

基于 ARIMA 和改进 LSTM 的突发事件舆情热度预测

许露萌

(上海工程技术大学 管理学院, 上海 201600)

摘要: 针对突发重大公共卫生事件的网络舆情热度预测模型存在较大误差问题, 文章选取百度指数作为舆情热度, 提出了 AMRIMA 预测和改进的 LSTM 预测方法。首先, 使用 ARIMA 模型预测新冠肺炎疫情以天为单位的舆情热度; 其次, 使用改进 LSTM 预测新冠肺炎疫情以小时为单位的舆情热度, 在 LSTM 中加入注意力机制, 实现了提高预测精度的目的; 最后, 得出预测结果。实验结果表明, ARIMA 算法和改进 LSTM 算法能更准确预测网络舆情热度值, 有效提高了网络舆情热度的预测精度。

关键词: 突发事件; 舆情热度; ARIMA; 改进 LSTM

中图分类号: TP301.6

文献标识码: A

DOI: 10.19358/j.issn.2097-1788.2023.03.009

引用格式: 许露萌. 基于 ARIMA 和改进 LSTM 的突发事件舆情热度预测 [J]. 网络安全与数据治理, 2023, 42(3): 53-57.

Prediction of public opinion popularity in emergencies based on ARIMA and aproved LSTM

Xu Lunmeng

(School of Management Studies, Shanghai University of Engineering Science, Shanghai 201600, China)

Abstract: In view of the large error in the prediction model of the network public opinion popularity for major public health emergencies, the article selects Baidu Index as the public opinion popularity, and proposes ARIMA prediction and aproved LSTM prediction methods. Firstly, ARIMA model is used to predict the public opinion popularity of COVID-19 epidemic in days; Secondly, aproved LSTM was used to predict the public opinion popularity of COVID-19 epidemic in hours. The attention mechanism is added to LSTM to improve the prediction accuracy; Finally, the prediction results are obtained. Experiments show that ARIMA algorithm and aproved LSTM algorithm can accurately predict the network public opinion popularity and effectively improve the prediction accuracy of the network public opinion popularity.

Key words: emergency; public opinion popularity; ARIMA; aproved LSTM

0 引言

预测网络舆情热度有助于及时掌握舆情动态, 提高应对舆情危机的能力。但突发重大公共卫生事件的网络舆情具有长周期性、快传播性以及强破坏性的特点, 其舆情扩散更容易导致舆情危机的出现、发酵使其难以控制^[1]。随着新冠肺炎疫情的出现, 国家和政府越来越重视“生物安全”, 尤其是可能迟滞或中断民族复兴历史进程的风险挑战, 当然也构成人类生存面临的重要威胁^[2]。过往对网络舆情的研究内容主要是对研究对象的文本内容进行分词与词性分析、演化规律以及影响机理进行分析^[3]。从风险的社会放大理论来看, 网络舆情会在短时

间内扩散并隐藏着潜在的风险^[4]。如果对突发事件的舆情预测处理不当, 会随着消极情绪的蔓延, 形成“涟漪效应”趋势^[5], 导致更为严重的舆情危机^[6], 加剧网络舆情的监管难度。

目前, 国内外的许多学者在突发事件舆情热度预测方面做了大量的研究。王努努^[7]等通过情感分类分析出网民的不同情感态度, 随着时间的变化对不同情感进行情感预测。舆情热度研究一般分为以单一变量的变化规律和趋势进行舆情热度的预测研究, 和以多个变量构建舆情热度模型进行舆情热度预测^[8]。在舆情热度预测方法上, 张和平等运用百度搜索指数作为舆情热度指标, 利用灰色马尔科夫模型对舆情热度进行预测^[9]; 王宁等

利用百度指数和微指数来衡量舆情热度的大小,使用灰色模型进行舆情热度预测^[10];陈福集等基于残差修正的多因素灰色模型来对网络舆情进行预测^[11];兰月新等首先建立舆情热度模型,然后在建立模型的基础上使用多维度 Logistic 模型对舆情热度预测^[12];Chen 使用模糊综合评价法选定舆情热度的指标并进行预测^[13];曾子明等构建微博舆情热度评价指标体系,利用信息熵确定各个指标的权重,使用 BP 神经网络的对突发传染病舆情热度趋势预测并验证模型的可行性^[14];徐敏捷等将 Logistic 模型、指数平滑法和灰色预测结合起来预测网络舆情^[15]。以上的国内外的网络舆情热度预测大多数采用的是机器学习算法和深度学习算法,大多数使用单一的模型或者组合模型,优点是相对简单,预测效率较高,缺点是不能适应不同类型时间序列数据的需要,一定程度上影响整体的预测效果。结合突发公共卫生事件舆情治理和目前已有的舆情预测研究成果,文章通过 ARIMA 算法和加入注意力机制 LSTM 对新型冠状病毒感染肺炎事件进行舆情热度预测,使用准确率更高的算法进行预测仿真得出舆情热度的发展趋势,并对舆情监测和引导提出可行性建议。

1 研究方法

1.1 ARIMA 算法

ARIMA 算法,也叫做差分自回归移动平均模型,是利用时间序列历史值预测未来对应值的一种方法,也是时间序列预测的一种典型方法^[16]。ARIMA 算法是由 Box、Jenkins 和 Reinsel 于 1976 年首次提出,该算法在时间序列中广泛应用,并且有很好的拟合效果。以下是 ARIMA (b, p, q) 算法的表达式:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

其中 y_t 是预测值, t 为事件变量, ε_t 是白噪声序列, φ_p 与 θ_q 是系数。

1.2 改进的 LSTM 算法

1.2.1 LSTM 神经网络

递归循环神经网络 RNN^[17]将前一时刻的输出作为当前时刻的输入,并在隐藏层内部保持着一个包含数据历史特征的隐藏状态,因此在时间序列预测上具有较大的优势。输入数据经过包含 N 个 RNN 单元的神经网络,将得到隐藏状态 $h_t = \{h_{t,1}, h_{t,2}, \cdots, h_{t,N}\} \in R^{r \times N}$ 和输出 y_t , 公式为:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (2)$$

$$y_t = W_{hy}h_t + b_y \quad (3)$$

式中 W_{xh} 、 W_{hh} 和 W_{hy} 为权值矩阵, b_h 、 b_y 为偏置向量。

理论和经验表明, RNN 在序列较长时的预测效果是很差的^[18]。LSTM 神经网络通过引入门结构的设计,拥有了回忆过去信息的能力,适用于解决较长时间序列的问题。LSTM 单元的门结构由遗忘门、输入门和输出门组成。输入数据 x_t 在 LSTM 单元中的前向传播过程可以由 (4) 至 (9) 表示:

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f) \quad (4)$$

$$i_t = \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \quad (5)$$

$$C'_t = \tanh(W_{ch}h_{t-1} + W_{cx}x_t + b_c) \quad (6)$$

$$C_t = i_t \odot C'_t + f_t \odot C_{t-1} \quad (7)$$

$$o_t = \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o) \quad (8)$$

$$h_t = \tanh[C_t] \odot o_t \quad (9)$$

式中 (4)、(5) 和 (8) 分别表示在遗忘门、输入门和输出门中的更新过程, f_t 、 i_t 、 o_t 为每个门结构的输出; W_{fh} 、 W_{fx} 、 W_{ih} 、 W_{ix} 、 W_{ch} 、 W_{cx} 、 W_{oh} 、 W_{ox} 和 b_f 、 b_i 、 b_c 、 b_o 为网络通过训练得到的权重矩阵和偏置向量; σ 、 $\tanh$ 分别表示 sigmoid 函数和双曲正切函数, C_t 为记忆状态, o_t 结合输出门信息得到当前 LSTM 单元输出的隐藏状态 h_t 。

1.2.2 改进的 LSTM 预测算法: 加入注意力模块

对于舆情热度预测,必须选取足够长的时间序列长度,才能体现出舆情热度的典型变化趋势。因此,仅利用 LSTM 无法很好地处理舆情热度预测问题。为了解决上述问题,将注意力机制与 LSTM 相结合,共同把握长时间序列的特征,注意力机制用来计算每一次历史舆情热度 $\{r_1, r_2, \cdots, r_{n-1}\}$ 与当前舆情热度 r_n 的相关性,从而构建当前时刻的上下文向量 c_n 。上下文向量 c_n 表征当前出行与历史出行的依赖关系,计算方法如式 (10) 和式 (11):

$$c_n = \sum_{k=1}^{n-1} \alpha_k h_k \quad (10)$$

$$\alpha_k = \text{softmax}(h_n W_c h_k) \quad (11)$$

式中 h_n 、 h_k 分别为当前时刻与 k 时刻 ($k < n$) LSTM 的隐藏层输出向量; W_c 为线性层;为 k 时刻特征对应的权重。得到上下文向量 c_n 后,将 c_n 与 h_n 拼接后输入线性层,并用 softmax 激活函数将线性层的输出转化为概率输出:

$$o_n = \text{softmax}(W_{out}u_n) \quad (12)$$

式中 W_{out} 为输出线性层, $u_n = c_n + h_n$ 。

2 案例研究

2.1 数据采集

2019 年末与 2020 年初开始出现的新冠肺炎疫情传播速度快、涉及范围广、影响深远。文章选择该事件的全

国舆情数据，在百度指数官网上获取新型冠状病毒感染肺炎事件的关键词，时间范围界定在 2020 年 1 月 9 日 - 2020 年 2 月 19 日，以百度指数为舆情热度数据作为 ARIMA 舆情热度预测模型的输出序列。然后在知微网站平台以小时为单位搜索微博数据作为舆情热度数据，用做改进的 LSTM 舆情热度预测。

2.2 ARIMA 模型分析

2.2.1 ADF 平稳性检验

ADF 检验如表 1 所示，检验所选取数据的平稳性可以得出：基于舆情热度，当差分为 0 阶时，显著性 P 值为 0.211，不能拒绝原假设，舆情热度序列为不平稳的事件序列；当差分为 1 阶和 2 阶时，显著性 P 值都为 0.000***，水平上呈现显著性，拒绝原假设，舆情热度序列是平稳的时间序列。综上，舆情热度序列满足平稳性的条件。

表 1 舆情热度 ADF 检验表

ADF 检验表							
变量	差分阶数	t	p	AIC	临界值		
					1%	5%	10%
舆情热度	0	-2.186	0.211	865.561	-3.616	-2.941	-2.609
	1	-6.335	0.000***	846.015	-3.606	-2.937	-2.607
	2	-9.068	0.000***	821.291	-3.616	-2.941	-2.609

注：***、**、* 分别代表 1%、5%、10% 的显著性水平

2.2.2 差分自相关和偏自相关分析

ADF 平稳性分析后，对舆情热度数据进行差分自相关和偏自相关分析，并将结果进行可视化。基于 AIC 信息准则自动寻找最优参数，得到 b, p, q 值，建立 ARIMA (0, 1, 0) 模型，如图 1 所示。

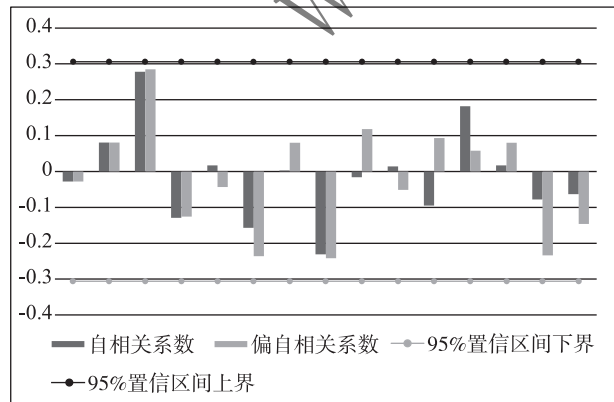


图 1 差分自相关和偏自相关图

由表 2 可以得到，Q6 在水平上不呈现显著性，不能拒

绝模型的残差为白噪声序列的假设，同时模型的拟合优度 R^2 为 0.841，表明模型表现优秀，能够满足预测的需求。

表 2 ARIMA 模型 (0, 1, 0) 检验表

项	符号	值
Q 统计量	Q6 (p 值)	0.035 (0.852)
	Q12 (p 值)	5.98 (0.425)
	Q18 (p 值)	11.396 (0.495)
	Q24 (p 值)	13.369 (0.769)
	Q30 (p 值)	14.235 (0.941)
拟合优度	R^2	0.841

2.2.3 舆情热度预测分析

当平稳性检验和相关性分析都满足条件后，ARIMA 模型的拟合优度 R^2 为 0.841，时间序列预测图如图 2 所示。拟合优度和拟合效果图反映出构建的 ARIMA 预测模型的拟合效果较好。用该模型预测未来 3 天的舆情热度趋势，结果发现舆情热度可能会呈上升趋势，相关部门要提前制定相应政策，避免出现舆情危机。



图 2 ARIMA 预测

2.3 改进的 LSTM 预测算法分析

2.3.1 模型训练

在 LSTM 模型训练之前，要对特征进行预处理，采用归一化的方法，形成无量纲等时距的数据样本，归一化计算式为：

$$x_t^* = \frac{x_t - x_{\min}}{x_{\max} - x_{\min}} \quad (11)$$

损失函数选用适合解决回归问题的均方误差函数 (MSE)，来检验模型训练集和验证集的预测值与真实值之间的误差。计算式为：

$$\xi = \frac{1}{N} \sum_{i=1}^N (y_n - \hat{y}_n)^2 \quad (12)$$

式中 x 为均方误差值, N 为样本数, y_n 为真实值, y_n 为预测值。 x 的值越大, 表明预测效果越差。

虽然 LSTM 具有提取长序列时序特征的能力, 但时序长度过长时, LSTM 的性能会快速下降。对于舆情热度预测, 必须选取足够长的时间序列长度, 才能体现出舆情热度的典型变化趋势。因此, 仅利用 LSTM 无法很好的处理舆情热度预测问题。为了体现本文提出方法的优越性, 在训练集、验证集与测试集的划分标准相同且网络训练过程中的迭代次数相同的情况下, 分别采用 LSTM 和改进

LSTM 模型进行舆情热度预测, 并进行对比分析, 如图 3 所示。

从模型拟合图可以得出, 改进的 LSTM 算法预测舆情热度在迭代的过程中拟合效果更好。从 RMSE 可以看出, 传统的 LSTM 模型的 RMSE 值为 52.67, 加入注意力机制的 LSTM 测试集 RMSE 值为 49.41, 传统的 LSTM 模型测试集的 RMSE 值比加入注意力机制的 LSTM 测试集的 RMSE 值高出 3.26, 表明改进的算法准确率有所提升, 也证明了加入注意力机制的 LSTM 舆情热度预测模型的可行性。

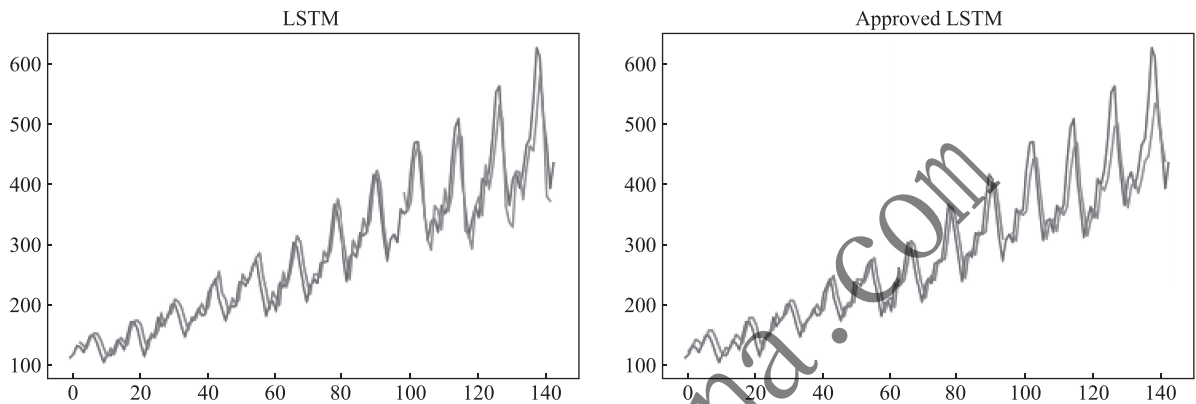


图 3 LSTM 和改进 LSTM 模型预测对比

表 3 两种预测 RMSE 对比

RMSE	训练集	测试集
LSTM	19.81	52.67
Approved LSTM	22.86	49.41

2.3.2 舆情热度预测

为了更好地说明本文提出方法的准确性与可行性, 选择以小时为单位的舆情热度数据对未来 20 小时的舆情热度进行预测 (如图 5 所示), 深色是真实值, 浅色是预测值。可以得出, 未来 20 小时舆情热度有下降且处于平稳状态, 但是舆情热度值仍处于高位, 需要采取措施进行疏导。

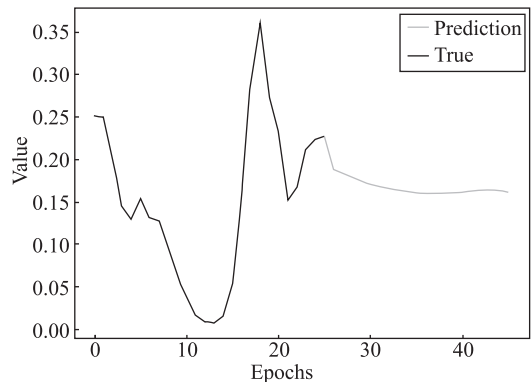


图 4 未来 20 小时舆情热度预测

3 结论

综上所述, 本文以百度指数和微博数据为对象, 以 2020 年 1 月 9 日 - 2020 年 2 月 19 日的新冠肺炎疫情舆情热度数据为基底, 通过使用 ARIMA 模型和加入注意力机制的 LSTM 模型来预测未来 3 天的舆情热度和未来 20 小时的舆情热度。结果表明, ARIMA 模型能够更为准确的预测舆情热度, 加入注意力机制的 LSTM 相比于 LSTM 准确度有所上升, 能够更准确的预测以小时为单位的舆情热度。

ARIMA 模型预测出未来 3 天舆情热度值是呈上升趋势, 改进的 LSTM 模型预测出未来 20 小时舆情热度虽然有所下降, 但是仍处于高位。如果政府想要降低舆情热度, 防止舆情危机的出现, 可以通过发布疫情的最新信息、科普疫情知识、公正公开正在施行的防疫举措等, 不断宣传防疫取得的阶段性成果, 疏导民众恐慌情绪, 从而能够更快的恢复正常生活。

参考文献

- [1] 张蒙. 突发公共卫生事件网络舆情发酵机理研究 [J]. 现代情报. 2020, 40(9): 20-31.
- [2] 周鑫, 栾媛媛. 情报分析视角下新型冠状病毒肺炎疫情研究 [J]. 科技情报研究, 2020, 2(4): 90-102.
- [3] 王连喜. 网络舆情领域相关概念分布及其关系辨析 [J]. 现代情报, 2019, 39(6): 132-141.

- [4] 全燕. 基于风险社会放大框架的大众媒介研究 [D]. 武汉: 华中科技大学, 2013.
- [5] 温芳芳, 马书瀚, 叶含雪, 等. “涟漪效应”与“心理台风眼效应”: 不同程度 COVID-19 疫情地区民众风险认知与焦虑的双视角检验 [J]. 心理学报, 2020, 52(9): 1087-1104.
- [6] 徐晓日. 网络舆情事件的应急处理研究 [J]. 华北电力大学学报: 社会科学版, 2007, (1): 89-93.
- [7] 王努努, 张伟佳, 钮亮. 基于 ARIMA 和 BP 神经网络模型的舆情情感预测 [J]. 电子科技, 2016, 29(5): 83-87.
- [8] 吴亚彬. 基于情感时间序列的 P2P 网贷企业舆情态势研判技术与实现 [D]. 北京: 北京邮电大学, 2019.
- [9] 张和平, 陈齐海. 基于灰色马尔可夫模型的网络舆情预测研究 [J]. 情报科学, 2018, 36(1): 75-79.
- [10] 王宁, 赵胜洋, 单晓红. 基于灰色系统理论的网络舆情预测与分级方法研究 [J]. 情报理论与实践, 2019, 42(2): 120-126.
- [11] 陈福集, 史蕊. 基于残差修正的多因素灰色模型的网络舆情预测研究 [J]. 情报科学, 2017, 35(9): 131-135.
- [12] 兰月新, 刘冰月, 张鹏, 等. 面向大数据的网络舆情热度动态预测模型研究 [J]. 情报杂志, 2017, 36(6): 105-110, 147.
- [13] CHEN X G, DUAN S, WANG L. Research on Trend Prediction and Evaluation of Network Public Opinion [J]. Concurrency Computation: Practice Experience, 2017, 29(24): e4212.
- [14] 曾子明, 黄城莺. 基于 BP 神经网络的突发传染病舆情热度趋势预测模型研究 [J]. 现代情报, 2018, 38(5): 37-44, 52.
- [15] 徐敏捷, 兰月新, 刘冰月. 基于组合预测的网络舆情数据预测模型研究 [J]. 情报科学, 2016, 4(12): 40-45, 87.
- [16] 王燕. 时间序列分析: 基于 R [M]. 北京: 中国人民大学出版社, 2015.
- [17] HYNDMAN R J, ATHANASOPOULOS G. Forecasting: principles and practice [M]. (2nd edition). Melbourne, Australia: OTexts, 2018.
- [18] 徐建华. 现代地理学中的数学方法 [M]. 北京: 高等教育出版社, 2002: 342-344.
- (收稿日期: 2023-01-03)
- 作者简介:**
许露萌 (1996-), 女, 硕士研究生. 主要研究方向: 信息管理。

(上接第 52 页)

- [12] 江涛. 民事诉讼效率研究——以程序设置为主要视角 [D]. 上海: 复旦大学, 2011.
- [13] 谢佑平, 万毅. 法理视野中的刑事诉讼效率和期间: 及时性原则研究 [J]. 法律科学 (西北政法学院学报), 2003(2): 8.
- [14] 毛其淋, 许家云. 中国企业对外投资是否促进了企业创新 [J]. 世界经济, 2014(8): 28.
- [15] COUMARELOS A, DEBORAH, MACCOURT, et al. How Are Legal Problems Finalised in Australia? [J]. Updating Justice, 2013(19): 1-3.
- [16] 张卫平. 在线诉讼: 制度建构及法理——以民事诉讼程序为中心的思考 [J]. 当代法学, 2022(3): 17.
- [17] HOLLANDER-BLUMOFF R, TYLER T R. Procedural Justice and the Rule of Law: fostering legitimacy in alternative dispute resolution [J]. Journal of Dispute Resolution, 2011(1): 9.
- [18] P CASHMAN, GINNIVAN E. Digital justice: online resolution of minor civil disputes and the use of digital technology in complex litigation and class actions [J]. Macquarie Law Journal, 2019(19): 39-79.
- [19] SELA A. Streamlining justice: how online courts can resolve the challenges of pro se litigation [J]. Cornell Journal of Law and Public Policy, 2016, 26(2): 331-388.
- (收稿日期: 2023-02-20)
- 作者简介:**
王立源 (1994-), 男, 博士研究生, 主要研究方向: 诉讼法学与司法制度。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com