

引用格式:尹鹏博,彭成,潘伟民. 基于集成学习的微博谣言早期检测[J]. 微电子学与计算机, 2021, 38(1): 83-88. [YIN P B, PENG C, PAN W M. Early detection of microblog rumors based on ensemble learning[J]. Microelectronics & Computer, 2021, 38(1): 83-88.]

基于集成学习的微博谣言早期检测

尹鹏博, 彭 成, 潘伟民

(新疆师范大学 计算机科学技术学院, 新疆 乌鲁木齐 830054)

摘 要: 微博谣言早期检测对于谣言防治有重要作用,而在谣言发生的早期缺乏相关信息,检测难度大. 该文通过构建检测特征和组合多种检测算法实现微博谣言的早期检测. 在检测特征选取方面,不直接使用微博的评论转发信息,而是通过对待检测微博文本和用户历史微博进行情感分析,构建刻画出用户和微博的情感特征. 在检测算法方面,采用集成学习方法作为谣言检测算法,算法的基模型由多个异构深度学习模型组成,元模型采取随机森林算法,以元模型在基模型的预测输出上进行二次训练的方式组合不同模型以提高检测准确率. 实验表明,该方法在谣言早期检测方面具有较好的检测效果.

关键词: 微博谣言; 早期检测; 集成学习; 深度学习

中图分类号: TP391.1

文献标识码: A

文章编号: 1000—7180(2021)01—0083—06

Early detection of microblog rumors based on ensemble learning

YIN Peng-bo, PENG Cheng, PAN Wei-min

(School of computer science and technology, XinJiang Normal University, Urumqi 830054, China)

Abstract: The early detection of microblog rumors plays an important role in the prevention and control of rumors. However, it is difficult to detect rumors due to the lack of relevant information in the early stage of rumor occurrence. In this paper, the early detection of microblog rumors is realized by selecting effective detection characteristics and combining multiple detection algorithms. In the term of selecting detection characteristics, this paper constructs the emotional characteristics of users and microblog through the emotional analysis of microblog text and user's historical text instead of using the information extracted from comments and forwards. In the term of detection algorithm, the ensemble learning method is adopted as the rumor detection algorithm. The base model is composed of multiple heterogeneous deep learning models. The random forest algorithm is used in the meta model to combine different models in the way of secondary training on the prediction output of the base model to improve the detection accuracy. Experiments show that this method has a good detection effect in the early detection of rumors.

Key words: microblog rumors; early detection; ensemble learning; deep learning

1 引言

谣言是广泛传播的、未经证实的陈述,其真实价值未经证实或故意为虚假的信息. 中国社会科学院发布的《新媒体蓝皮书》显示,中国有 59% 的网络谣言来自新浪微博. 对此,微博于 2010 年推出其官方

辟谣平台,通过人工查证识别谣言. 根据新浪微博社区管理中心网站数据显示,截止 2020 年 1 月 1 日,共辟谣 38186 条,换句话说,9 年来,平均每天辟谣 11 条,人工检测的效率可见一斑,因此,微博谣言自动检测研究逐渐成为研究热点.

本文在文献[1-12]的基础上提出基于集成学习

的微博谣言早期检测方法,该方法分为两个方面:一是通过用户历史行为分析,构建有效的检测特征;二是提出以多个异构深度学习模型为基模型的集成学习检测模型,该模型使用随机森林算法以在基模型的输出上进行二次训练的方式组合基模型(Random Forest stacking based on deep learning, RFS-BD).

2 特征分析

特征分析的目的是为了选取出最有效的检测特征以提高检测效果. 谣言早期检测要求尽早进行检测,虽然微博的转发、评论等信息是非常有效的检测特征,但是由于这些信息是在微博发布后随时间流逝而逐渐增长的,具有很大的时延和不确定性,不适用于早期检测任务. 故本文选择通过分析用户历史行为并结合微博文本进行微博谣言的早期检测.

2.1 微博特征分析

很多谣言微博在语言结构、逻辑表达等方面与正常微博有一定的差异,例如谣言微博通常含有较多的主观推断成分、事实依据含糊不清、含有较多的口语化表述,这也就意味着微博文本在一定程度上就表达了该条微博是否是谣言微博,因此可以通过博文分析进行谣言检测.

通过对大量谣言微博的分析发现,大多数谣言微博的话题或者观点具有较强的情绪. 如表 1 所示,

表 1 谣言微博的分类

类别	示例	情感值
揭秘爆料性谣言	【赶紧防范! 别出门!】河南省商丘市市政府医院,昨天凌晨 2:30,13 名男女生感染 H7N9 病毒死亡,最大的 32 岁,最小的 5 岁	0.989
求助性谣言	捡到一张准考证,刘明婷,考点在一中,请朋友们转发,让刘明婷联系这个号码 15375268418,一定帮他群发一下,这孩子一家肯定急死了,扩散,扩散,别耽误孩子高考!	0.904
伪科普性谣言	【为了家人和朋友,转发】如果你被匪徒挟持要求输入取款机密码,你可以用倒转输入密码的方式去间接知会警方. 例如你的密码是 1234 的话,你可以输入 4321,取款机会识别到你是以倒转方式输入密码,取款机会按你要求吐出金额,但是会在匪徒不知情的情况下通知警方.	0.06
事实离奇性谣言	【难过】香港检测出韩国农心辛拉面面向中国内地销售的产品不合格!,塑化剂超标 50 倍,接近极限. -居然毒害了咱那么多年阿[泪]	0.192

由于带有争议性的微博的情感值通常较大或较小,因此可以选取用户的历史微博进行情感分析,记情感值大于或小于某阈值的微博为引战微博,引战微博的总数称为用户的引战言论数,将其作为用户的历史行为特征. 由于历史微博会拥有一定数目的评论,而对微博评论进行情感分析比直接对微博文本进行情感分析更为准确,因此,本文选择通过计算微博评论的情感进而判断微博文本的情感. 首先,计算微博评论的情感,将

展示了谣言微博的几种类别及其示例. 这些谣言微博常常具有极性情感:正面程度很高或者很低,通常这是由谣言微博具有争议性观点所致. 因为带争议性的观点能激发人们的情绪,引起较激烈的反响^[11]. 因此,本文通过计算待检测微博的情感引入一个新的特征:情感极性(Q). 情感极性由情感计算而来,在常规情感计算中,情感值的范围为[0,1],较小和较大的值分别表示消极和积极,中间的值表示中性. 使用公式(1)将情感值量化为情感极性,采用抛物线公式的目的是将不连续的情感表示转化为以情感激烈程度量化的连续情感值.

$$Q = 4x^2 - 4x + 1$$

(1)

式中, $x \in [0,1]$, $Q \in [0,1]$,且 Q 越大情感越激烈.

2.2 用户特征分析

用户特征对于谣言早期检测具有重要作用. 分析发现,很多发布谣言的用户不止发布过一条谣言,且这些用户的日常热度较高,保持热度的有效做法之一是发布具有争议性的微博,因为带争议性的观点通常能激发人们的情绪,引起较激烈的反响^[11],谣言微博也被归于此类. 并且,这些用户的非谣言微博中,带有争议性观点的微博比例也要高于普通用户,这就说明,用户发布谣言并不是偶然的,并且与之前的行为是有所关联的,因此用户的历史行为特征就为谣言检测提供了很强的参考性.

大于一定阈值的记为积极评论(pos),小于一定阈值的记为消极评论(neg),且 neg、pos 均不为 0,中性评论忽略不考虑. 然后通过 pos、neg 的数量计算微博的情感极性(C). 当 pos=neg 时,认为评论中的争议最大,而当 pos→0 或 neg→0 时,认为争议性最小,因此,可以用式(2)计算表示微博的情感极性:

$$C = \frac{\text{pos}}{\text{neg}} + \frac{\text{neg}}{\text{pos}} = \frac{\text{pos}^2 + \text{neg}^2}{\text{pos} * \text{neg}}$$

(2)

但是在公式(2)中,对 pos 和 neg 的数量级不敏感,同时 C 的取值范围不在一个确定的区间,不直观,因此为其添加评论数权重和归一化函数得到公式(3).

$$C = f\left(\frac{\text{pos}^2 + \text{neg}^2}{\text{pos} * \text{neg}} * f(|\text{neg} + \text{pos} - a|)\right) \quad (3)$$

公式(3)中, a 是偏置系数,用于过滤有效评论过少的数据.当 $C = 0$ 时,微博为中性,当 $C = 1$ 时,微博情感极性最强.当 C 大于一定阈值时,可判定该条微博为引战微博.

3 算法设计

本文的 RFS-BD 算法使用多个异构的深度学习模型作为基模型,基模型使用原始训练集做训练.选取随机森林算法作为元模型,在基模型的预测输出上进行二次训练,元模型的输出为最终的结果,这种集成方式被称为 Stacking,已有的研究已经从理论上证明了 Stacking 的优越性^[13].如何选择“好而不同”的基模型,是集成学习的核心.在相关研究中,基模型一般选取的是决策树、随机森林等树形算法,而本文则选取了深度学习算法,如 CNN, RNN,使得检测效果得到了很大的提升.

3.1 基本定义

一条微博样本数据包含多个数据特征和一个代表其是否为谣言微博的标签.定义由特征和标签组成的原始数据集(D_1)如下:

$$D_1 = \{x_{ij}, y_{ij}\}$$

x_i^j ($i = 1, 2, 3 \dots m; j = 1, 2, 3 \dots n$) 表示第 i 个样本数据的第 j 个特征; y_i^j 表示第 i 个样本的第 j 个特征的标签, $y_i^j \in \{0, 1\}$. 0 和 1 分别代表是谣言和不是谣言.

算法包含若干个基分类器 h_i ($i = 1, 2, 3 \dots T$), 每个基分类器接收样本数据 D_1 的某些特征,输出为二次训练样本 D_2 ,将其定义为:

$$D_2 = \{o_i, y_i\}, \text{ 其中 } o_i = \{h_1(x_i), h_2(x_i) \dots h_T(x_i)\}, h_T(x_i) \text{ 是第 } T \text{ 个基模型的预测输出.}$$

算法 1 是检测模型工作流程的伪代码描述.

算法 1 检测算法工作流程的步骤描述

输入:初始训练集 D_1

输出:集成分类模型 H

1: Step1: 训练基模型

2: for $t = 1$ to T do

3: 在 D_1 上训练基模型 h_t

4: end for

5: Step2: 生成二次训练集 D_2

6: for $i = 0$ to m do

7: $D_2 = \{o_i, y_i\}$, 其中 $o_i = \{h_1(x_i), h_2(x_i) \dots h_T(x_i)\}$

8: end for

9: Step3: 二次训练

10: 用 D_2 训练元分类器 H

11: return H

3.2 基分类器选择

在集成学习中,通常基分类器会选择随机森林、极限树、梯度提升树等树形算.其特点是解释性强且结果具备随机性,能较好地满足集成学习“好而不同”的要求.但是由于这些基分类算法的单个准确率并不是很高,因此限制了集成模型的整体准确率的提升.研究表明,深度神经网络在自然语言处理方面效果要显著好于传统的机器学习算法,单个神经网络模型相较于单个的树形算法效果会更好,实验对比如图 1 所示.因此本文选择采取神经网络模型作为基模型.

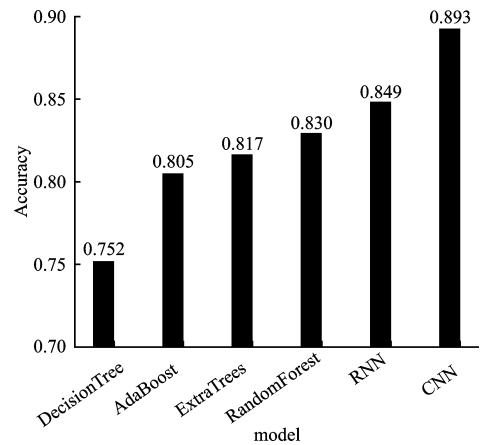


图 1 树形算法和深度学习算法准确率对比

不同模型对于不同特征的敏感度不同.每个基模型都接受训练样本的全部特征 x_{ij} ($i = 1, 2, 3 \dots m; j = 1, 2, 3 \dots n$), 还是只接受部分特征 x_{ij} ($i = 1, 4, 5 \dots m; j = 2, 3, 6 \dots n$) 对于检测效果会产生不同的影响.图 2 展示了 CNN 和 RNN 在不同特征下的准确率.通过结果可以看出,相同的特征作用于 CNN 和 RNN 上有不一样的效果, CNN 对于用户信息等离散数据更为敏感,而 RNN 则对于连续的微博文本信息更敏感,因此,按照不同的基模型使用不同的样本特征会取得不同的效果.

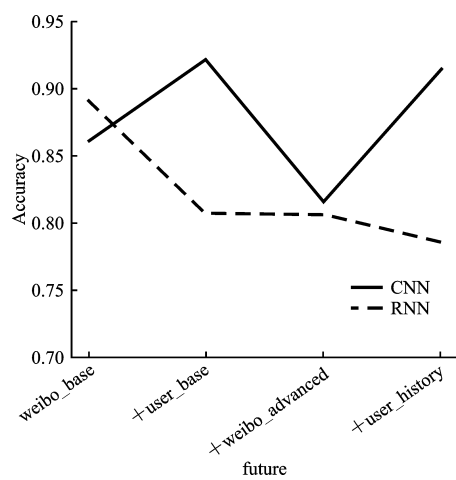


图 2 不同特征对于检测准确率的影响

3.3 分类器集成

常见集成学习例如 Bagging, 采取的是基模型投票法, 即每个基模型都对样本进行预测其分类, 以得票最多的分类为样本的最终分类. 该方法能减小整体预测的方差, 但是对于预测效果并无太大的改进. 本文选取随机森林作为元模型, 合并基模型的输出为二次训练集, 在元模型上进行二次训练, 最终的预测结果由元模型提供. 算法示意如图 3 所示.

为了充分利用数据, 采取 K-折交叉验证法构建二次训练集. 首先, 将初始训练样本分成训练集和测试集, 通过 K-折交叉验证在训练集上训练基模型, 使用测试集进行模型最终的测试. 对于每一轮训练,

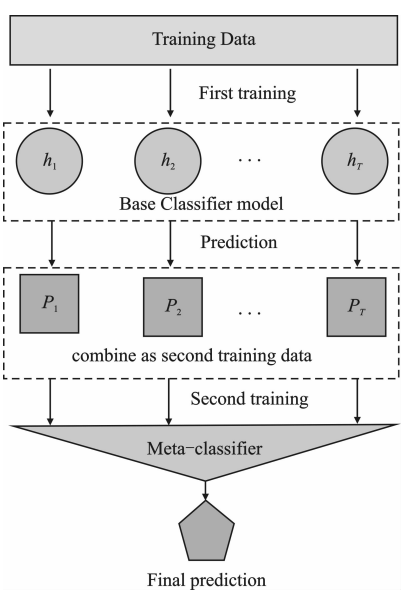


图 3 算法示意图

都有 $K-1$ 份数据被用于训练, 1 份数据用于测试, 当 K 轮训练之后, 每个模型对于所有训练样本都有了其预测结果; 然后将所有基模型的预测结果合并就得到新的二次训练样本, 二次训练样本的标签就是初始样本数据的标签, 这样通过 K -折交叉验证就可以在防止过拟合的情况下得到所有样本的二次训练数据. 二次训练数据集的构建过程如图 4 所示, 最后将二次训练集划分为训练集和测试集在元模型上进行训练与测试.

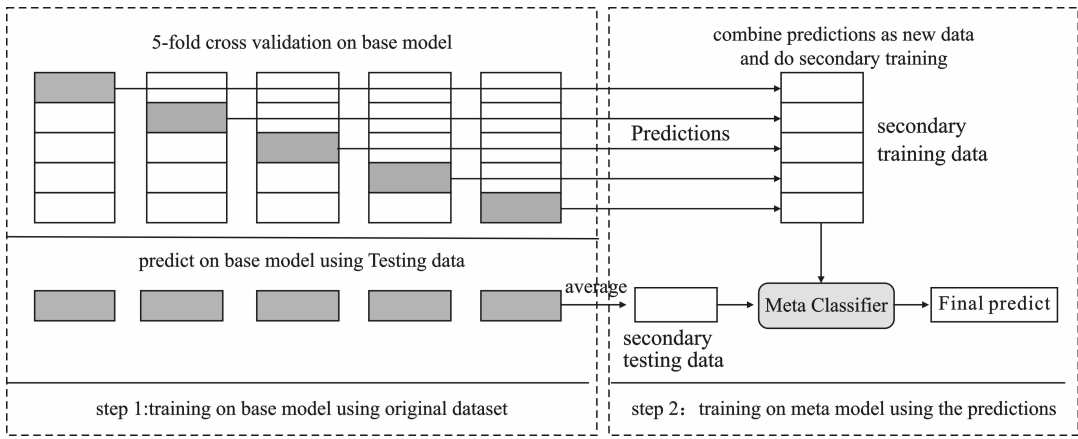


图 4 通过 5-折交叉验证构建在基模型上构建二次训练流程图

4 实验

4.1 数据集和检测特征选取

首先选取公开的微博谣言数据集, 它已被使用在多个相关研究中^[1, 5, 6]. 数据集中包含 4 664 条微

博和 2 309 名用户, 其中有谣言微博 2 313 条, 非谣言微博 2 351 条. 由于该数据集中, 用户信息不是最新的数据, 且不包含用户的历史微博, 因此本文设计了网络爬虫, 在新浪微博上爬取这 2 309 名用户的最新个人信息和他们的历史微博. 除去被微博封号

的无效用户,实际共爬取了 2 145 名用户的最新个人信息与他们的历史微博共 135 991 条以及这些微博的评论信息. 最终有 2 183 条谣言微博和 2 012 条非谣言微博参与训练.

微博文本不仅表达了语义信息,还记录了发布者的语言习惯,情感偏向等信息. 研究表明,谣言微博通常都是在非移动设备上发布,在通过对谣言微博的综合后选择出有效的检测特征. 表 2 展示了微博特征和用户特征的选取,其中引战言论数通过对用户历史微博进行情感分析得来.

表 2 检测特征的选取

微博特征	用户特征
微博文本、微博文本的情感极性、微博的发布设备	用户昵称、性别、所属地区、引战言论数、个人描述、关注数、粉丝数、发布微博数

4.2 实验设置

用户特征分析中,通过实验表明取 $C=0.65$ 作为阈值效果最好. 当 $C>0.65$ 时,认为该条微博的情感较为激烈,记为引战微博,将所有引战微博数记为用户引战言论数.

基模型的选取情况如表 3 所示. 选取随机森林作为元模型组合基模型形成最终的分类模型. 元模型使用 100 个决策树,决策树节点划分标准选用信息熵,允许进行 Bootstrap 抽样.

表 3 基模型的选取与说明

模型名	说明
CNN_v1	单层卷积神经网络
CNN_v2	双层卷积神经网络
LSTM	长短期记忆循环神经网络
GRU	门控循环神经网络
CNN-LSTM	串联 CNN 和 LSTM 的组合模型
CNN-GRU	串联 CNN 和 GRU 的组合模型

实验选取 Snownlp 作为情感分析工具,pukseg 作为中文文本分词工具,模型训练采取 5-折交叉验证法.

4.3 实验结果

表 4 展示了基模型和集成模型的准确率对比. 对比准确率可以看出,通过对不同模型的组合,可以提高检测的准确率,使集成模型的效果优于单个基模型.

为了实验具有可比性,本文实现了刘政等人^[8]的 CNN 模型,Chen 等人^[14]的 RNN 模型,李力钊等人^[7]的 C_GRU 模型,Xiong X 等^[10]的 GTB-RD 集

成学习模型,以及基模型为树模型的集成学习模型 (Linear stacking based on tree algorithm,LSBT),将这些模型应用在本文所构建的数据集上,分别比较了 Precision, Accuracy, Recall, F1-score. 表 5 展示了本文的 RFS-BD 模型与其他模型的效果对比.

表 4 基模型与集成模型的准确率

模型名	准确率
CNN_v1	0.905
CNN_v2	0.919
LSTM	0.896
GRU	0.899
CNN-LSTM	0.869
CNN-GRU	0.871
RFS-BD	0.929

表 5 实验结果对比

Model	Precision	Accuracy	Recall	F1-score
RNN	0.878	0.887	0.89	0.884
GTB-RD	0.896	0.896	0.88	0.889
LSBT	0.897	0.897	0.886	0.892
C_GRU	0.909	0.902	0.885	0.897
CNN	0.923	0.915	0.901	0.912
RFS-BD	0.927	0.929	0.923	0.925

通过实验结果可以看出,本文所提出的方法具有较好的效果. 从算法的角度看有三个特点:(1)使用集成模型的检测效果比使用单个深度学习模型好;(2)基模型采取深度学习模型的检测效果要优于树形模型;(3)采用二次训练的集成方法要优于梯度提升的集成方法. 而从特征选取的角度看,用户特征对于谣言早期检测效果的提高有重要作用.

虽然使用集成模型的效果要好于单个的基模型,但是在开销上也要大于单个基模型. 主要是性能的开销和时间的开销增大,其中以时间开销大作为显著特点. 以单层 CNN 为例,隐藏层大小设置为 256,批大小为 128,学习率 0.001,优化器为 adam,在当前数据集上使用内存为 8G、I5 处理器的机器进行训练时,其模型收敛时间在 8 分钟左右,而当采用 6 个这样的 CNN 模型为基模型,随机森林为元模型,以 5-折交叉验证进行训练时,收敛时间就提升到 240 分钟左右. 因此需要谨慎的做出权衡,一是要选择少而性能优异的基模型,二是要对于检测特征进行抽样以降低数据规模,这样可以在保证取得良好性能的基础上尽可能降低时间开销. 同时,值得注意的是,使用集成的方式提升模型性能更适合的

是小样本学习任务,在谣言检测任务中,避免使用大规模数据集进行训练,反之使用小样本,渐进式训练,逐步优化模型性能是非常好的发展方向。

5 结束语

基于用户特征进行微博谣言早期检测可以尽早发现谣言;因为不使用传播特征作为检测依据可以在谣言产生之初就进行检测,使检测周期不再依赖不确定的传播因素,取而代之的是特征的分析提取;使用集成学习方法结合多种检测算法能更好的综合不同算法的优点,提高检测准确率。不足之处是使用集成学习方法会提高检测算法的时间复杂度,因此在时间复杂度和检测效果之间需要一个平衡点,在今后的工作中,将进一步研究更高效的算法和挖掘更有效的用户特征,目标是检测更早、检测更准。

参考文献:

- [1] YANG F, LIU Y, YU X H, et al. Automatic detection of rumor on Sina Weibo[C]//Proceedings of ACM SIGKDD Workshop on Mining Data Semantics. Beijing, China: ACM, 2012: 1-7. DOI: 10.1145/2350190.2350203.
- [2] VOSOUGHI S, MOHSENVAND M N, ROY D. Rumor Gauge: predicting the veracity of rumors on twitter[J]. ACM Transactions on Knowledge Discovery from Data, 2017, 11 (4): 1-36. DOI: 10.1145/3070644.
- [3] ZHANG Z L, ZHANG Z Q, LI H Y. Predictors of the authenticity of Internet health rumours[J]. Health Information and Libraries Journal, 2015, 32 (3): 195-205.
- [4] GUPTA M, ZHAO P X, HAN J W. Evaluating event credibility on twitter[C]//Proceedings of 2012 SIAM International Conference on Data Mining. Anaheim: SIAM, 2012: 153-164. DOI: 10.1137/1.9781611972825.14.
- [5] ZHAO Z, RESNICK P, MEI Q Z. Enquiring minds: early detection of rumors in social media from enquiry posts[C]//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy: ACM, 2015: 1395-1405. DOI: 10.1145/2736277.2741637.
- [6] 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 on Conference on Information and Knowledge Management. Melbourne, Australia: ACM,

- 2015: 1751-1754. DOI: 10.1145/2806416.2806607.
- [7] 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. DOI: 10.6040/j.issn.1672-3961.0.2018.189.
- LI L Z, CAI G Y, PAN J. A microblog rumor events detection method based on C-GRU [J]. Journal of Shandong University (Engineering Science), 2019, 49 (2): 102-106. DOI: 10.6040/j.issn.1672-3961.0.2018.189.
- [8] 刘政, 卫志华, 张韧弦. 基于卷积神经网络的谣言检测[J]. 计算机应用, 2017, 37(11): 3053-3056.
- LIU Z, WEI Z H, ZHANG R X. Rumor detection based on convolutional neural network[J]. Journal of Computer Applications, 2017, 37(11): 3053-3056.
- [9] ROY A, BASAK K, EKBAL A, et al. A deep ensemble framework for fake news detection and classification[J]. arXiv: 1811.04670, 2018.
- [10] XIONG X, YANG B, KANG Z F. A gradient tree boosting based approach to rumor detecting on sina weibo[J]. arXiv: 1806.06326, 2018.
- [11] LIN D Z, MA B, CAO D L, et al. Chinese microblog rumor detection based on deep sequence context[J]. Concurrency and Computation: Practice and Experience, 2019, 31 (23): e4508. DOI: 10.1002/cpe.4508.
- [12] KWON S, CHA M, JUNG K. Rumor detection over varying time windows[J]. PLoS One, 2017, 12(1): e0168344. DOI: 10.1371/journal.pone.0168344.
- [13] SILL J, TAKACS G, MACKAY L, et al. Feature-weighted linear stacking[J]. arXiv: 0911.0460, 2009.
- [14] NGUYEN T N, LI C, NIEDERE C. On early-stage debunking rumors on twitter: Leveraging the wisdom of weak learners[C]//Proceedings of the 9th International Conference on Social Informatics. Oxford: Springer, 2017: 141-158. DOI: 10.1007/978-3-319-67256-4_13.

作者简介:

尹鹏博 男, (1996-), 硕士研究生. 研究方向为自然语言处理、数据挖掘(网络舆情方向)。

彭成(通讯作者) 男, (1971-), 博士, 副教授, 硕士生导师. 研究方向为教育信息化、娱教技术。

E-mail: pcxjnu@163.com.

潘伟民 男, (1963-), 硕士, 教授, 硕士生导师. 研究方向为计算机应用技术、网络信息安全。