$表示$t$时刻的输入;$r_t$和$z_t$分别表示$t$时刻重置门和更新门的输出;$\overline{h_t}$表示$t$时刻状态更新的内容。

3.4　语义时序特征权重学习

为使模型更加关注语义时序数据中具有谣言特征的部分,本文利用Attention机制对DBiGRU最后一层各时刻的输出y_t赋予不同的权重,然后进行加权求和,得到最终分层语义特征F,如公式(6)所示。

$$F = \sum_{i=1}^{t} \alpha_i y_i \tag{6}$$

其中,α_i表示y_i的权重,其计算过程如公式(7)和公式(8)所示[26]。

$$\boldsymbol{u}_i = \tanh(\boldsymbol{W}_w y_i + b_w) \tag{7}$$

$$\alpha_i = \frac{\exp(\boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{u}_w)}{\sum_{i=1}^{t} \exp(\boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{u}_w)} \tag{8}$$

其中,$\boldsymbol{W}_w$表示神经网络连接参数;b_w表示偏置项参数;$\boldsymbol{u}_w$表示随机初始化权重。

3.5　结果预测

把特征F输入Softmax进行结果预测,如公式(9)所示。

$$p = \mathrm{Softmax}(\boldsymbol{W} \cdot F + b) \tag{9}$$

本文基于最小化交叉熵损失函数和Adam优化算法对整个模型的参数进行训练,交叉熵损失函数计算如公式(10)所示[27]。

$$L = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log p_i + (1 - y_i)\log(1 - p_i)) + \frac{\lambda}{2}\|\boldsymbol{W}\|_2^2 \tag{10}$$

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;$\frac{\lambda}{2}\|\boldsymbol{W}\|_2^2$为L2正则化,降低模型的过拟合程度,并对深层双向GRU采用Dropout正则化方法。

4　实验及分析

4.1　实验数据和评价指标

实验数据选取Ma等[28]公开的Weibo谣言数据集,该数据集被广泛应用于社交媒体谣言检测。该数据集共包含4 664个事件,其中谣言事件2 313个,非谣言事件2 351个,具体统计信息如表1所示。

表1　Weibo谣言数据集统计信息

Table 1　Statistics of the Weibo Rumor Dataset

统计项	数量
事件总数	4 664
谣言事件数	2 313
非谣言事件数	2 351
所有事件转发评论总数	3 805 656
事件平均转发评论数	816
事件最大转发评论数	59 318
事件最小转发评论数	10

使用K折交叉验证的准确率、谣言查准率、谣言查全率、谣言F1值、非谣言查准率、非谣言查全率、非谣言F1值作为模型检测评价指标。

4.2　主要参数设置

本文利用实验探索的方式对相关参数进行设置,具体参数设置如表2所示。

4.3　实验结果分析

(1) 不同深度的DBiGRU实验结果对比

设置DBiGRU的深度分别为1、2、3、4、5、6,其他模型参数和表2保持一致,实验结果如表3所示。

通过实验结果对比发现,当DBiGRU的深度从1逐渐增加到3时,模型的各项指标值也逐步提高,准确率从92.39%分别提高到94.21%(深度为2)和95.39%(深度为3)。而当DBiGRU的深度从3逐渐增加到6时,模型的各项指标值却逐步降低,准确率从95.39%分别降低到93.68%(深度为4)、90.14%

表2　主要参数设置

Table 2　Main Parameter Setting

参数类别	参数名称	参数值
模型参数	BERT模型的层数*L*	12
	BERT模型的Multi-head个数*A*	12
	BERT模型的输出维度*H*	768
	事件信息分块数*N*	50
	CNN卷积核高度*h*	3,4,5
	同一高度下的卷积核个数*m*	80
	双向GRU的层数*L*	3
	正则化参数λ	2e-4
	Dropout的*keep-prob*	0.8
训练参数	学习率*learning_rate*	0.001
	最大迭代次数*epoch_num*	300
	批量训练的*batch_size*	64
	交叉验证*K-fold*	6

表3　不同深度的双向GRU实验结果对比

Table 3　Experimental Results of Bidirectional GRU with Different Layers

深度	准确率/%	类别	查准率/%	查全率/%	F1/%
1	92.39	R	91.18	93.74	92.44
		N	93.65	91.06	92.34
2	94.21	R	92.87	95.68	94.26
		N	95.61	92.77	94.17
3	95.39	R	93.75	97.19	95.44
		N	97.13	93.62	95.34
4	93.68	R	92.26	95.25	93.73
		N	95.16	92.13	93.62
5	90.14	R	89.38	90.93	90.15
		N	90.91	89.36	90.13
6	85.64	R	84.93	86.39	85.65
		N	86.36	84.89	85.62

（深度为5）和85.64%（深度为6）。出现这种结果的原因可能是DBiGRU的深度较浅时，语义时序特征模型相对简单，不能挖掘深层次的语义时序特征，导致模型发生欠拟合，所以深度的适当增加提高了模型的学习能力。然而，当DBiGRU的深度不断增加时，模型的复杂度越来越高，导致模型发生过拟合，所以当DBiGRU的深度超过某个范围继续增加时模型的泛化能力逐步降低。

（2）不同分块方法的实验结果对比

为了验证基于时间域的动态划分方法相比于基于数量均等划分方法的有效性，对比其在分块数量为1、5、10、15、20、25、30、35、40、45、50的模型准确率，实验结果如图2所示。

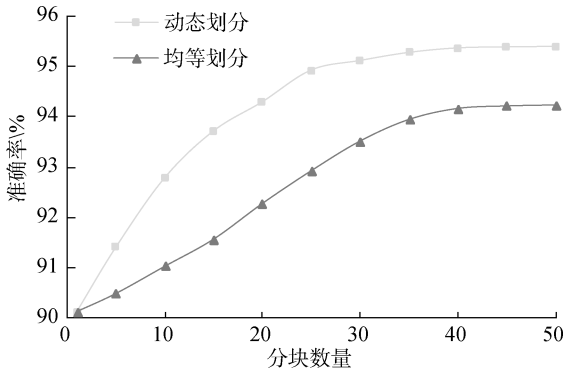

图2　不同分块方法的实验结果对比

Fig.2　Experimental Results of Different Blocking Methods

随着分块数量在一定范围内增加，基于时间域的动态划分方法和基于数量均等划分方法的模型准确率从90.12%分别逐步增加到95.39%和94.23%，因为随着分块的粒度越来越细粒化，模型越能捕获到事件传播过程中的舆情动态变化特征。在相同的分块数量下，基于时间域的动态划分方法的准确率都高于基于数量均等划分方法，这是因为在事件舆情发展的不同阶段，用户的总体态度和参与热度是不一样的，在舆情的发展阶段用户的转发评论信息比较密集，而在舆情的开始阶段和结束阶段转发评论信息比较稀疏，在某个粒度的时间内用户总体态度相对一致，基于数量均等划分方法很容易把相同舆情阶段的信息划分到不同的块中，而基于时间域的动态划分方法把相同阶段的转发评论信息尽量划分到相同的块中，可以缓解块内信息表意相差较大的问题。

（3）不同组合模型的实验结果对比

为了验证BCGA组合模型的有效性，把BCGA模型和以下6种不同组合模型进行实验对比。

① Doc2Vec_CNN_DBiGRU_Attention（DCGA）：把BCGA模型中的BERT替换为Doc2Vec，由Doc2Vec直接生成文本特征向量作为底层特征输入。

② BERT_CNN_DGRU_Attention（BCGA_s）：把BCGA模型中的双向GRU替换为单向GRU。

③ BERT_CNN_DBiLSTM_Attention（BCLA）：把BCGA模型中的双向GRU替换为双向LSTM。

④ BERT_CNN_Self-Attention_Attention(BCSA)：把BCGA模型中的双向GRU替换为Self-Attention。

⑤ BERT_CNN_DBiGRU（BCG）：去除BCGA模型中的基于Attention的语义时序特征权重学习机制。

⑥ BERT：直接基于BERT预训练模型进行分类学习，考虑到BERT模型最大输入长度的限制，对事件信息分块输入。

实验结果如表4所示。

表4　不同组合模型的实验结果对比

Table 4　Experimental Results of Different Combination Models

组合模型	准确率/%	类别	查准率/%	查全率/%	F1/%
BERT	89.92	R	88.20	92.01	90.06
		N	91.78	87.87	89.78
DCGA	93.25	R	92.19	94.38	93.28
		N	94.34	92.13	93.22
BCG	93.35	R	91.86	95.03	93.42
		N	94.93	91.70	93.29
BCGA_s	93.57	R	92.07	95.25	93.63
		N	95.15	91.91	93.51
BCSA	94.00	R	92.31	95.90	94.07
		N	95.80	92.13	93.93
BCLA	94.86	R	92.96	96.98	94.93
		N	96.89	92.77	94.78
BCGA	95.39	R	93.75	97.19	95.44
		N	97.13	93.62	95.34

① 单纯基于BERT预训练模型的谣言检测准确率也高达89.92%，表明BERT预训练模型具有超强的语义表征能力，但模型没有考虑谣言事件传播过程中的语义时序特征，所以和其他6种模型相比准确率最低。

② BCGA模型的准确率比DCGA模型的准确率高2.14个百分点，表明基于BERT的文本语义特征抽取能力强于传统的Doc2Vec。

③ BCGA模型的准确率比BCG模型的准确率高2.04个百分点，表明基于Attention学习语义时序特征权重的有效性，Attention机制使模型更加关注语义时序数据中具有谣言特征的部分。

④ BCGA模型的准确率比BCGA_s模型的准确率高1.82个百分点，表明双向GRU的学习能力强于单向GRU，因为双向GRU使序列某点的输出同时结合过去和未来的信息，提升了语义特征的学习能力。

⑤ BCGA模型的准确率比BCLA模型的准确率高0.53个百分点，实验出现这种结果一定程度上是因为GRU相对LSTM具有更少的模型参数，在有限的训练样本下更容易收敛。

⑥ BCGA和BCLA两种模型的准确率都高于BCSA模型。实验出现这种结果一定程度上是因为Self-Attention丢失了位置信息，虽然位置编码部分弥补了这一缺点，但相比于严格保留输入前后关系的GRU和LSTM模型，BCSA模型的时序特征学习能力相对较弱。Self-Attention的优势在于并行计算和远距离依赖关系的计算，但本文对输入信息进行了分块，序列长度相对较短，传统的GRU和LSTM就可以处理。

⑦ BCGA模型的各项评价指标都高于其他6种组合模型，证明了该组合的有效性。

（4）和基准模型的实验结果对比

为了进一步验证本文模型的有效性，利用本文数据集对以下几种谣言检测基准模型进行复现。其中，DTC[2]、SVM-RBF[3]和SVM-TS[5]三种模型是基于人工特征工程的统计机器学习谣言检测的经典模型，DTC和SVM-RBF模型只利用了静态统计特征，SVM-TS模型引入了动态特征；TGBiA[13]、GRU-2[28]和CNN-GRU[29]三种模型是目前基于语义特征的深度学习谣言检测的标杆模型，TGBiA模型将谣言检测视为单文本分类问题，利用生成对抗网络强化谣言文本语义特征的学习，GRU-2和CNN-GRU模型融合转发评论信息进行谣言检测，但没有构建分层语义特征学习模型结构。

① DTC：基于决策树的分类模型。

② SVM-RBF：基于RBF核函数的SVM分类模型。

③ SVM-TS：基于动态时间序列结构的SVM分类模型。

④ TGBiA：基于生成对抗网络的分类模型。

⑤ GRU-2：基于双层GRU网络的分类模型。

⑥ CNN-GRU：基于CNN和GRU相结合的分类模型。

6种基准模型和文本BCGA模型的实验结果对比如表5所示。

表5　和基准模型的实验结果对比

Table 5　The Experimental Results Compared with the Benchmark Models

模型	准确率/%	类别	查准率/%	查全率/%	F1/%
DTC	82.96	R	84.70	80.13	82.35
		N	81.41	85.74	83.52
SVM-RBF	81.56	R	82.26	80.13	81.18
		N	80.91	82.98	81.93
SVM-TS	85.85	R	85.14	86.61	85.87
		N	86.58	85.11	85.84
TGBiA	91.21	R	89.28	93.52	91.35
		N	93.30	88.94	91.07
GRU-2	89.28	R	87.42	91.58	89.45
		N	91.29	87.02	89.11
CNN-GRU	90.68	R	89.17	92.44	90.77
		N	92.27	88.94	90.57
BCGA	95.39	R	93.75	97.19	95.44
		N	97.13	93.62	95.34

① TGBiA、GRU-2、CNN-GRU和BCGA这4种深度学习模型的准确率明显高于DTC、SVM-RBF和SVM-TS这3种传统统计机器学习模型。实验出现这种结果一定程度上是因为深度学习模型具有较强的语义特征学习能力，而基于人工特征工程的统计机器学习方法，人工特征缺乏全面性和灵活性，难以表示谣言的深层语义特征。

②在DTC、SVM-RBF、SVM-TS这3种统计机器学习模型中，SVM-TS模型的准确率最高，出现这种结果一定程度上是因为DTC和SVM-RBF模型忽视了特征随时间的变化，而SVM-TS模型利用了动态时间序列结构，捕获多种上下文特征随时间的变化。

③ 在TGBiA、GRU-2、CNN-GRU、BCGA这4种深度学习模型中，BCGA模型的准确率最高，比前面三种基准模型的准确率分别提高了4.18、6.11、4.71个百分点。实验出现这种结果一定程度上是因为BCGA模型相比于其他三种基准模型不仅考虑了文本语义特征，还融合了事件传播过程中的语义时序特征，通过分层模型实现了从微观层面到宏观层面的谣言语义判别，从而提高了模型的准确率。

（5）早期谣言检测效果对比分析

为了验证BCGA模型在早期谣言检测任务中的效果，选取原始微博发布后的9个时刻（1h、3h、6h、12h、24h、36h、48h、72h、96h）作为检测点。在每个检测点，输入模型的测试数据集为该时间范围内的相关事件信息。选择基于事件传播数据的SVM-TS、GRU-2、CNN-GRU作为基准模型，和本文模型对比在各检测点的准确率，并以微博社区管理中心的平均官方辟谣时间（Official Report Time）作为对照点，具体实验结果如图3所示。

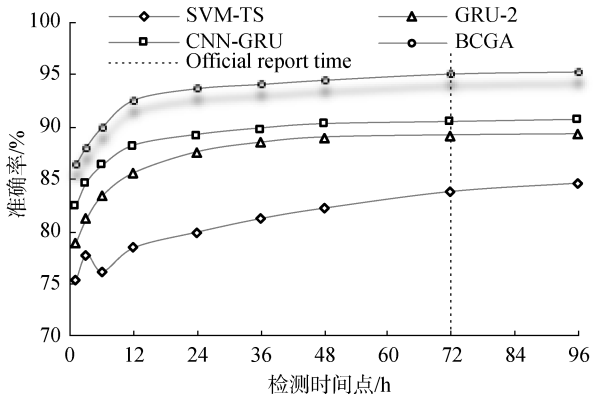

图3　早期谣言检测结果对比

Fig.3　Comparison of Early Rumor Detection Results

随着时间的推移，各模型的准确率总体上呈现不同梯度的上升，GRU-2、CNN-GRU和BCGA这3个模型相比于SVM-TS模型的准确率上升更加快速和稳定，到达36h后模型准确率基本稳定，而SVM-TS模型到达96h后才接近表5中的实验结果。BCGA模型在各检测时间点上的准确率都高于其他三个模型，说明本文模型依赖于更少的事件传播信息，在早期谣言检测任务上具有更好的效果。BCGA模型在事件发生12h后的准确率即达到92.5%，比SVM-TS、GRU-2、CNN-GRU模型分别高出14.0、6.9和4.3个百分点，并且比平均官方辟谣时间提早了约60h，由此证明了本文模型在早期谣言检测任务上的有效性。

5　结　语

社交媒体在方便人们信息分享的同时也导致了网络谣言的滋生和传播，谣言自动检测模型对遏制谣言的传播具有重要意义。目前基于语义特征的深度学习谣言检测方法重点关注博文内容，把谣言检测视为单文本分类问题，这种方法有时并不能有效识别谣言。为了提高基于语义特征的谣言检测方法的有效性，本文提出一种基于分层语义特征学习模型的微博谣言事件检测方法，在单文本语义特征的基础上融合转发和评论信息，依次采用 BERT、CNN、GRU、Attention 分别学习单文本语义特征、文本集合语义相关性特征、事件传播语义时序特征和语义时序特征权重。在公开数据集上的实验表明，本文模型的检测准确率达到 95.39%，检测时延在 12h 以内，相比于其他基准模型具有高的准确率和时效性。在未来研究中，将融合用户特征和社交网络关系，减少模型对传播数据的依赖，从而进一步提升模型在早期谣言检测任务中的效果。

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. (Gao Yujun, Liang Gang, Jiang Fangting, et al. Social Network Rumor Detection: A Survey[J]. Acta Electronica Sinica, 2020, 48(7): 1421-1435.)

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

[3] Yang F, Liu Y, Yu X H, et al. Automatic Detection of Rumor on Sina Weibo[C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. 2012: 1-7.

[4] 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. (He Gang, Lv Xueqiang, Li Zhuo, et al. Automatic Rumor Identification on Microblog[J]. Library and Information Service, 2013, 57(23): 114-120.)

[5] Ma J, Gao W, Wei Z Y, et al. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites [C]// Proceedings of the 24th ACM International Conference on Information and Knowledge Management. 2015: 1751-1754.

[6] 祖坤琳, 赵铭伟, 郭凯, 等. 新浪微博谣言检测研究[J]. 中文信息学报, 2017, 31(3): 198-204. (Zu Kunlin, Zhao Mingwei, Guo Kai, et al. Research on the Detection of Rumor on Sina Weibo[J]. Journal of Chinese Information Processing, 2017, 31(3): 198-204.)

[7] 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. (Ma Ming, Liu Yun, Liu Dijun, et al. Rumor Detection in Microblogs Based on Topic and Prevention Model[J]. Transactions of Beijing Institute of Technology, 2020, 40(3): 310-315.)

[8] Wu K, Yang S, Zhu K Q. False Rumors Detection on Sina Weibo by Propagation Structures[C]//Proceedings of 2015 IEEE 31st International Conference on Data Engineering. 2015: 651-662.

[9] 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 (Volume 1: Long Papers). 2017: 708-717.

[10] 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. (Zeng Ziming, 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.)

[11] Yu F, Liu Q, Wu S, et al. A Convolutional Approach for Misinformation Identification[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017: 3901-3907.

[12] Ajao O, Bhowmik D, Zargari S. Fake News Identification on Twitter with Hybrid CNN and RNN Models[C]//Proceedings of the 9th International Conference on Social Media and Society. 2018: 226-230.

[13] 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.(Li Ao, Dan Zhiping, Dong Fangmin, et al. An Improved Generative Adversarial Network for Rumor Detection[J]. Journal of Chinese Information Processing, 2020, 34(9): 78-88.)

[14] Ma J, Gao W, Wong K F. Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning[C]//Proceeding of the 2019 World Wide Web Conference. 2019: 3049-3055.

[15] 黄学坚, 王根生, 罗远胜, 等. 融合多元用户特征和内容特征的微博谣言实时检测模型[J]. 小型微型计算机系统, 2022, 38(12): 2518-2527. (Huang Xuejian, Wang Gensheng, Luo Yuansheng, et al. Weibo Rumors Real-time Detection Model Based on Fusion of Multi User Features and Content Features[J]. Journal of Chinese Computer Systems, 2022, 38(12): 2518-2527.)

[16] Tu K F, Chen C, Hou C Y, et al. Rumor2vec: A Rumor Detection Framework with Joint Text and Propagation Structure Representation Learning[J]. Information Sciences, 2021, 560: 137-151.

[17] Ke Z W, Li Z, Zhou C Z, et al. Rumor Detection on Social Media via Fused Semantic Information and a Propagation Heterogeneous Graph[J]. Symmetry, 2020,12(11): 1806.

[18] Ma T H, Zhou H H, Tian Y, et al. A Novel Rumor Detection Algorithm Based on Entity Recognition, Sentence Reconfiguration, and Ordinary Differential Equation Network[J]. Neurocomputing, 2021, 447: 224-234.

[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.(Yin Pengbo, Pan Weimin, 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] 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.(Xie Bailin, Jiang Shengyi, Zhou Yongmei, et al. Misinformation Detection Based on Gatekeepers' Behaviors in Microblog[J]. Chinese Journal of Computers, 2016, 39(4): 730-744.)

[21] 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.(Liu Zhiyuan, Zhang Le, Tu Cunchao, et al. Statistical and Semantic Analysis of Rumors in Chinese Social Media[J]. Scientia Sinica (Informationis), 2015, 45(12): 1536-1546.)

[22] Vosoughi S, Roy D, Aral S. The Spread of True and False News Online[J]. Science, 2018, 359(6380): 1146-1151.

[23] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[24] Liu F G, Zheng J Z, Zheng L L, et al. Combining Attention Based Bidirectional Gated Recurrent Neural Network and Two Dimensional Convolutional Neural Network for Document-Level Sentiment Classification[J]. Neurocomputing, 2020, 371: 39-50.

[25] Dey R, Salem F M. Gate-variants of Gated Recurrent Unit (GRU) Neural Networks[C]//Proceedings of 2017 IEEE 60th International Midwest Symposium on Circuits and Systems. 2017: 1597-1600.

[26] Parikh A, Täckström O, Das D, et al. A Decomposable Attention Model for Natural Language Inference[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2249-2255.

[27] Ho Y, Wookey S. The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling[J]. IEEE Access, 2019, 8: 4806-4813.

[28] Ma J, Gao W, Mitra P, et al. Detecting Rumors from Microblogs with Recurrent Neural Networks[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016: 3818-3824.

[29] 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.(Li Lizhao, Cai Guoyong, 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.)

作者贡献声明:

黄学坚:提出研究思路,进行实验,起草论文;
马廷淮:对论文提出修订意见,论文最终版本修订;
王根生:设计研究方案。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

[1]黄学坚. Weibo谣言事件检测数据集. https://www.scidb.cn/s/bA3Qny.

收稿日期:2022-06-14
收修改稿日期:2022-07-18

Detecting Weibo Rumors Based on Hierarchical Semantic Feature Learning Model

Huang Xuejian[1,2]　Ma Tinghuai[1]　Wang Gensheng[3]

[1](College of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

[2](VR College of Modern Industry, Jiangxi University of Finance and Economics, Nanchang 330013, China)

[3](College of Humanities, Jiangxi University of Finance and Economics, Nanchang 330013, China)

Abstract: [**Objective**] This paper tries to improve the accuracy and timeliness of Weibo rumor detection. [**Methods**] We proposed a rumor detection method based on the hierarchical semantic feature learning model (BCGA). Firstly, we extracted the semantic features of a single text in an event based on the BERT model. Secondly, we dynamically grouped the event propagation data based on the time domain. Next, we used the convolutional neural network to learn the semantic correlation features of the text sets in each time domain. Fourth, we input the semantic correlation features in each time domain into the deep bidirectional gated recurrent neural network to learn the deep semantic temporal features of the event propagation process. Finally, we integrated the attention mechanism to make the model focus on the rumor feature in semantic temporal features. [**Results**] Experiments on the Weibo public data sets show that the detection accuracy of the model reached 95.39%, while the detection delay was within 12 hours. [**Limitations**] The model requires a certain amount of forwarding and commenting information and the detection effect is not prominent when the event is not popular enough. [**Conclusions**] The hierarchical semantic feature learning model achieves a learning process from local to global semantics, improving the performance of Weibo rumor detection.

Keywords: Rumor Detection　Deep Learning　Semantic Features　Temporal Data　Hierarchical Semantic