情报学报　2019 年 1 月　第 38 卷　第 1 期
Journal of the China Society for Scientific and Technical Information, Jan. 2019, 38(1): 89-96

DOI: 10.3772/j.issn.1000-0135.2019.01.010

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

曾子明[1,2]，王　婧[1,2]

（1. 武汉大学信息资源研究中心，武汉　430072；2. 武汉大学图书情报实验教学中心，武汉　430072）

摘　要　网络谣言的肆虐对人们的日常生活和社会稳定造成了较为严重的负面影响，为了辅助网络谣言管控的有效推进，本文以2016年雾霾谣言为例，根据微博数据和以往研究定义了用户可信度和微博影响力特征变量，采用LDA主题模型深入挖掘微博文本的主题分布特征，并基于以上特征变量采用随机森林算法进行谣言识别的模型训练。实验表明，LDA提取的文档-主题分布特征在谣言识别中发挥了重要作用，且基于LDA的随机森林模型能够有效提高谣言识别的准确率。

关键词　微博；谣言识别；LDA；随机森林；雾霾

Research on Microblog Rumor Identification Based on LDA and Random Forest

Zeng Ziming[1,2] and Wang Jing[1,2]

(1. *Center for the Study of Information Resources, Wuhan*　430072;
2. *Laboratory Center for Library and Information Science, Wuhan*　430072)

Abstract: The spread of Internet rumors has a negative impact on everyday life and social stability. In order to assist in rumor control, this paper analyzes information about the "haze" rumors on the Sina Weibo microblogging platform in 2016, and constructs reliability and influence variables based on Weibo data and history research. In addition, the LDA model is used to gather the topic distribution of the experimental text data. Based upon the reliability variable, the influence variable, and the probability of topics, the paper uses random forest for classification to achieve rumor identification. The experiment results show that the probability of topics plays an important role in rumor identification, and that the random forest model, based on LDA, can lead to an improvement in the accuracy of rumor identification.

Key words: Weibo; rumor identification; LDA; random forest; haze

1　引　言

随着互联网和移动通信设备的飞速发展，在线社交平台成为人们发布和获取信息、发展和维系社会关系的重要渠道。微博凭借其便捷的交互方式、友好的互动体验和入驻名人的影响力吸引了大量用户。根据2017年8月联通沃指数显示，微博月活跃用户达3.3亿。作为我国活跃的社交平台之一，微博汇集了大量碎片式用户生成信息。由于社交平台的信息呈现严重混沌状态，个体认知的不确定性提

收稿日期：2017-11-03；**修回日期**：2018-08-01

基金项目：教育部人文社会科学重点研究基地重大项目“大数据资源的智能化管理与跨部门交互研究——面向公共安全领域”（16JJD870003）。

作者简介：曾子明，男，1977年生，博士，教授，博士生导师；王婧，女，1994年生，硕士研究生，主要研究领域为大数据环境下的信息资源管理、信息服务，E-mail：360548430@qq.com。

升，网络谣言因而滋生[1]。研究发现，造成较大社会影响的谣言大部分源自微博平台[2]。谣言是在含有潜在威胁的语境下产生且未经官方证实的阐述[3]。在官方渠道缺失的情境下，谣言能够在一定程度缓解人们的认知焦虑。然而，谣言的肆虐往往引发负面消极的网络舆论风波，对社会稳定和公民安全构成潜在的威胁，网络谣言的识别工作尤为关键。为提高谣言识别的准确度，本文以2016年曝光的雾霾谣言为例，首先采用LDA主题模型提取微博内容的深层次语义主题特征，并构建用户可信度特征和微博影响力特征；其次，为研究各类特征指标在谣言识别中的作用，利用随机森林方法选取不同特征训练多类模型并进行性能对比；最后使用Logistic回归和支持向量机方法进行分类训练，将其作为评估随机森林谣言识别性能的基准分类器，利用ROC曲线评估和对比三类分类器在雾霾谣言识别问题上的性能表现。

2 研究现状

谣言具有传播速度快、影响范围广等特点，因而微博谣言通常会经由复杂的在线社会网络引发更大范围的网络舆论事件[4]。目前有关谣言识别的研究主要围绕以下三方面展开：①谣言文本特征研究。张志安等[5]利用微信后台提供的共享数据探讨了微信谣言的话题特色、语言风格以及信源特色等。贺刚等[2]利用微博中hashtag、url、@等文本统计特征以及关键词分布特征构建特征群，采用SVM进行微博分类。武庆圆等[6]则针对短文本词语稀疏的特点提出一种多标签双词主题模型用于识别属于谣言、欺诈、情色、诱导分享类别的文本。②谣言发布用户特征研究。研究表明基于谣言发布用户特征的方法在谣言识别的准确度方面表现出色，其中Zhang等[7]则提取了三个消息发布用户的特征用于识别谣言。刘雅辉等[8]综合以往研究总结出谣言识别中可使用的用户基本特征、用户名、用户网络特征和行为特征。③传播网络特征研究。由于谣言与非谣言的传播模式不同，Wu等[9]基于新浪微博研究发现谣言通常被普通用户发布，再由意见领袖转发，继而辐射更大范围的普通用户。王理等[10]利用公共事件数据详细分析网络谣言生成及传播机制，其中由微博等新媒体曝光的公共事件更容易滋生谣言。蒙在桥等[11]提出一种基于在线社交网络的谣言传播模型，并使用仿真实验评估传播节点的影响力等。

在以往的谣言识别研究中，谣言内容的深层语义特征、传播用户可信度以及行为特征尚未得到很好的运用。因而本文利用LDA主题模型提取微博文本主题作为谣言识别模型训练的文档-主题特征，结合用户可信度、微博影响力特征，采用随机森林方法实现谣言识别。

3 雾霾谣言识别模型和方法设计

微博谣言识别的方法如图1所示，共分为6步：①采集目标数据；②数据预处理，清洗无效数据，并进行数据变换；③采用LDA模型训练得到微博的主题分布；④构建特征变量，利用变换后的数据构建用户可信度和微博影响力变量；⑤随机森林谣言识别训练；⑥使用SVM和LR作为基准分类器评估模型性能。

3.1 LDA主题模型

潜在狄利克雷分布模型（Latent Dirichlet Allocation）是Blei等[12]于2003年提出的一种文档主题生成模型。由于LDA能够降低文本表示维度，在语义挖掘领域得到了广泛应用[13]。LDA模型是一个三层贝叶斯网络模型，其核心思想是每个文档对应一个服从Dirichlet分布$\vec{\theta}$主题分布，每个主题对应的词分布服从Dirichlet分布$\vec{\varphi}$，其中文档-主题分布α参数和主题-词分布β参数服从Dirichlet分布$\vec{\alpha},\vec{\beta}$。

设采集M条微博文本，共有N个词，微博文本主题个数为K；从Dirichlet分布$\vec{\alpha}$中取样生成微博

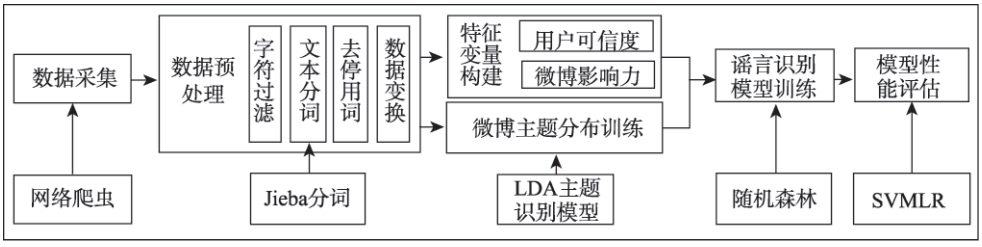

图1 谣言识别方法设计

文本的$\vec{\theta}$主题分布，根据主题分布，取样生成微博词对应的主题z；从 Dirichlet 分布$\vec{\beta}$中取样生成微博主题的$\vec{\varphi}$词分布，根据词分布，取样生成相应的词w。模型不断重复上述过程，直至所有微博文本采样完毕，最终得到每条微博文本的主题分布及各主题的词分布。

LDA 主题模型是一种无监督模型，其中主题个数是模型重要的输入参数。本文采用困惑度（perplexity）确定文档的最优主题数目。困惑度是用于评估模型优劣的标准，可用于调节主题个数，其计算公式如下[12]：

$$\text{perplexity}(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \quad (1)$$

式中，w_d表示词，$p(w_d)$表示文档中词的概率，N_d表示词的数量，M表示文档的数量，D表示文档中所有词的集合。

使用困惑度进行评估时，主题越多，困惑度数值会逐渐下降；而主题数越多，LDA 模型计算代价越大[14]。同时为了避免模型过拟合，应综合考虑选取困惑度数值和主题数目，选择困惑度最小和主题数最少的数值作为 LDA 模型训练的最优数目。

3.2　随机森林

随机森林是一种集成分类算法，它是 Breiman[15] 于 2001 年提出的由多颗随机采样变量和数据生成的分类树组成的分类器，其分类结果取决于模型所有分类树中分类结果最多的类别。随机森林在处理缺失数据和不平衡数据时表现稳健，分类性能良好，且模型训练和分类的速度快[16]，因而在文本和语言处理等领域得到了广泛应用[17]。

在谣言识别模型训练中，随机森林从采集的全部实验数据中有放回的多次重复抽样作为模型的训练集，随后从实验数据的M个特征中随机选取m个（$m<M$），并确定最佳分裂方式。按照以上的子树生成模式不断生成新的子树形成随机森林，模型的最终分类结果由子树进行投票（少数服从多数原则）得到。

4　数据获取与处理

4.1　数据采集及预处理

本文以我国环境保护部宣传教育中心和北京市环境保护宣传中心于 12 月 30 日联合曝光的 2016 年雾霾谣言和新浪微博的微博辟谣官方账号发布的辟谣微博作为谣言评判基准，采用爬虫软件爬取以关键词搜索的 2016 年新浪微博数据以及同时间范围内的非谣言微博，共采集到 1032 条微博数据，数据包含微博内容信息（文本内容、点赞数、转发数、评论数）和发布微博的用户信息（微博数、关注数、粉丝数）。在筛选删除原始数据池中的缺失、冗余和无关等数据后，最终获得 872 条有效数据。本文根据曝光的雾霾谣言对采集的数据进行人工标注，其中谣言数据 351 条，非谣言数据 521 条，数据组成如表 1 所示。

表 1　实验微博数据组成

类别	关键词	数量
2016 年雾霾谣言数据	微距 雾霾	156
	肺泡 雾霾	149
	病菌 雾霾	46
2016 年雾霾非谣言数据	雾霾	521

表 1 中“微距+雾霾”代表环保部门曝光的“网传视频用 4000 流明灯光微距镜头下显示出的北京雾霾”，“肺泡+谣言”是环保部门曝光的“80 个 PM 2.5 微粒可以堵死一个肺泡”，“病菌+雾霾”是微博辟谣官方账号 2016 年 12 月 21 日曝光的“雾霾产生的病菌侵害主体为 12 岁以下的儿童”。

数据预处理包括无关字符过滤、文本分词、去停用词、数据变换。微博文本内容分词处理是运用 LDA 主题模型前的必要步骤。因为微博内容包含大量表情、url、标点和 hashtag、@等符号，在分词处理之前应过滤微博内容中的无关字符。因而本文通过正则表达式从原始数据集中提取出文本内容，接着采用 Jieba 分词处理文本数据。分词后的数据中存在诸多停用词，为了降低主题无关词的干扰，本文根据停用词表去除文本中没有意义的词。为了提高后续模型训练的有效性，本文采用 z-score 对采集的 5 类数值属性（粉丝数、关注数、转发数、评论数、点赞数）进行规范化后用于模型训练[18]。在完成数据的预处理之后，将得到的分词结果保存在本地数据库中，以备后续分析使用。

4.2　微博谣言识别特征构建

微博强大的社交功能使得用户能够更为简单快速地将信息分享给其他用户。在复杂的社交网络结构背后，微博信息的评论、转发和点赞等互动机制

表 2 实验数据变量

特征	文档-主题分布			用户特征			微博特征			
特征指标	0	…	n	是否认证	粉丝数	关注数	已发微博	转发数	评论数	点赞数
变量	p_{mo}	…	p_{mn}	verify	follower	following	num	repost	comment	like

激励了更多用户参与社交，这也使得微博信息的覆盖范围进一步扩大，影响力也不断提高。当人们缺乏专业知识和官方渠道的信息时，谣言发布者则利用此时人们对未知事件的恐惧和焦虑心理传播谣言。结合 LDA 主题识别模型获得的文档-主题分布，本文的数据变量如表 2 所示。其中 verify_i 表示用户 u_i 是否通过新浪微博个人认证，若通过，则 verify_i 为 1，否则为 0。

中文谣言特征研究发现大量发布谣言微博的用户极有可能是网络水军[19]，谣言信息则借助微博社交网络加速扩散，网络水军通常关注数多而粉丝数少[20]，因而粉丝数比关注数的比值能够较好的反应用户的可信度。又因为发表过较多微博的用户发布谣言的可能性小[8]。因而本文定义用户可信度 Reliability(u)为：

$$\text{Reliability}(u)=\log\left(e^{\text{follower}-\text{following}}+e^{\text{num}}\right)+\text{verify} \quad (2)$$

式中，follower、following、num 分别是对粉丝数、关注数和已发微博数进行 z-score 规范化后的数值。用户可信度数值越大，表示用户的可信度越高。

本文使用用户的粉丝数、转发数、评论数和点赞数作为影响力的评估指标，粉丝数越大，则该微博涉及的人群越多，转发数和评论数能够反应用户互动程度，而点赞在一定程度上显示出用户对内容的肯定和认可，因而将微博影响力 Influence(t)定义为：

$$\text{Influence}(t)=\log\left(e^{\text{follower}}+e^{\text{repost}}+e^{\text{comment}}\right)+\text{like} \quad (3)$$

式中，follower、repost、comment、like 分别是对粉丝数、转发数、评论数和点赞数进行 z-score 规范化后的数值，其值越大，表示微博的影响力越大。

5 研究结果分析

5.1 LDA 主题识别

5.1.1 主题个数的选择

本文依照困惑度公式，计算出 2 到 30 区间内（间隔为 1）不同主题个数的困惑度数值，实验结果如图 2 所示。横轴显示主题个数，纵轴显示困惑度，从图中可以看出，随着主题个数的增加，困惑度波动变化。当主题个数为 7 时，存在一个极小值点。根据困惑度最小和主题个数最少的原则，本文选取 7 作为 LDA 模型的主题参数值。

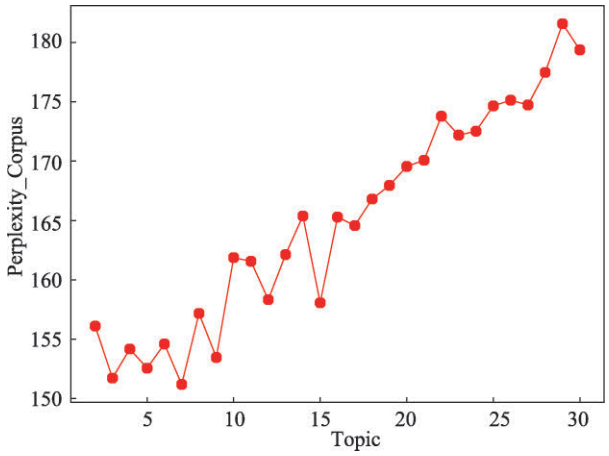

图 2 Perplexity-Topic 折线图

5.1.2 LDA 主题分类结果

在确定最优主题个数后，将分词后的数据用于 LDA 模型训练，最终得到文档-主题分布和主题-词分布，本文将获取到的 LDA 文档-主题分布作为谣言识别的文本深层语义特征。利用 LDA 模型训练得到的 7 个主题结果如表 3 所示，各主题均选取前 5 个词表示。

本文随机选取 8 篇文档绘制文档-主题分布图，如图 3 所示。从图中可以看出所选文档的某 1 个或 2

表 3 主题词分布

主题 0	瘀血	心衰	呼吸道	直接	患者
概率	0.021	0.016	0.013	0.012	0.012
主题 1	携带	病菌	专家	躲不开	传染性
概率	0.013	0.012	0.008	0.008	0.006
主题 2	发烧	特征	病菌	以下	12
概率	0.022	0.021	0.014	0.013	0.012
主题 3	疗法	肺里	吸到	网友	改变
概率	0.008	0.007	0.007	0.010	0.010
主题 4	真相	焦点	心情	改变	或许
概率	0.005	0.005	0.005	0.004	0.004
主题 5	微距	4000	流明	灯光	镜头
概率	0.033	0.031	0.029	0.028	0.027
主题 6	堵死	肺泡	PM 2.5	环境	一年
概率	0.045	0.031	0.028	0.018	0.018

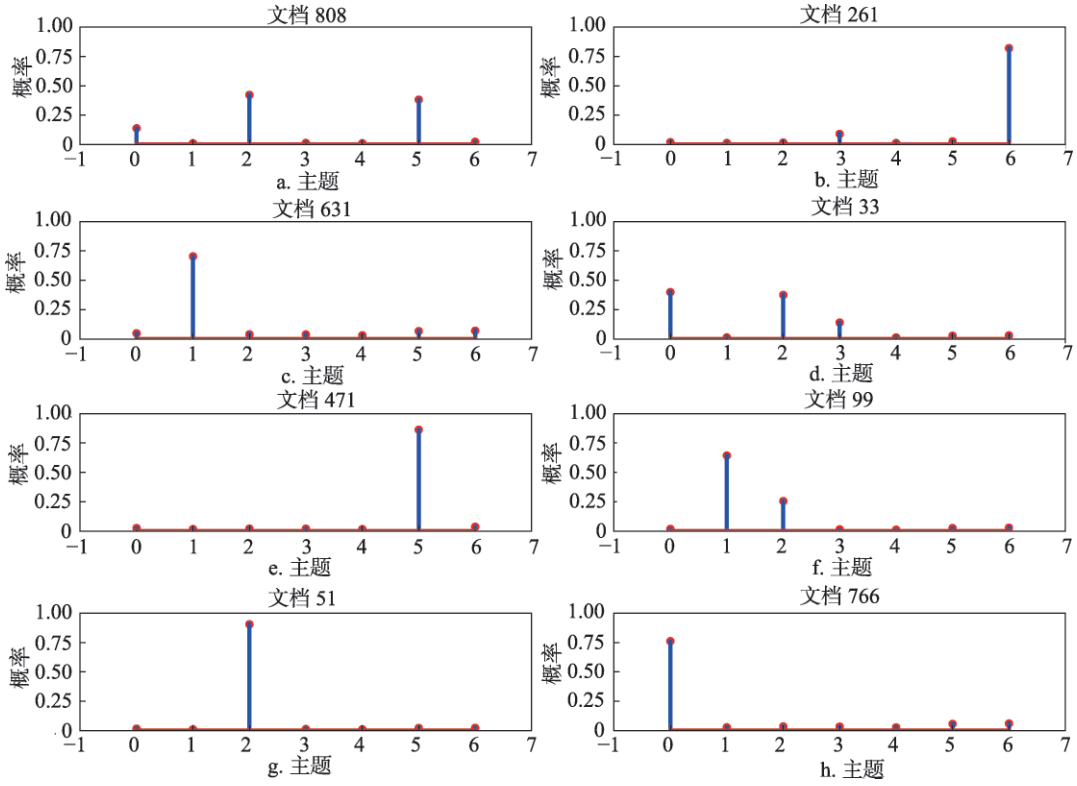

图 3　文档-主题分布

个主题概率相比其他主题概率高，即文档均存在主要主题，说明模型较好实现了微博文本主题的划分。

同时，本文根据文档-主题分布计算了所有微博的主题概率的平均值，并将结果按升序排列，概率值依次为：主题 0，0.155；主题 1，0.078；主题 2，0.115；主题 3，0.104；主题 4，0.079；主题 5，0.220；主题 6，0.248。具体如图 4 所示。横轴显示主题类别，纵轴显示概率值。其中主题 6 平均概率最大，其次是主题 5、主题 0、主题 2、主题 3、主题 4、主题 1。其中概率最高的两类主体：主题 5 和主题 6。通过对照表 2 和环保部门及微博辟谣曝光的雾霾谣言可知，主题 5 是有关微距雾霾的谣言（谣言一：微距镜头中的北京雾霾），主题 6 是有关肺泡的谣言（谣言十：雾霾堵死肺泡？）。

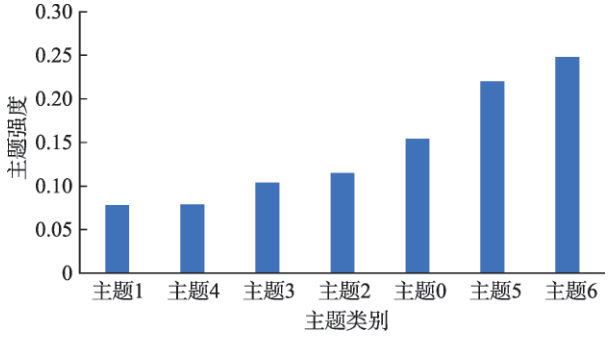

图 4　文档-主题平均概率

5.2　随机森林分类

为研究各特征变量对微博谣言识别的影响，本文基于随机森林算法选择不同特征构造五个模型，并选取准确率、召回率和 F 值比较模型分类结果。为评估随机森林模型在雾霾谣言识别的分类性能，本文采用 SVM 和 Logistic 回归作为基准分类器与之对比，将 LDA 主题模型得到的文档-主题分布和用户可信度以及微博影响力指标作为各分类器的分类特征，使用 ROC 曲线评估分类器的表现。

5.2.1　随机森林的参数调节

随机森林是由多个决策树组成的分类器。随机森林中的每颗决策树是一个弱分类器，而汇集多颗随机决策树分类结果使得模型的准确度和稳定性提高。为了有效调节模型的效率和速率，需要多次测试调整随机森林的关键参数。随机森林算法中重要的参数包括子树最大深度（max_depth）、最小样本划分数量（min_samples_split）、子树数量（n_estimators）、最小叶子节点数量（min_sample_leaf）。

其中子树最大深度是指到子树叶子节点距离的最大值。最小样本划分数量指明当样本数小于最小值，则子树不再划分。最小叶子节点数量是指子树的末端节点个数的最小值，若子树的叶子节点小于该值则会被剪枝，其值越小越容易使模型受到噪声数据干扰，导致过拟合。本文采用10折交叉验证的网格搜索方法调节随机森林的参数，并使用袋外分数（OOB score）和AUC值等来评估调参模型的优劣[21]，其中袋外分数反映了模型的泛化能力，两者数值越大，说明模型的性能越优。

5.2.2 基于不同特征的随机森林模型

其中，模型1的训练集包含全部特征变量数据，模型2仅使用样本数据中的用户可信度和微博影响力特征训练，模型3仅使用文档-主题分布特征，模型4使用文档-主题分布和用户可信度特征变量，模型5使用文档-主题分布和微博影响力特征变量，如表4所示。

表4 模型-特征变量

模型	特征		
	文档-主题分布 Distribution	用户可信度 Reliability	微博影响力 Influence
模型1 Model_1	√	√	√
模型2 Model_2		√	√
模型3 Model_3	√		
模型4 Model_4	√	√	
模型5 Model_5	√		√

针对每个模型选用基于10折交叉验证的网格搜索方法计算模型的最优参数。具体结果如表5所示。

表5 模型最优参数

模型	Parameter			
	n_estimators	min_samples_split	max_depth	min_sample_leaf
Model_1	55	10	9	4
Model_2	45	190	7	3
Model_3	30	20	9	5
Model_4	70	10	9	4
Model_5	30	10	9	4

根据表5的参数训练随机森林模型，并使用OOB Score、AUC、Precision、Recall、F-Score评估采用不同特征的模型性能，结果如表6所示。

从表6中可以看出，仅使用用户可信度和微博影响力特征的模型2的所有评估值均显著低于使用了文档-主题分布特征的模型1，说明使用LDA模型训练得到的文档-主题分布特征显著提升了随机森林在雾霾谣言识别中的分类效果。而模型3仅使用了文档-主题分布特征，其袋外分数数值小于模型1，且在雾霾谣言分类的表现显著不如模型1性能优异。虽然模型5的袋外分数与模型3基本相等，但模型4、模型5的AUC值、准确率、召回率和F值均略高于模型3，说明用户可信度特征或微博影响力特征对于雾霾谣言识别起到了正向作用。模型4的袋外分数略高于模型5的值，但模型4的AUC值、准确率、召回率和F值均低于模型5，说明在谣言识别中微博影响力特征相比用户可信度特征更有效。而模型4、模型5的各类评估数值均小于模型1的数值，说明用户可信度特征与微博影响力特征的共同使用对谣言识别的效果提升最佳。综上所述，基于文档-主题分布特征、用户可信度特征和微博影响力特征的模型表现最优。

表6 模型评估

模型	OOB_Score	AUC	Precision	Recall	F-Score
Model_1	86.483%	0.924	93.010%	93.013%	92.990%
Model_2	64.834%	0.664	68.543%	69.072%	68.421%
Model_3	85.223%	0.897	89.714%	89.691%	89.185%
Model_4	85.682%	0.912	91.737%	91.753%	91.737%
Model_5	85.223%	0.920	92.545%	92.554%	92.534%

为详细探究基于LDA训练得到的文档-主题分类特征的各模型的变量对谣言分类的影响，本文利用随机森林提供的特征选择方法计算模型1、模型3、模型4、模型5中各特征变量的重要度。研究表明，平均精确率减少（mean decrease accuracy）方法相比于平均不纯度减少方法（mean decrease impurity）效果更好[22]，因而本文选择平均精确率减少方法衡量变量的重要度。平均精确率减少的主要思想是通过随机打乱原始各特征变量的值，计算乱序后的特征值对模型准确率的影响。若特征变量对模型分类越重要，则乱序后的数据会使模型精确率降低越多。各类模型的特征指标重要度如图5所示。

由图5可以看出，主题5、主题6在四类谣言识别模型训练中均起到了重要作用。在模型1和模型5的特征变量重要度条状图中，微博影响力特征重要度仅次于主题5、主题6；在模型1和模型4中，用户可信度特征重要度均大于主题4的重要度，说明论文构建的微博影响力和用户可信度变量对谣言识别是有效的。与此同时，模型4中主题5、主题6的重要度比例（某特征变量重要度/所有特征变量

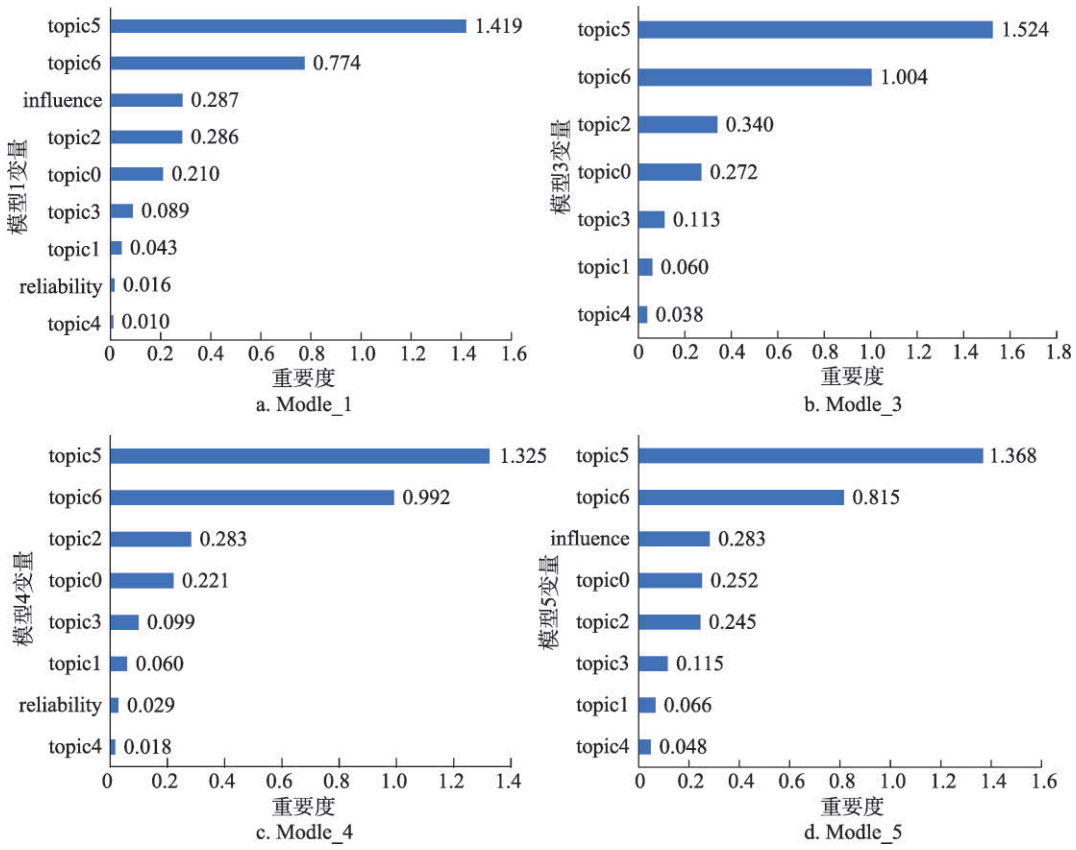

图 5　基于 LDA 的模型特征变量排序

重要度的和）分别为 43.785%、32.762% 均大于模型 5 中相应的主题重要度比例 42.844%、25.538%。而模型 4 中用户可信度特征重要度比例为 0.965% 显著低于模型 5 中微博影响力重要度比例 8.849%，因此微博影响力特征对谣言识别的性能提升作用高于用户可信度特征。

5.2.3　随机森林分类性能比较

为了评估随机森林在雾霾谣言识别问题中的性能，本文选择 SVM 和 Logistic 回归作为性能比较的基准分类器，使用 ROC 曲线进行分类模型效能评估。又因为基于 LDA 的随机森林谣言识别模型中，使用 LDA 文档-主题分布、用户可信度和微博影响力特征变量的分类器表现最优，所以将其作为随机森林、SVM 和 Logistic 回归的输入特征变量，并随机抽取 80% 的样本数据作为模型的训练数据。SVM（Support Vector Machine）支持向量机是以统计学习理论为理论为基础的一种有监督学习模型，能够有效处理小样本、高维数据，拥有较好的泛化能力[23]。本文使用基于高斯核函数的 SVM 进行模型训练，同样采用基于 10 折交叉验证的网格搜索方法确定模型的最优参数。罗吉斯回归是一种简单实用的广义线性回归模型，适合处理大规模数据，并且在二分类问题中得到了广泛应用，本文采用基于 sigmoid 函数的 Logistic 回归方法进行模型训练。经过参数优化后的模型训练结果如图 6 所示。

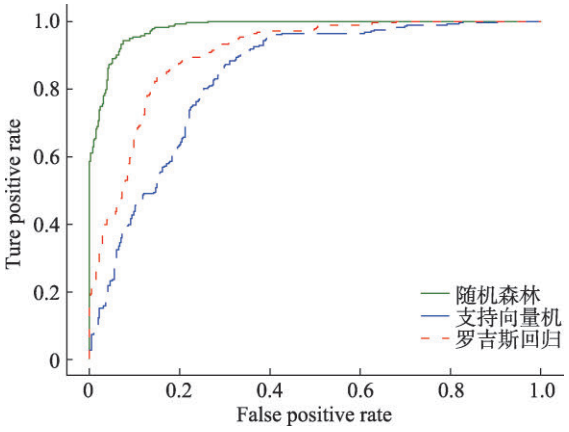

图 6　ROC 曲线

图 6 中横轴是 FPR，纵轴是 TPR，实线表示的是随机森林，虚线代表支持向量机，点状线代表罗吉斯回归。在 ROC 曲线中越靠近左上角，则模型

的性能越好。由图可知，随机森林的表现最优，其次是罗吉斯回归，最后是支持向量机。本文分别计算了各模型ROC曲线下的面积（即AUC值），随机森林的AUC值为0.915，罗吉斯回归的AUC值为0.827，支持向量机的AUC值为0.767。综上所述，在使用主题-文档分布、用户可信度特征和微博影响力特征变量的分类器中，随机森林模型在雾霾谣言分类问题中表现优异，能够较为准确地实现谣言识别。

6 结 论

在社交网络蓬勃发展的现代社会，便捷互联的社交平台成为网络谣言加速扩散的关键渠道。在官方信息缺失和知识不足的情况下，网络谣言降低了用户的信息焦虑。谣言传播过程中，部分的谣言接受者会转为谣言的扩散者，谣言得以在用户的个人社交网络中继续传播，从而引发更广范围的人群恐慌或愤怒等负面情绪。因此，社交平台谣言识别工作迫在眉睫。本文以环保部门和微博辟谣曝光的2016年雾霾谣言为例，从新浪微博着手展开谣言识别研究。文章基于LDA主题识别模型深入挖掘微博文本语义信息，获取文档-主题分布特征，将其与定义的用户可信度和微博影响力特征变量作为随机森林的输入变量进行分类训练，并进一步探讨了特征变量的重要度和模型性能。实验表明，基于LDA主题模型的谣言识别分类器效果显著优于未使用LDA文档-主题分布特征的分类器。在所有特征变量中，文档-主题分布特征对模型准确率贡献较高。在模型性能评估中，基于LDA的随机森林模型性能显著优于基于LDA的Logistic回归和支持向量机模型。综上所述，本文采用的基于LDA的随机森林谣言识别方法能够有效挖掘谣言文本深层语义信息，且较为准确地实现了谣言识别，为社交平台网络谣言管控提供了一种有价值的模式参考。未来的研究可以针对谣言传播的时序特征、网络特征或谣言内容中的情感信息进行深入挖掘。

参 考 文 献

[1] 李桂华, 王亚男, 朱一凡. 网络谣言的信息接收反应机制及其风险治理[J]. 情报学报, 2014, 33(3): 305-312.

[2] 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.

[3] 闵庆飞, 刘晓丹. 谣言研究综述: 基于媒介演变的视角[J]. 情报杂志, 2015, 34(4): 104-109.

[4] 李丹丹, 马静. 复杂社会网络上的谣言传播模型研究综述[J]. 情报理论与实践, 2016, 39(12): 130-134.

[5] 张志安, 束开荣, 何凌南. 微信谣言的主题与特征[J]. 新闻与写作, 2016(1): 60-64.

[6] 武庆圆, 何凌南. 基于多标签双词主题模型的短文本谣言分析研究[J]. 情报杂志, 2017, 36(3): 92-97.

[7] Zhang Q, Zhang S, Dong J, et al. Automatic detection of rumor on social network[M]//Natural Language Processing and Chinese Computing. Cham: Springer, 2015: 113-122.

[8] 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): 1536-1545.

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

[10] 王理, 谢耘耕. 公共事件中网络谣言传播实证分析——基于2010~2012年间网络谣言信息的研究[J]. 上海交通大学学报(哲学社会科学版), 2014, 22(2): 86-99.

[11] 蒙在桥, 傅秀芬, 陈培文, 等. 基于OSN的谣言传播模型及影响力节点研究[J]. 复杂系统与复杂性科学, 2015, 12(3): 45-52.

[12] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[13] Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering[J]. Machine Learning, 2001, 42(1): 143-175.

[14] 张志飞, 苗夺谦, 高灿. 基于LDA主题模型的短文本分类方法[J]. 计算机应用, 2013, 33(6): 1587-1590.

[15] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[16] Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author)[J]. Statistical Science, 2001, 16(3): 199-231.

[17] 邓生雄, 雒江涛, 刘勇, 等. 集成随机森林的分类模型[J]. 计算机应用研究, 2015, 32(6): 1621-1624.

[18] Han J W, Kamber M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001.

[19] 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.

[20] 袁旭萍, 王仁武, 翟伯荫. 基于综合指数和熵值法的微博水军自动识别[J]. 情报杂志, 2014, 33(7): 176-179.

[21] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 33-37.

[22] Wolfe F, Clauw D J, Fitzcharles M A, et al. The American college of rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity[J]. Arthritis Care & Research, 2010, 62(5): 600-610.

[23] 汪海燕, 黎建辉, 杨风雷. 支持向量机理论及算法研究综述[J]. 计算机应用研究, 2014, 31(5): 1281-1286.

（责任编辑 王克平）