基于主题和预防模型的微博谣言检测

马鸣[1,3]， 刘云[1,2]， 刘地军[3]， 白健[3]
(1. 北京交通大学 电子信息工程学院,北京 100044; 2. 北京交通大学 北京市教委通信与信息系统重点实验室,北京 100044;3. 保密通信重点实验室,四川,成都 610041)

摘 要: 针对微博短文本存在的特征提取困难及微博谣言传播浪费网络资源的问题,提出了基于主题和预防模型的微博谣言检测. 对微博进行主题提取,按主题分类后提取基于用户、传播结构、内容三方面的统计特征. 将样本与官方谣言子集中的微博进行相似度计算,将其值与传统特征进行特征融合之后作为统计特征进入有监督的机器学习. 实验结果表明,相对于传统的有监督机器学习,该方法将微博谣言检测的性能提升了 3%左右,同时实现了谣言预防.

关键词: 微博; 谣言检测; 主题发现; 谣言预防
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 1001-0645(2020)03-0310-06
DOI: 10.15918/j.tbit1001-0645.2018.384

Rumor Detection in Microblogs Based on Topic and Prevention Model

MA Ming[1,3], LIU Yun[1,2], LIU Di-jun[3], BAI Jian[3]
(1. College of Electronic Information Engineering, Beijing Jiaotong University, Beijing 100044, China; 2. Beijing Municipal Education Commission Key Laboratory of Communication and Information System, Beijing Jiaotong University, Beijing 100044, China; 3. Confidential Communication Laboratory, Chengdu, Sichuan 610041, China)

Abstract: In view of the difficulty in extracting features of Weibo's short texts and the existence of a large number of officially certified microblogs that had not been used efficiently, a microblog rumors detection method was proposed based on topics and prevention model. Firstly, the official rumors were extracted and categorized according to the subject and were organized according to the user, spread frame and content characters, forming the subsets of official rumors based on a certain topic. And then, the similarity between the microblog and the official rumors with an identical topic was calculated. Merging the values with the traditional features, the result was taken as statistical features put into supervised machine learning. Finally, some experiments were carried out to validate the detection method. The results show that, compared with the traditional supervised machine learning, the method can improve the performance of Weibo rumors detection by about 3%, and can achieve rumor prevention.

Key words: microblog; rumor detection; topic discovery; rumor prevention

根据中国互联网信息中心 2018 年 1 月发布的中国互联网发展统计报告[1],截至 2017 年 12 月,我国网民规模达 7.72 亿,普及率达到 55.8%. 在 2018 年第一季度,微博月活跃用户已达 4.11 亿. 面对如此多用户、大流量的平台现状,许多不法分子在微博平台上编造并传播了大量难于验证的谣言信息,给用户体验、平台发展和国家稳定带来巨大危害. 谣言检测已经成为国内乃至国际研究的重点及方向[2].

收稿日期: 2018-10-10
基金项目: 国家重点研发计划(2016YFC0801004);中央高校基本科研业务费专项资金课题(2017JBZ017)
作者简介: 马鸣(1993—),女,硕士,E-mail:16120107@bjtu.edu.cn;刘云(1955—),女,博士,教授.

传统的谣言检测主要从机器学习算法的选择和创新、区分谣言和非谣言的特征提取两方面做出贡献. Yang 等[3]在传统特征的基础上提出了用户发布谣言使用的客户端类型和是否包含 URL 两个新特征,使用 J48 算法来构建谣言检测模型. Ma 等[4]在谣言传播结构的基础上提出了基于核函数的传播树检测模型. Wu 等[5]在文献[4]的基础上,提出了基于图-核函数的传播树检测模型. Jin 等[6]则从传播结构中的另一方面入手,采集该微博全部转发和评论,对转发和评论内容分析,提出基于主题下的情感-评论对,通过计算该微博下所有回复的情感-评论对的相似度,确定该微博是否为谣言.

然而,之前的工作都忽略了微博官方辟谣账户的工作强度,仅将其作为确定谣言的数据源. 实际上,当前微博上的多数谣言为了引人注目而采用的噱头和过去一个月、两个月甚至一年内的谣言基本一样. 图 1 显示了当前的 3 个谣言微博内容,可以发现其描述的虽然是同一个事件,但是经过不同用户的传播,对谣言内容进行了夸张夸大的修饰,包括但不仅限于改变城市地址、改变医院名称、改变病毒名称等行为(如红色边框标记单词所示). 这种传播模式不仅加大了谣言的争议性范围,还加大了谣言检测难度. 单用传统的谣言特征,如微博用户是否认证,微博用户的关注数和粉丝数,谣言微博的转发评论数等,只增加了检测工作强度,并不能得到较高的谣言检测正确率.

图 1　谣言内容比较图

Fig. 1　Comparison of rumors

Ma 等[7]首次使用循环神经网络(necurrent neural networks,RNN)对特征进行深度学习. 考虑到 RNN 的特性,Ma 等提出将事件的数据流根据时间算法进行切片,再将切片后的数据通过循环单元和隐藏层进行学习,并将学习结果与传统的谣言分类结果比较,发现使用神经网络可大大提高谣言检测准确性. 杨文太等[8]提出基于突发话题和领域专家的谣言检测方法,该方法借鉴物理学中的动力学理论对话题特征进行建模. Jin 等[9]提出多模型的谣言检测算法. 针对图片和视频的谣言属性,实现了图片的特征向量和文本的特征向量的融合算法.

本文提出的基于主题和谣言预防的谣言检测模型,则很好地利用了谣言数据库. 将已存储的谣言集主题提取并分类,对新样本确定主题、构建特征向量并与数据库中该主题下得到的谣言子集特征平均值进行距离计算,将计算结果作为统计特征放入样本的特征向量中,最后经过训练器二类分类得到分类结果. 实验表明,该模型大大提高了谣言检测的准确性,相较于传统的谣言检测,该模型可有效地在初期发现谣言,实现谣言预防.

1　基于主题和预防的谣言检测模型

1.1　基本定义

下面将对谣言检测中的相关问题进行定义并阐释其含义.

谣言:错误的信息或者是故意捏造的信息. 比如,在 2017 年“八宝粥”事件中,一些用户为了出

名，故意捏造一些可能造成社会恐慌的微博信息，如“好可怕，八宝粥都是用海绵做的”，这类信息不仅严重伤害了用户的感情，还对社会产生了负面影响.

消息：文本信息，与之对应的单词为事件，事件是由多个消息构成. 以“5.6滴滴司机杀害空姐事件”为例，该事件的主要内容是“5月6日，一空姐在河南郑州航空港区搭乘了一辆滴滴顺风车赶往市内后遇害，警方称，作案人员是一名滴滴司机，凶手仍在潜逃”. 该事件微博下的转发评论信息为相关消息. 如网友评论“这次空姐滴滴打车遇害事件，凶手仍在潜逃，滴滴道歉，太可怕了…”则为消息内容.

谣言子集：该主题下相关的谣言数据集.

1.2 模型结构

定义谣言子集如式(1)所示.

$$\begin{cases} m_{ij} \in \text{topic}_i \\ i \in [1,n], \quad j \in [1,k] \\ \boldsymbol{v}_{\text{topic}} = F_{\text{topic}_i}(f_1, f_2, \cdots, f_j) \end{cases}. \tag{1}$$

式中：topic_i 表示谣言数据库中存在的第 i 个主题，谣言数据库中共包含 n 个主题；m_{ij} 表示 topic_i 下第 j 个消息，topic_i 下共包含 k 个消息；$\boldsymbol{v}_{\text{topic}}$ 表示主题 i 下消息的特征向量；f_j 表示第 j 个特征值.

基于主题和预防的谣言检测模型包含4个单元，如图2所示.

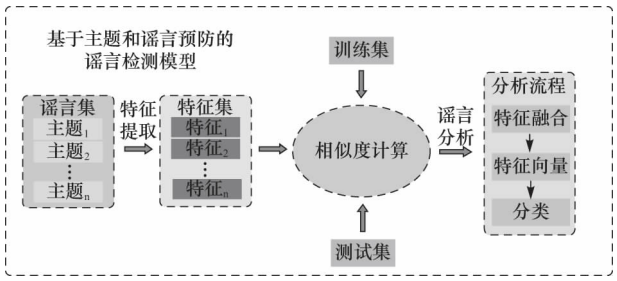

图2 基于主题和谣言预防的谣言检测模型

Fig. 2 Rumor detection model based on subject and rumor prevention

第一部分为对已存储的谣言集主题提取并按主题确定谣言子集；第二部分对谣言子集进行特征提取，得到特征向量的平均值；第三部分为相似度计算，将训练集/测试集得到的数据特征提取，并与第二部分得到的谣言子集特征向量的平均值进行距离计算，得到距离值；第四部分为分类流程，将第三部分得到的距离值作为统计特征放入特征向量中，再经过分类器分类.

1.2.1 谣言子集主题提取

考虑到旧谣新传时关键词的稳定性，同时考虑到LDA主题生成模型会将相关性较小的旧谣言归于同一主题下，本文采用消息的TF-IDF值对谣言数据库进行主题分类. TF-IDF是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度. 选用TF-IDF值作为消息的主题表达是谣言检测中常用的表达方式[10,12]. 对谣言数据库中的每条消息得到TF-IDF值，选择TF-IDF值最大的前3个单词作为主题表达.

1.2.2 特征选择

微博的谣言特征经过多年研究，共分为3类：基于消息的特征、基于传播结构的特征和基于用户的特征. 为了更好地检验本文提出的消息相似度特征，本文使用计算得出的 x^2 值与单个特征 F_1 值排前7的特征构成特征向量. 这7个特征分别为VIPRank、Repost _rate、Reputation、Is _specific、Is _Url、Is _Verify和Participation. VIPRank表示用户VIP等级. Repost _rate表示转发率(转发数/(转发数+评论数)). Reputation(粉丝数/(粉丝数+关注数))反应了用户的线上声望. Is _specific表示文本是否包括#等特殊符号. Is _Url表示文本是否包括URL地址符. Is _Verify表示用户是否认证. Participation(评论数/(评论数+转发数))表示用户参与该微博讨论的热情. 同时，考虑到情感分数在文本相似度计算中有重要意义，以特征向量中和情感相关的特征值为例，若该文本特征值偏向正向情感分数，而另一文本特征值偏向负向情感分数，则文本间的距离较大，反之则较小. 因此，本文加入了第8个特征：平均情感得分 U_{ass}.

U_{ass} 的计算公式如下：

$$U_{\text{ass}} = \frac{P_i - N_i}{l}.$$

式中：P_i 为文本中积极情感的词数；N_i 为消极情感的词数；l 为文本的长度. 本文采用的是Hownet情感词典. 但由于不同领域有不同的情感词，需要根据具体领域构建针对性的情感词典. 因此，在微博谣言检测中，将P图、PS、去年、前年、以前、过去等词语在谣言检测中设置为正向情感，因为这些词语多出现于反驳谣言的非谣言文本中. 同时将差评、当心、小心等词语设置为负向情感.

1.2.3　相似度计算

相似度计算公式如下:

$$Sim(m_{new},m_{topic_i})=\sum_{j=1}^{k}T_j\times(\|f_{t_j}-f_{m_j}\|_p^2)+d(m_{new},m_{topic_i}).$$

式中:m_{new}为样本消息;m_{topic_i}为样本对应的数据库中该主题下的谣言子集;f_{t_j}为谣言子集对应的第 j 个特征值;f_{m_j}为样本对应的第 j 个特征值;T_j 为该特征值的权重,$T_j\in[-1,1]$. 考虑到传统特征的现实意义,定义相似度的距离为特征向量的欧式距离与文本距离之和. 文本距离之和公式如下.

$$d(m_{new},m_{topic_i})=Aveg[d(m_{new},m_j)].$$

文本与谣言子集的距离为文本与谣言子集中的每一条消息进行距离计算,最后得到其平均值. 消息可看成是由一串词向量构成,即 $\boldsymbol{d}=[w_1\ \ w_2\ \ \cdots\ \ w_l]$. 因此计算文本间的距离,单词出现的位置及其前后词语的相关性有较大关系. 以“总统奥巴马参加演讲”与“美国总统参加演讲”为例,主要内容为总统演讲,若单考察词向量间的距离,那么词向量错位会造成计算结果大量偏失. 因此,文本距离还应考虑单词的错位计算,本文采用词向量间最小值与相移距离乘积计算结果. 文本与每一条消息的距离计算公式如下.

$$d(m_{new},m_j)=\sum_{u=1}^{l}d_{step}\cdot\min\{d(w_u,w_v)\}.$$

式中:w_u 为文本中的第 u 个单词;w_v 为谣言子集中第 j 个消息中的第 v 个单词;d_{step}为文本间单词的相对距离,如下所示.

$$d_{step}=\left|\frac{u-v}{l}\right|.$$

两个单词间的距离公式计算如下:

$$d(w_u,w_v)=\|w_u-w_v\|_p^2.$$

综上,文本距离计算算法如下:

算法 1　基于主题模型下的文本距离计算算法.

输入　训练集合 train,谣言子集 topic_rumors

输出　文本与谣言子集的距离

```
1  begin
2  for all lines∈train do
     /*得到训练集中的每行数据*/
3    TF-IDF top3-topic
     /*确定该行消息主题*/
4    for lines_j∈topic_rumors do
     /*表示谣言主题子集中的一行数据*/
5      for queryword_u∈lines do
6        for exixtword_v∈lines_j do
7          d(queryword_u,
       exixtword_v)
8            min d(queryword_u,
             exixtword_v)
9            step=|(u-v)/n|
10         end for
11         step*min
12     end for
13     ave=average d(lines,lines_j)
       /*得到新文本与谣言主题子集文本
       相似度的平均值*/
14     end for
15     return ave
16   end for
17   end
```

算法 1 开始时,得到训练集中的每行数据,每行数据表示一个文本,对该文本计算 TF-IDF 值,选择 top3 单词作为主题表达,以确定谣言子集. 对该文本与谣言子集中的每条消息分词,将文本分词后的结果遍历谣言子集,求得最小词向量间的距离并与位移作乘积运算.

2　实验和结果分析

2.1　微博数据收集与处理

2.1.1　数据收集

实验数据部分包括两部分:① 官方谣言数据库;② 微博谣言及非谣言样本.

新浪微博包含两个官方辟谣账户,其一为微博社区管理中心,另一个为微博辟谣账户,本文编写网络爬虫爬取近 1 年来官方辟谣账户的微博内容,其中随机选取 300 条微博作为已证实的官方微博谣言存储集,随机选取 1 500 条谣言微博作为训练和测试数据.

同时为了获得非谣言样本,使用网络爬虫随机爬取了微博 2 000 名用户的微博文本及用户特征等相关数据,并手动提取了非谣言数据集. 为了使谣言与非谣言在数据集中的比例近似为 1∶1,随机选择了 1 500 条非谣言微博作为训练和测试数据.

2.1.2　数据预处理

对爬取的所有数据进行数据预处理,包括 3 部分:分词、去停用词、学习词向量.

① 分词.

对主题子集下所有消息进行分词,本文使用了ansj中文分词系统,ansj是一个基于中科院ICTALAS中文分词算法的中文分词系统.

② 去停用词.

为了提高词向量的学习效率,对分词后的结果利用正则表达式去掉标点符号、特殊符号、url、的、了等共包括8项在内的停用词.

③ 学习词向量.

随后,为了使计算词语之间的语义相似度更加准确,本文中使用Mikolov等提出的SkipGram算法以整个数据集(约10万个单词)为语料库学习100维的词向量表示.

2.2 实验设置

本文选择SVM作为分类器,SVM模型利用了台湾大学林智仁副教授的LIBSVM工具,根据5折交叉验证后的结果,选用RBF核函数类型,C取值为100,gamma取值为0.14/0.125,其余采用默认参数设置.

在谣言检测任务中采用常用的评价标准:正确率、准确率、召回率和F_1值4种度量指标对实验结果进行评估.

① 正确率η表示被预测正确的消息占所有消息数据的比例,

$$\eta=\frac{t_p+t_n}{t_p+f_p+f_n+t_n}.$$

式中t_p,t_n,f_p,f_n分别表示真正例、真负例、假正例、假负例.

② 准确率α以谣言类别为例,表示该模型预测为谣言的消息数中实际为谣言的比例,

$$\alpha=\frac{t_p}{t_p+f_p}.$$

③ 召回率β以谣言类别为例,表示在实际谣言中被模型预测为谣言的比例,

$$\beta=\frac{t_p}{t_p+f_n}.$$

④ F_1值表示准确率和召回率的调和平均值,

$$F_1=\frac{2\alpha\beta}{\alpha\beta}.$$

为了检测单个消息相似度特征对谣言检测准确性的影响,本文实现了文献[11]的对比试验,将9个特征分别经过SVM分类器后得到的F_1值,与未加入特征(特征值随机)得到SVM分类器的F_1值做比较,实验结果如表1所示.

表1 单个特征F_1值

Tab.1 Single feature F_1 value

编号	特征	单个特征F_1
1	VIPRank	0.625
2	Repost_rate	0.491
3	Reputation	0.619
4	Is_specific	0.489
5	Is_Url	0.520
6	Is_Verify	0.613
7	Participation	0.553
8	Ass	0.462
9	Similarity	0.667
10	Random	0.342

从实验结果可以看出,随机初始化训练得到的SVM分类结果F_1仅为0.342,选择的9个特征的F_1值均高于随机初始化得到的基线;在9个特征中,仅使用消息相似度得到的F_1值最高,为0.667,说明使用此特征对样例检测的准确性提升最高;未加入消息相似度时VIPRank特征的F_1值最高,为0.625,说明VIP等级越高的用户可信度越高,发布谣言的概率更小.

为了检测引入消息相似度特征与未引入消息相似度特征对谣言检测准确率提升的影响,将未引入消息相似度特征的的8个特征训练SVM分类模型,得到的分类器结果与加入新特征时进行对比,得到表2和表3的实验结果.

表2 未引入消息相似度特征的结果

Tab.2 Results of the message similarity feature not introduced

检测状态	准确率	召回率	F_1值
rumors	0.858	0.864	0.861
non_rumors	0.865	0.860	0.863
正确率		0.863	

表3 引入消息相似度特征的结果

Tab.3 Results of introducing message similarity characteristics

检测状态	准确率	召回率	F_1值
rumors	0.918	0.847	0.881
non_rumors	0.860	0.926	0.892
正确率		0.887	

从实验结果可以看出,未加入消息相似度特征时,谣言检测的准确率为0.858,召回率为0.864,F_1值为0.861;加入消息相似度特征后,谣言检测准确率为0.918,召回率为0.847,F_1值为0.881.加入消息相似度特征后,整体正确率由0.863提升至0.887,提升了大约2.4%的精度,大大提高了谣言识别的准确率.

为了验证此模型谣言预防的效果,收集部分消息各个时间节点的属性值,将各时间节点的特征向量送入 SVM 训练器学习,训练结果如图 3 所示.

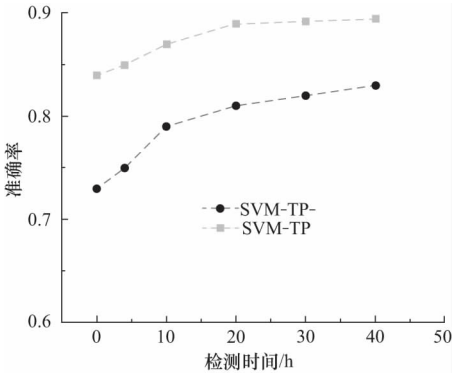

图 3　谣言预防结果图

Fig. 3　Proverbs prevention results map

结果显示,加入消息相似度特征值后,该模型可在第一时间检测出原微博是否为谣言,与其随时间变化的属性值关联较小.

3　结束语

本文提出基于主题和预防模型的谣言检测算法来解决现有谣言检测工作存在的问题. 主要包括两方面: ① 使用官方谣言数据库. 由于对数据库中谣言集按主题划分,故提高了谣言检测效率,可及时发现该主题下相关谣言. ② 提出一个新的统计特征:消息相似度. 该统计特征融合了传统谣言检测方法中提出的基于用户、社交、内容三方面的特征,因而对谣言的预防检测效果较好. 实验结果表明,本文提出的分类特征有效提高了检测谣言的正确率.

参考文献:

[1] CNNIC. 中国互联网络发展状况统计报告(2018 年 1 月)[R]. 北京:中国互联网信息中心,2018.
CNNIC. Statistical report on internet development in China(January 2018)[R]. Beijing:China Internet Information Center,2018. (in Chinese)

[2] 毛二松,陈刚,刘欣,等. 基于深层特征和集成分类器的微博谣言检测研究[J]. 计算机应用研究,2016,33(11):3369-3373.
Mao Ersong,Chen Gang,Liu Xin,et al. Research on microblog rumors detection based on deep features and integrated classifier[J]. Journal of Computer Applications,2016,33(11):3369-3373. (in Chinese)

[3] Yang F,Liu Y,Yu X,et al. Automatic detection of rumor on Sina Weibo[C]// Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. [S. l.]:ACM,2012:13.

[4] Ma J,Gao W,Wong K F. Detect rumors in microblog posts using propagation structure via Kernel learning [C]// The Meeting of the Association for Computational Linguistics. Vancouver,Canada:[s. n.],2017.

[5] Wu K,Yang S,Zhu K Q. False rumors detection on sina weibo by propagation structures[C]// Data Engineering (ICDE),2015 IEEE 31st International Conference on. [S. l.]:IEEE,2015:651-662.

[6] Jin Z,Cao J,Zhang Y,et al. News verification by exploiting conflicting social viewpoints in microblogs[C]// AAAI. Phoenix,Arizona:[s. n.],2016:2972-2978.

[7] Ma J,Gao W,Mitra P,et al. Detecting rumors from microblogs with recurrent neural networks[C]// IJCAI. New York,USA:[s. n.],2016:3818-3824.

[8] 杨文太,梁刚,谢凯,等. 基于突发话题和领域专家的微博谣言检测方法[J]. 计算机应用,2017(10):2799-2805.
Yang Wentai,Liang Gang,Xie Kai,et al. Weibo rumors detection method for unexpected topics and domain experts[J]. Computer Application,2017(10):2799-2805. (in Chinese)

[9] Jin Z,Cao J,Guo H,et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs [C]// Proceedings of the 2017 ACM on Multimedia Conference. [S. l.]:ACM,2017:795-816.

[10] Kown S,Cha M,Jung K,et al. Prominent features of rumor propagation in online social media[C]//Proceedings of the 13th IEEE International Conference on Data Mining(ICDM′2013). Dallas,Texas,USA:[s. n.],2013.

[11] 郭凯. 基于评论情感的微博谣言检测研究[D]. 大连:大连理工大学,2014.
Guo Kai. Research on Weibo proverbs detection based on commentary emotion[D]. Dalian:Dalian University of Technology,2014. (in Chinese)

[12] 赵胜辉,李吉月,徐碧瑢,等. 基于 TFIDF 的社区问答系统问句相似度改进算法[J]. 北京理工大学学报(自然科学版),2017,37(9):982-985.
Zhao Shenghui,Li Jiyue,Xu Birong,et al. Improved TFIDF-based question similarity algorithm for the community interlocution systems[J]. Transactions of Beijing Institute of Technology,2017,37(9):982-985. (in Chinese)

(责任编辑:刘芳)