基于时间编码LSTM的高校舆情热点趋势预测研究

易杰[1]，曹腾飞[1]，黄明峰[2]，黄肖翰[1]，张子震[1]
1. 青海大学计算机技术与应用系，青海 西宁 810016；
2. 云上贵州大数据产业发展有限公司，贵州 贵阳 550081

摘要
随着互联网技术的发展，网络舆情热点信息能在短时间内迅速传播。预测舆情热点的发展趋势，有助于高校对学生思想健康状况进行分析管理，也是当下网络舆情信息研究领域的重要课题。针对微博中的舆情信息文本，构建基于时间编码长短期记忆网络（LSTM）的高校舆情热点趋势预测模型，并与支持向量机、循环神经网络两种模型的预测效果进行对比，验证了基于时间编码的LSTM算法在舆情趋势预测上的准确率。最后，利用微博中的高校实时舆情事件对构建的模型预测效果进行评估，并动态调整评估参数，实现了对评估性能的优化，预测效果得到了显著提升。

关键词
长短期记忆网络；热点预测；高校舆情；时序数据；时间编码

中图分类号：TP391 文献标志码：A doi：10.11959/j.issn.2096-0271.2022034

Research on trend prediction of time-coded LSTM based public opinion hot spots in universities

YI Jie[1], CAO Tengfei[1], HUANG Mingfeng[2], HUANG Xiaohan[1], ZHANG Zizhen[1]
1. Department of Computer Technology and Applications, Qinghai University, Xining 810016, China
2. Guizhou-Cloud Big Data Industry Development Co., Ltd., Guiyang 550081, China

Abstract
With the development of Internet technology, network public opinion hot information can be quickly spread in a short time. Predicting the development trend of public opinion hot spots is helpful to the analysis and management of college students' ideological health, and it is also an important issue in the field of network public opinion information research. Aiming at the public opinion information text in microblog, the hot spots trend prediction model of universities based on time-coded long short-term memory (LSTM) was constructed. Compared with the prediction effect of support vector machine, and recurrent neural network through experiments, the superiority of time-coded LSTM was verified.

Finally, the prediction effect of time-coded LSTM was evaluated by using the real-time public opinion events of colleges and universities in microblogs, and the evaluation parameters were dynamically adjusted to optimize the performance of the evaluation, and the prediction effect was improved significantly.

Key words
LSTM, hot spots prediction, public opinion in universities, time series data, time-coded

0 引言

随着互联网通信技术的快速发展，多样的新媒体平台（如微博、抖音、贴吧等平台）将信息及时推送给用户，使得社会发生的实时新闻能迅速传播。据中国互联网络信息中心发布的第47次《中国互联网络发展状况统计报告》统计，截至2020年12月，我国网民规模达9.89亿，其中学生占比最高[1]，达到21%。新时代下高校学生作为互联网用户的主要群体，在网络上的参与度以及活跃度比较高。当高校突发一些热点事件时，由于高校学生思想活跃并且乐于表达自我看法，实时热点问题会引发激烈的讨论[2]。若舆情信息的价值取向是负面的，则极易带偏高校学生的思想观念，从而引发一系列高校舆情管理问题，高校舆情管理的重要性不言而喻[3]。

近些年，高校舆情事件频频发生，其中舆情信息的主题主要围绕社会时事、校园安全、师风师德、学术造假等方面，例如研究生校内身亡、高校实验室爆炸、学生违纪违法等事件。在事情的真相还未正式公布时，网络上各种评论的助推极易导致错误的舆情发展方向，引发一系列高校以及社会舆情管理问题[4]。高校舆情发展一般是阶段性的，初期由个别大学生在网络上发布自己对某个问题的想法，而后随着时间的推移，逐渐引起大范围的关注，引发更多的讨论。一般情况下，网络舆情的发展趋势遵循新闻传播学中的“沉默螺旋效应”，大多数人支持的意见会因为更多的人赞同而越来越流行；而少数人支持的观点会逐渐减少直至最后消失[5]。基于此原理，若舆情的发展趋势能比较及时、准确地被预测，高校有关部门就能在短时间采取相应的应对措施，合理地解决问题，以达到对舆情发展进行管控的目的。因此，对高校舆情的发展趋势进行预测，有助于新媒体时代下的大学校园完善管理体系，及时预测舆情发展趋势并加以正确的引导[6]，能极大地提升高校对突发舆情事件的处置水平[7-8]。

基于上述分析，对舆情热度的预测分析显得尤为重要，不仅关乎高校学生的思想健康发展，而且关乎整个社会的价值观取向和稳定性。由于舆情信息的发展一般会随着时间变化，当获取到舆情的时序数据后，需要对数据进行分析处理，找到数据的变化和发展趋势，对未来舆情事态的发展做出预测，以便及时管控。随着互联网技术的发展，在大数据和人工智能等技术的推动下，时序数据处理的有效性逐渐提高。因此，本文利用长短期记忆网络（long short-term memory，LSTM）对时序数据处理的有效性，研究基于时间编码的LSTM模型。LSTM对时序数据的处理具有极大的优势，但是其只考虑了数据相对的先后顺序，不包含绝对的时间意义，如LSTM在自然语言处理任务上的应用[9]。对输入数据加入时间编码，即在使用LSTM处理数据时，同时考虑热点话题发生的具体时间，以实现对高校舆情热点的精准预测。与支持向量机（support vector machine，SVM）和循环神经网络

（recurrent neural network，RNN）两种模型的预测结果进行对比发现，基于时间编码的LSTM在热度预测准确率上具有明显优势。

1 相关工作

1.1 舆情分析和预测

高校舆情经常引发全社会的广泛关注，而舆情的正确引导对于高校管理以及社会的稳定发展有着十分重要的意义。参考文献[10]针对高校在舆情管理和引导工作中遇到的挑战与问题，构建以“大数据”为支撑、新媒体为载体、机制创新为保障的“三位一体”的舆情管理和引导的工作模式，营造了良好的校园舆论生态环境。该参考文献考虑了高校舆情对学生意识形态管理的意义，并提出引导舆情向正确方向发展的策略，然而其在舆情发展趋势预测方面的考虑不足，导致难以有效地引导舆情的发展[11-13]。在网络舆情预测的研究方面，参考文献[14]针对区间犹豫模糊集在描述决策信息时会导致决策信息重要性程度降低这一问题，构建了一种基于概率区间犹豫模糊几何算子的多属性群决策模型，且通过网络舆情预测系统的选择实例验证了所提决策模型是可行和有效的。秦涛等人[15]提出一种基于排序学习的舆情事件演化趋势重要性评估算法。在模型训练过程中，充分利用标注数据中的专家知识以及有标签数据和无标签数据的关联关系，筛选出重要舆情事件并进行管控，提升了资源的利用效能。参考文献[16]以网络流文本为对象，通过分析网络话题内容焦点的迁移特性，提出了网络话题内容焦点的识别方法。上述方法由于模型训练未考虑舆情事件的动态变化性，预测准确率不高，还需要进一步增强模型的适应性。

1.2 舆情时序数据处理

由于舆情热度数据按照时间序列变化，刘定一等人[17]针对单一模型预测精度不高和社交媒体对舆情走势影响较大的问题，提出了融合微博热点分析和LSTM的舆情预测方法。然而特征集的数量较少，网络舆情谣言识别的准确率还有待提高。笃程成等人[18]利用深度循环神经网络对社交消息的传播过程进行建模，提出了SMOP模型。该模型由于优化目标单一，未考虑通过联合建模优化来进一步提升预测准确率。彭丹蕾等人[19]针对如何高效挖掘处理大量评论数据并进行情感分析的问题，采用SVM和LSTM分别对从京东网站爬取的商品评论进行建模。由于情感分析涉及的学科跨度比较大，并且采集的数据集比较单一，该模型适应性不强。为了有效监控和管理新型冠状病毒肺炎疫情引起的网络舆情，景楠等人[20]基于差分自回归移动平均（autoregressive integrated moving average，ARIMA）模型以及LSTM预测和分析舆情数据，对舆情模型进行参数估计、模型诊断和模型评价。由于未考虑各地区疫情发展的影响因素不同，该模型适应性不足。张陶等人[21]针对无属性社交网络的节点分类问题，提出了一种基于图嵌入与SVM的社交节点分类方法。由于采用静态的社交网络数据集进行模拟，该方法对动态社交网络的适应性不足，应用范围受到限制。针对方面情感，宋婷等人[22]提出基于方面情感分析的深度分层注意力网络模型，利用改进的LSTM获取句子内部和句子间的情感特征。由于未包含跨领域的词汇

和网络用语句子的方面情感分析，该模型的情感分类效果有待进一步提高。

根据以上分析，现有的时序数据处理模型存在算法预测精确度不够高、特征集和数据集比较单一的问题，并且很少结合舆情数据动态更新预测值。本文在对高校舆情数据进行处理时，利用关键词匹配全面考虑高校的相关信息，目的在于提高对高校舆情信息预测的准确率。结合时间编码方法，对舆情热度数据的绝对时间因素进行分析，可以解决LSTM处理时序数据时仅考虑数据先后关系的问题。同时利用实时舆情数据动态更新预测值，使得预测精确率进一步提升。本文提出基于时间编码LSTM的高校舆情热点趋势预测研究方法，动态调整评估参数。本文研究主要包括以下5个方面：一是获取微博热搜数据集合；二是通过降维、筛选、升维3种方法对数据集进行处理；三是将热点话题的时间编码加入数据集并进行归一化处理；四是生成训练集和测试集，利用训练集训练模型并生成预测模型，再利用测试集进行模型预测；五是对比分析预测值与真实值，最后评估各个模型的性能，验证了基于时间编码的LSTM在舆情热点时序数据处理方面的优越性。

2 模型介绍

2.1 LSTM模型

RNN在时序数据处理过程中会保留之前所有输入数据信息。一方面，随着后序数据的输入，先前的输入对模型隐含层的影响会越来越小，即长距离的依赖问题；另一方面，一些不重要的信息将被RNN保留。为了克服上述困难，LSTM被提出，该模型具有保持长期记忆性的特点，在时序数据处理方面具有良好的性能，LSTM结构如图1所示。LSTM模型的构建如下。

首先，对输入数据x_{i-1}和隐含状态h_{i-1}进行运算，得到LSTM的遗忘门，如式（1）所示：

$$C=\mathrm{sigmoid}(W_1(x_{i-1},h_{i-1})+b_1) \quad (1)$$

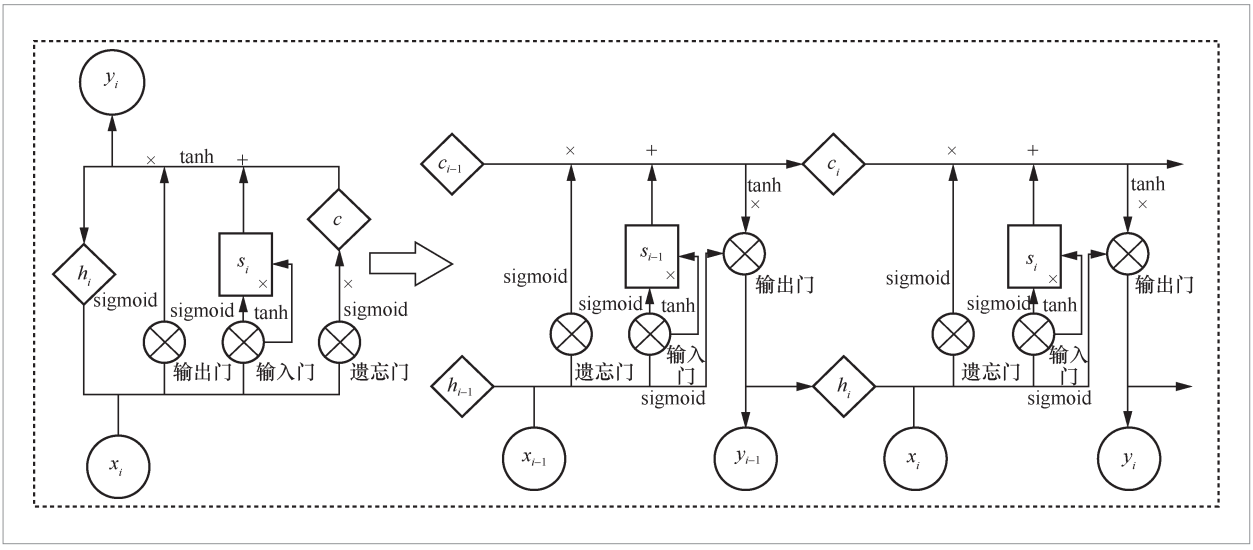

图1 LSTM 结构

在式（1）中，对输入数据x_{i-1}、隐含状态h_{i-1}与遗忘门的权重W_1进行线性运算，b_1表示引入的偏置项，再经过sigmoid激活函数引入非线性元素，此时$C \subset (0,1)$。C越大，记忆的部分越大。将C与当前的长期记忆状态c_{i-1}相乘并输出，即遗忘门的输出表示对长期记忆状态的记忆程度，如式（2）所示：

$$C' = \text{sigmoid}(W_1(x_{i-1}, h_{i-1}) + b_1) \times c_{i-1} \quad (2)$$

接下来计算LSTM的输入门部分，**图1**中s_i表示输入门的sigmoid神经网络层，符号×表示点乘运算操作，激活函数tanh将输入的新信息归一化到(−1,1)，通过点乘运算对信息进行缩放，决定保留哪些新信息，如式（3）、式（4）所示：

$$C_1'' = \text{sigmoid}(W_2(x_{i-1}, h_{i-1}) + b_2) \quad (3)$$

$$C_2'' = \tanh(W_3(x_{i-1}, h_{i-1}) + b_3) \quad (4)$$

其中，C_1''与C_2''均为临时变量，W与b分别表示可学习的权重与偏置。与遗忘门相似，$C_1'' \subset (0,1)$可被视为比例系数。在输入门中将输入的新信息经过tanh函数归一化表示为C_2''，$C_2'' \subset (-1,1)$，其可控制新增的量是正的或负的。经过点乘运算后，当前时刻输入门的网络输入表示为C''，如式（5）所示：

$$C'' = C_1'' \times C_2'' \quad (5)$$

通过上述过程，LSTM模型可以完成对已有长期记忆元素的更新，如式（6）所示：

$$c_i = C' + C'' \quad (6)$$

其中，c_i将被用于下一层LSTM的计算。与RNN相比，LSTM在输出时也进行了一定的改进。LSTM在输出时综合考虑了当前长期记忆和当前输入数据的影响，如式（7）所示：

$$y' = \text{sigmoid}(W_3(x_{i-1}, h_{i-1}) + b_3) \quad (7)$$

类比式（1），y'能够控制输出长期记忆的大小，$y' \subset (0,1)$。最终的输出如式（8）所示：

$$y_{i-1} = y' \times \tanh(c_i) \quad (8)$$

使用tanh函数激活当前长期记忆结果的值，得到LSTM的实际输出，y_{i-1}表示当前时刻的话题热度。

2.2 损失函数与优化

在训练模型时，需要将损失函数的值降至较低水平，以提高模型性能。损失函数是衡量神经网络性能的重要参考指标，通常损失函数在测试集上的结果越小，模型的性能越好。常用的损失函数有适用于回归问题的均方误差（mean square error，MSE）损失函数和适用于分类问题的交叉熵（cross entropy）损失函数等。对高校热点舆情话题的预测属于回归问题，因此将MSE作为损失函数，如式（9）所示：

$$\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}(\hat{y}_i - y_i)^2 \quad (9)$$

其中，$\hat{y}_i$表示预测值，y_i表示真实值。

时序数据通常具有隐含关系。通过神经网络的训练，挖掘数据的潜在特征，从而实现对数据的预测，即神经网络目标是一个回归任务。因此选择MSE对损失函数进行优化，提升模型性能。

2.3 评价指标

除了MSE，还可将平均绝对误差（mean absolute error，MAE）和平均绝对百分比误差（mean absolute percentage error，MAPE）作为模型的评价指标。

MAE表示预测值与真实值之间的误差平均绝对值，如式（10）所示：

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (10)$$

MAE能更好地反映预测值误差的实际情况（已经经过归一化）。模型测试数

据的MAE越大，预测误差越大。MAPE表示预测值与真实值的平均差距百分比，如式（11）所示：

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \qquad (11)$$

值得注意的是，一个较好的评估模型的MAPE值应该为0，其表示预测值与真实值之间没有差别。

3 实验与分析

3.1 数据集的获取

本研究使用的数据集来自新浪微博热搜榜，由于微博是一个实时信息交流分享平台，一旦舆情信息出现，将在短时间内迅速传播，因此分析微博热搜数据的热度变化趋势具有重要的研究意义。

具体的热度数据采集过程如下，首先通过Python框架对页面内容进行解析，并定义需爬取的字段，接着从某个时刻开始，每间隔15 min对热搜榜数据进行一次爬取，例如从0:00起，采集0:00、0:15、0:30等时间点的数据。如果热点话题仍在热搜榜上，就继续采集并添加热度值；否则，将新上榜的热搜数据添加至表中，最后保存所收集的数据，用于后续实验分析。热搜榜每次显示50个热点话题，按照其搜索热度进行排名。

由于研究对象是包含时间序列关系的话题名和热度值，因此将收集到的数据按照时间变化存储，对于某时刻不在热搜榜的话题，其在该时刻的热度为空。最终用于实验仿真的数据包含排名、关键词、热度、热度标识、时间5个维度，共15 000余条数据。整理收集到的数据，部分热搜数据示例见**表1**。

3.2 数据集的处理

3.2.1 降维

本文针对话题热度值进行预测，在获得原始数据集后，首先对数据集进行降维处理，删减与本文研究内容无关的维度，去掉冗余变量，这有助于提高算法的准确率。一方面，微博热搜榜的“爆”“沸”“热”等热度标识来源于实时热度值高低，热度标识与热度值意义重复，关系冗余，因此将数据集的热度标识维度删除。另一方面，微博话题热度序号只显示热度排名前50的话题，而热搜榜在实时变化，某一时刻的热度排名只与该时刻的话题热度有关，在对热度进行预测时，该时刻的相对排名对于研

表1 新浪微博部分热搜数据示例

排名	关键词	热度	热度标识	时间
置顶位	全国城市20年节水近千立方米	—	热	2021/5/11 15:45
1	张柏芝想为结婚对象再生小孩	3 229 372	沸	2021/5/11 15:45
2	成都警方回应四十九中学生坠亡	1 745 869	新	2021/5/11 15:45
3	吴磊范丞丞不敢看钟楚曦	1 639 683	新	2021/5/11 15:45
4	校方否认学生坠亡与化学老师有关	1 412 405	沸	2021/5/11 15:45
5	我国仍是世界第一人口大国	1 062 243	—	2021/5/11 15:45
6	洱海坠机牺牲机组人员信息	1 024 249	沸	2021/5/11 15:45
7	技嘉道歉	965 581	沸	2021/5/11 15:45

究意义不大，因此将热度排名维度删除。通过对数据集的降维处理，可以节约高校舆情热点趋势预测方法的训练时间。降维后的热搜数据示例见**表2**。

3.2.2 升维

在收集舆情数据时，按照15 min的间隔进行收集。热点话题在热搜榜不断地出现或消失，在存储某一时刻的热搜榜数据时，下一时刻该话题的热度数据可能消失，新的话题可能出现，因此在数据的整合和存储方面，需要考虑热搜话题变化带来的数据维度不一致问题。

针对上述问题，对原数据集进行升维操作，在原始数据基础上增加时间序列维度，按时间顺序记录每一数据爬取时刻的热度值。若话题热度不够高，未进入热搜榜或热度已经下降并离开热搜榜，则该时刻的热度值为空。升维后的热搜数据示例见**表3**。

3.2.3 筛选

微博热搜榜的话题包含娱乐、体育、民生、时政等多个话题类型，本文针对高校舆情类话题进行研究，因此需要对热搜话题进行筛选。针对与高校舆情相关的话题，通过关键词方式进行筛选。在获得的数据集中，将“高校”“大学”“学院”等与高校相关的话题关键词表示为集合

表 2 降维后的热搜数据示例

关键词	热度	时间
全国城市20年节水近千立方米	—	2021/5/11 15:45
张柏芝想为结婚对象再生小孩	3 229 372	2021/5/11 15:45
成都警方回应四十九中学生坠亡	1 745 869	2021/5/11 15:45
吴磊范丞丞不敢看钟楚曦	1 639 683	2021/5/11 15:45
校方否认学生坠亡与化学老师有关	1 412 405	2021/5/11 15:45
我国仍是世界第一人口大国	1 062 243	2021/5/11 15:45
洱海坠机牺牲机组人员信息	1 024 249	2021/5/11 15:45
技嘉道歉	965 581	2021/5/11 15:45

表 3 升维后的热搜数据示例

关键词	0:00	0:15	0:30	0:45	1:00	1:15	1:30	1:45	2:00	2:15	2:30
彭昱畅钉钉子钉了个寂寞	—	273 523	1 102 081	1 019 060	820 537	637 017	463 160	363 044	280 718	242 421	200 413
张艺兴说我现在太封闭了	—	—	664 143	1 712 153	1 737 095	1 652 408	1 347 315	1 072 099	855 029	714 722	587 527
黄旭熙 酒吧	—	—	—	354 148	350 332	364 802	291 973	194 279	170 444	140 730	124 714
首批全国禁毒示范城市名单公布	—	—	—	260 183	279 053	381 825	314 804	262 217	211 792	210 833	171 351
龚俊黑色衬衫	—	—	—	164 698	146 597	122 397	117 979	105 316	88 146	73 791	61 520
职场中要不要做老好人	—	—	—	—	146 471	175 040	202 605	190 093	161 795	132 898	106 573

$K=\{k_1,k_2,k_3,\cdots\}$，若话题与集合无交集，则为无关话题，对无关的话题进行忽略处理。对舆情话题进行分词处理，分词前后的话题见**表4**。

接着将分词后的舆情话题与高校关键词集合K进行匹配与筛选，保留与高校舆情相关的热点话题以及热度值变化情况，去除与高校舆情无关的数据信息，筛选后的高校热点数据见**表5**。

3.3 时间编码与归一化

在对数据进行归一化处理之前，首先加入时间编码，具体过程是对收集数据的每个时刻进行编码，例如从0:00开始收集数据，0:15的时间编码参数是0.25，0:30的时间编码参数是0.5，0:45的时间编码参数为0.75……具体的时间编码参数设置过程是将每小时分为4个部分，每部分占比为25%，若当前时刻为H时M分，编码参数设置为$\left(H+\frac{M}{60}\right)\bmod 24$。对高校舆情数据进行时间编码的优势是，若热度持续时间大于或等于24 h，此时时间编码大于或等于24，规定从0进行编码。在时间编码后，舆情数据之间的前后关系由于含有绝对时间的编码参数，舆情热度会随着时间发生变化，因此进一步结合不同时间段的舆情数据进行分析，可以提高舆情热度预测的准确率。接着对数据进行归一化操作，对以

表 4 分词前后的话题

分词前	分词后
盲人姐妹花双双考取名校研究生	盲人 姐妹花 双双 考取 名校 研究生
袁弘儿子在国际不打小孩日捣蛋	袁弘 儿子 在 国际 不 打 小孩 日 捣蛋
金海心 有没有一首歌会让你想起我	金海心 有没有 一首歌 会 让 你 想起 我
南昌大学办学100周年文艺晚会	南昌大学 办学 100 周年 文艺晚会
有酒店浴室挂钩暗藏摄像头	有 酒店 浴室 挂钩 暗藏 摄像头
残联回应残疾女硕士未通过教资认定	残联 回应 残疾女 硕士 未 通过 教资 认定

表 5 筛选后的高校热点数据

关键词	0:00	0:15	0:30	0:45	1:00	1:15	1:30	1:45	2:00	2:15	2:30
盲人姐妹花双双考取名校研究生	194 118	273 696	357 995	381 243	329 089	262 359	203 493	—	—	—	—
河南一高校设立失恋博物馆	—	188 306	222 776	310 417	303 227	274 726	258 909	240 363	—	—	—
背母上学的他毕业后重回大山教书	—	—	—	225 036	217 219	421 282	355 193	345 712	412 703	550 195	256 125
全国超2.18亿人具有大学文化程度	—	—	—	—	277 599	607 318	1 517 899	1 786 493	1 749 609	1 657 169	1 580 934
残联回应残疾女硕士未通过教资认定	—	—	—	—	—	—	—	178 871	269 884	651 628	401 901
南昌大学办学100周年文艺晚会	—	—	—	—	—	—	—	—	—	—	295 304

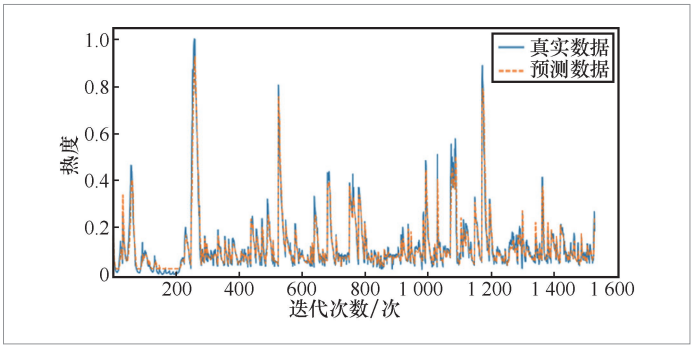

图 2 RNN 训练集预测效果

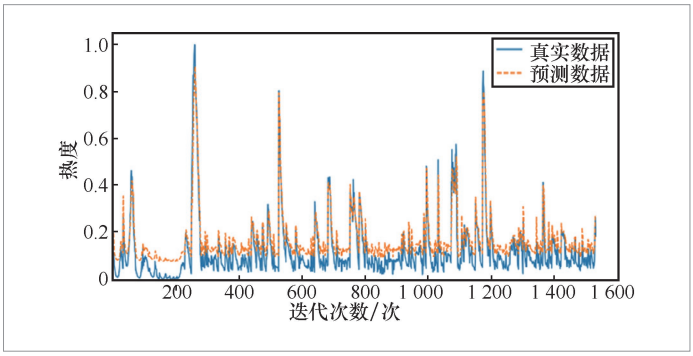

图 3 SVM 训练集预测效果

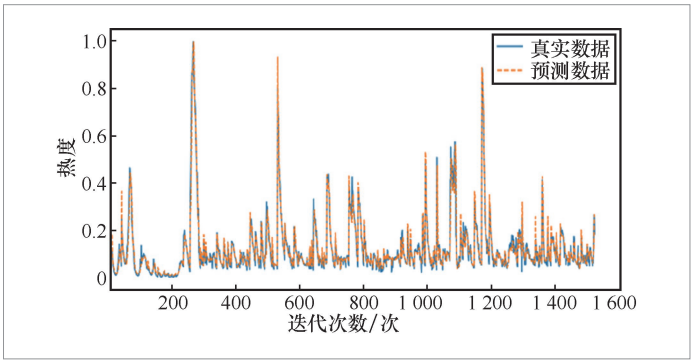

图 4 LSTM 训练集预测效果

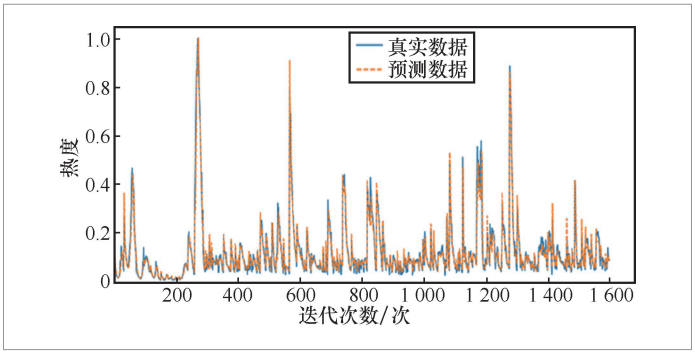

图 5 基于时间编码的 LSTM 训练集预测效果

十万甚至百万为单位的热度数据进行归一化处理，能加速模型收敛，提高模型精度。热度数据归一化可以对每个热点话题的时间变化序列数据进行归一化，也可以针对所有热度值进行归一化。考虑到每个话题的热度变化范围不同，某些话题的峰值热度可能仍低于其他话题的中等热度，导致归一化后的相对热度表示误差较大，因此采用整体归一化的思路进行处理。数据归一化表示为：

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{12}$$

在式（12）中，x表示某一时刻的热度值，$x_{\min}$表示该话题的最小热度值，$x_{\max}$表示该话题的最大热度值。通过热度值的归一化，有效地减小了热度值范围跨度。在后续神经网络训练及预测时，数据不会因为微小扰动而产生巨大误差，因此数据拟合与损失函数的收敛速度将进一步提高。

3.4 RNN、SVM、LSTM和基于时间编码的LSTM的舆情热度预测

首先将数据集分为训练集和测试集两类，将数据集的70%作为训练集，30%作为测试集。训练时设置学习批次大小为128，使用随机梯度下降法对模型进行优化，学习率设置为0.01，损失函数使用MSE。为了进一步对比验证时间编码的优势，与不含时间编码的LSTM进行对比。

实验一共进行100轮训练，使用梯度下降法反向传播误差，更新隐含层权重W'。

经过100轮训练，损失函数已经趋于零，说明模型性能基本达到最优。RNN、SVM、LSTM和基于时间编码的LSTM在训练集上的预测效果分别如**图2**、**图3**、**图4**、**图5**所示。其中，SVM使用高斯核函数，该核函数的参数gamma设置为0.1。

从图2~图5可以看出，基于时间编码的LSTM的预测性能略优于普通的LSTM，同时预测结果也比SVM、RNN更加准确，原因在于基于时间编码的LSTM不仅对时间序列数据保持长期的记忆性，而且具有更新数据信息的能力。同时，加入时间编码后，LSTM在舆情热度值与绝对时间之间建立了相应的联系，因此其在舆情趋势预测方面具有较好的性能。

接着使用100轮训练后LSTM、RNN、SVM和基于时间编码的LSTM模型，在测试集上进行预测，依次对每个话题的数据集进行测试，预测数据与真实数据的误差在较低水平。RNN、SVM、LSTM和基于时间编码的LSTM在测试集上的预测效果分别如图6、图7、图8、图9所示。

对比4种模型在测试集上的预测效果，基于时间编码的LSTM性能最优。SVM在热度较低时的预测结果偏高，在热度值最高点的预测结果偏低。由于热度变化是有规律的，可以根据前序数据得到后序数据，而SVM没有考虑前序数据的变化特征，导致其回归精度不够高。出现高校舆情后，通过基于时间编码的LSTM对舆情热度趋势进行预测，及时引导舆论发展的方向，将有利于高校对学生思想健康的管理，提升高校处理舆情事件的水平。

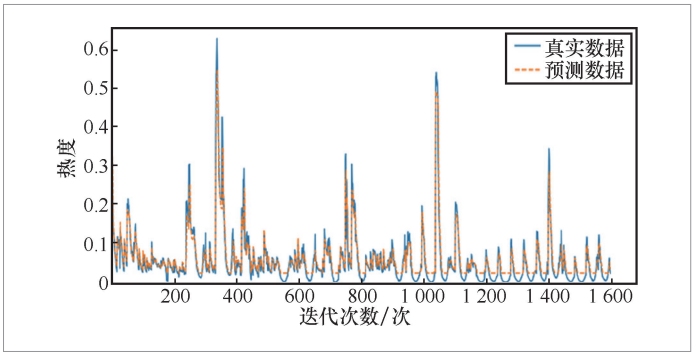

图 6 RNN 测试集预测效果

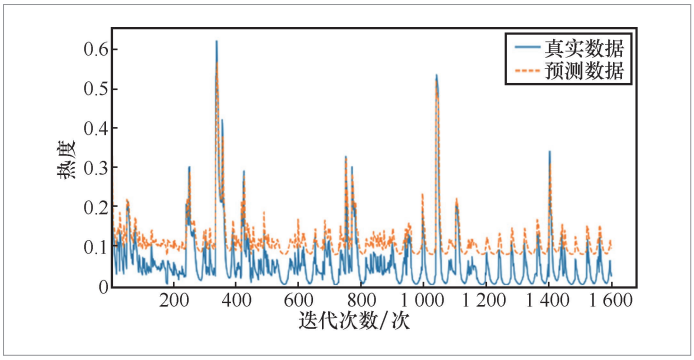

图 7 SVM 测试集预测效果

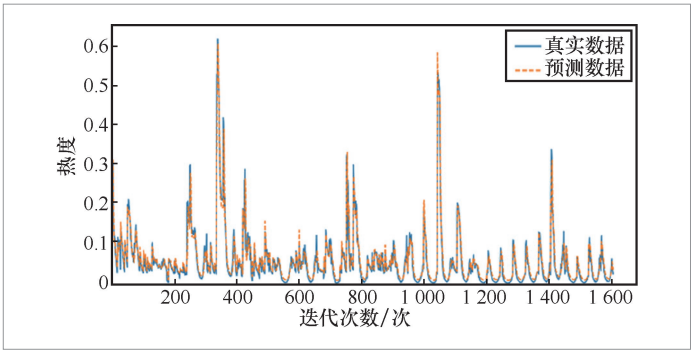

图 8 LSTM 测试集预测效果

4 模型评估

4.1 基于时间编码的LSTM模型真实集评估

以某真实事件为例，对新浪微博中该事件的真实热度数据每间隔15 min采集一次，并对收集到的数据进行整理保存。首

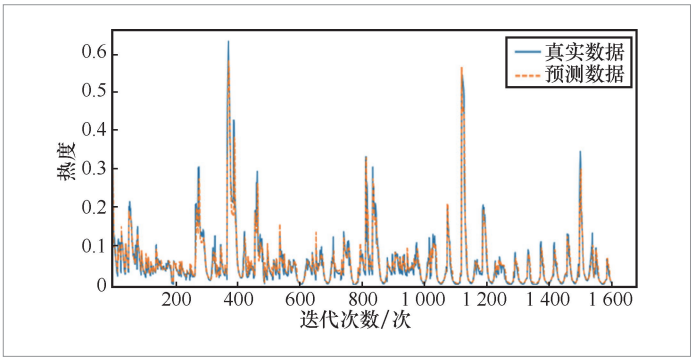

图 9 基于时间编码的 LSTM 测试集预测效果

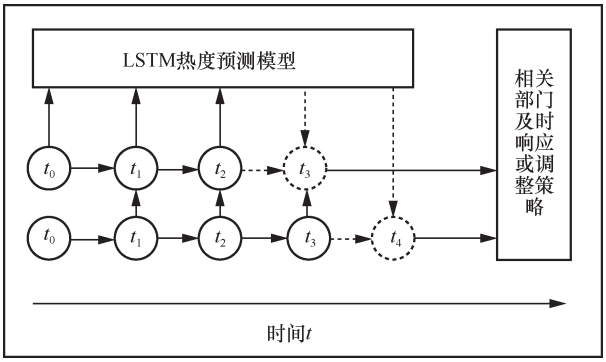

图 10 动态调整评估参数

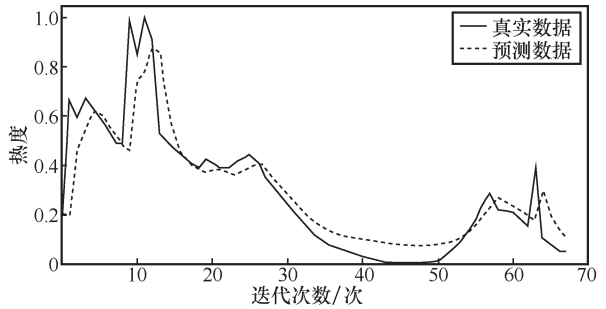

图 11 SVM 真实集预测效果

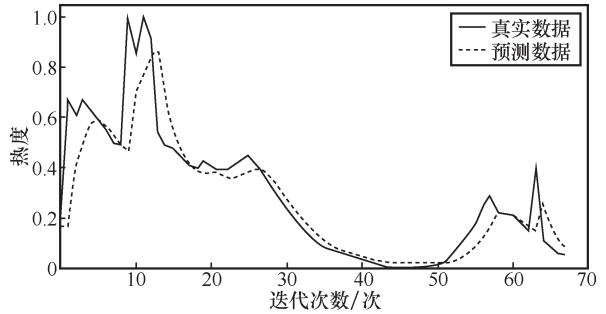

图 12 RNN 真实集预测效果

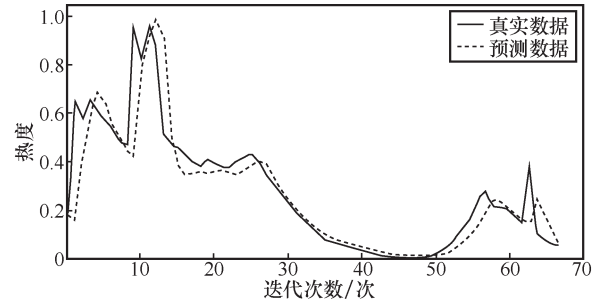

图 13 LSTM 真实集预测效果

先，对上述数据进行预处理归一化，并加入时间编码参数，将前45 min的数据作为神经网络的输入，预测得到下一时刻的输出。由于舆情数据受多种因素的影响，单独使用模型进行预测的效果不理想，故需要结合舆情实时的动态变化性对评估参数进行调整。在预测下一时刻的热度值时，可以根据舆情变化做出相应的处理，在获得真实数据后，结合真实热度数据进行预测，即进行动态的校正与下一步预测，动态调整过程如**图10**所示。

结合动态调整策略，对真实事件持续处于热搜榜的18.5 h（即74个时刻）进行预测，分别使用SVM、RNN、LSTM以及基于时间编码的LSTM进行预测，结果分别如**图11**、**图12**、**图13**、**图14**所示。分析实验结果可知，基于时间编码的LSTM模型能得到事件在具体时刻的热度，结合动态调整策略，其适应性得到提高。与其他3种算法相比，基于时间编码的LSTM的预测准确率是最高的。当高校舆情热点趋势即将进入爆发期时，相关部门及时响应或调整策略，对舆情热点发展趋势进行管控，有助于高校完善舆情管理体系。

随着时间推进，相关部门可根据预测结果，提前为舆情的发展做出判断和回应。然而，神经网络不能自动判断预测停止时间。通过实验和数据分析可以得出，当热度数据预测值低于最低话题热度值时，可认为话题热度低于热搜榜上榜要求，停止预测。

4.2 误差评估分析

对基于时间编码的LSTM、LSTM、RNN和SVM 4种模型在不同数据集上的MAPE和MAE进行对比，结果如**图15**所示。从**图15**可知，MAPE和MAE的数

值越小，预测值与真实值的误差越小，即预测结果越接近真实值。在4个模型中，基于时间编码的LSTM的预测效果是比较准确的。从MAPE对比实验结果分析：基于时间编码的LSTM在训练集和测试集上的预测效果明显优于其他3种模型。从MAE对比实验结果分析，基于时间编码的LSTM模型预测效果在训练集、测试集和真实集上明显优于RNN和SVM。但受到真实事件的动态变化性以及不确定因素的影响，基于时间编码的LSTM模型在部分预测集上的效果略差，后续研究需进一步提升模型的稳定性。综合比较，基于时间编码的LSTM还是具有明显优势的，在测试集上的预测效果优于其他模型。因此，使用基于时间编码的LSTM对高校舆情热点趋势进行预测具有较高的准确率，可以降低舆情带来的不利影响。

5 总结与展望

本文通过爬取新浪微博中高校的舆情热点数据，使用基于时间编码的LSTM学习舆情数据热度的时序变化情况，并对时序数据进行建模。将经过多轮训练和参数调优的基于时间编码LSTM的高校舆情热点趋势预测模型与RNN、SVM和LSTM 3种模型的预测结果进行对比分析，实验结果表明，基于时间编码的LSTM在训练集、测试集、真实集上的预测结果误差较小，具有良好的实时预测效果。本文可为相关部门预测热点事件的舆情趋势变化提供一定的参考，从而及时做出相应的决策。未来研究将从热点问题的内容与评论入手，进一步研究基于时间编码的LSTM模型的稳定性，建立更完善的舆情预测模型，挖掘更深层次的舆情趋势的发展规律。

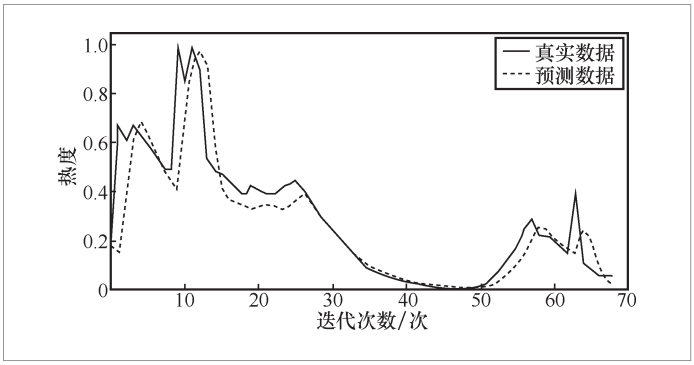

图 14 基于时间编码的 LSTM 真实集预测效果

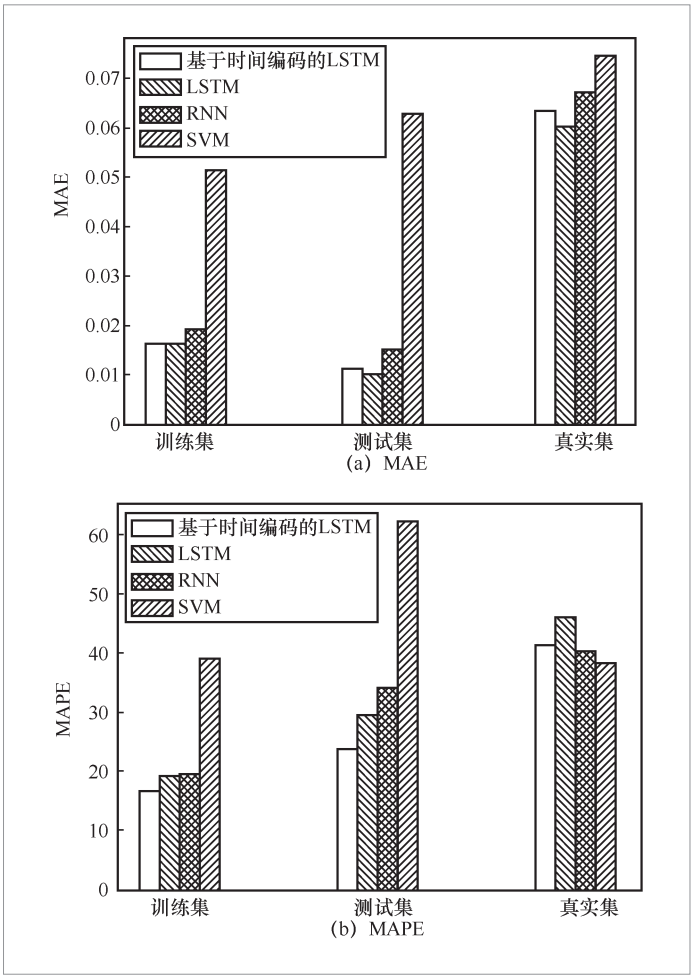

图 15 模型预测效果

参考文献

[1] 第47次《中国互联网络发展状况统计报告》发布[J]. 中国广播，2021(4)：38.

The 47th statistical report on Internet development in China released[J]. China Broadcasts, 2021(4): 38.

[2] 黄苏芬, 司雯, 穆亭钰. 自媒体时代高校网络舆情管控与引导机制创新研究[J]. 情报科学, 2021, 39(4): 62-67, 91.
HUANG S F, SI W, MU T Y. Innovation and research of online public opinion control and guidance mechanism in universities in the self-media era[J]. Information Science, 2021, 39(4): 62-67, 91.

[3] 赵妍妍, 秦兵, 刘挺. 社会焦点透视镜系统: 大数据视角下的舆情观测平台[J]. 大数据, 2016, 2(2): 46-55.
ZHAO Y Y, QIN B, LIU T. Social event sensor: a public opinion platform from the big data perspective[J]. Big Data Research, 2016, 2(2): 46-55.

[4] 罗文. 微传播时代高校网络舆情风险管理研究[J]. 新闻研究导刊, 2021, 12(5): 31-33.
LUO W. Research on the risk management of internet public opinion in universities in the micro-communication era[J]. Journal of News Research, 2021, 12(5): 31-33.

[5] 聂辉, 吕吉. 高校大学生突发性舆情事件应对机制与策略研究: 基于沉默螺旋理论的分析[J]. 江苏高教, 2021(2): 49-53.
NIE H, LYU J. Research on the coping mechanism and strategy for sudden public opinion events of college students[J]. Jiangsu Higher Education, 2021(2): 49-53.

[6] 陈淑娟, 徐雅斌. 面向主题社团的意见领袖挖掘方法[J]. 计算机工程与应用, 2021, 57(2): 156-163.
CHEN S J, XU Y B. Opinion leader mining method for theme community[J]. Computer Engineering and Applications, 2021, 57(2): 156-163.

[7] PENG L J, SHAO X G, HUANG W M. Research on the early-warning model of network public opinion of major emergencies[J]. IEEE Access, 9: 44162-44172.

[8] ZHANG B Y, ZHU X F, HUANG X Y, et al. A novel microblog sentiment classification method based on top-k pooling[C]// Proceedings of 2021 4th International Conference on Artificial Intelligence and Big Data. Piscataway: IEEE Press, 2021: 335-341.

[9] ZHANG H L, XU H B, SHI J Q, et al. Word level domain-diversity attention based LSTM model for sentiment classification[C]//Proceedings of 2020 IEEE 5th International Conference on Data Science in Cyberspace. Piscataway: IEEE Press, 2020: 164-170.

[10] 周洋洋. 网络强国战略下高校网络舆情管理与引导[J]. 网络安全技术与应用, 2021(7): 115-117.
ZHOU Y Y. Management and guidance of network public opinion in colleges and universities under the strategy of network powering the country[J]. Network Security Technology & Application, 2021(7): 115-117.

[11] YIN W D. Public opinion prediction based on Markov model[C]//Proceedings of 2021 6th International Conference on Communication and Electronics Systems. Piscataway: IEEE Press, 2021: 218-221.

[12] YU N, LIU K, MA K. Analysis of public opinion heat trend in universities on the basis of Markov chain[C]//Proceedings of 2018 IEEE 15th International Conference on e-Business Engineering. Piscataway: IEEE Press, 2018: 218-222.

[13] JIAO H, MA Y H. Prediction of Weibo event dissemination attention based on Markov model[C]//Proceedings of 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering. Piscataway: IEEE Press, 2021: 611-615.

[14] 周小领, 马庆功. 概率犹豫模糊算法及其网络舆情预测模型选择[J]. 计算机工程与应用, 2019, 55(4): 179-184, 192.
ZHOU X L, MA Q G. Probabilistic hesitant fuzzy algorithm and its application for selection method of network public opinion prediction model[J]. Computer Engineering and Applications, 2019, 55(4): 179-184, 192.

[15] 秦涛, 沈壮, 刘欢, 等. 基于排序学习的网络舆情演化趋势评估方法研究[J]. 计算机研究

与发展, 2020, 57(12): 2490-2500.
QIN T, SHEN Z, LIU H, et al. Learning to rank for evolution trend evaluation of online public opinion events[J]. Journal of Computer Research and Development, 2020, 57(12): 2490-2500.

[16] 周亚东, 刘晓明, 杜友田, 等. 一种网络话题的内容焦点迁移识别方法[J]. 计算机学报, 2015, 38(2): 261-271.
ZHOU Y D, LIU X M, DU Y T, et al. A method for identifying the evolutionary focuses of online social topics[J]. Chinese Journal of Computers, 2015, 38(2): 261-271.

[17] 刘定一, 沈阳阳, 詹天明, 等. 融合微博热点分析和LSTM模型的网络舆情预测方法[J]. 江苏大学学报(自然科学版), 2021, 42(5): 546-553.
LIU D Y, SHEN Y Y, ZHAN T M, et al. Network public opinion forecasting method fusing microblog hotspot analysis and LSTM model[J]. Journal of Jiangsu University (Natural Science Edition), 2021, 42(5): 546-553.

[18] 笱程成, 秦宇君, 田甜, 等. 一种基于RNN的社交消息爆发预测模型[J]. 软件学报, 2017, 28(11): 3030-3042.
GOU C C, QIN Y J, TIAN T, et al. Social messages outbreak prediction model based on recurrent neural network[J]. Journal of Software, 2017, 28(11): 3030-3042.

[19] 彭丹蕾, 谷利泽, 孙斌. 基于SVM和LSTM两种模型的商品评论情感分析研究[J]. 软件, 2019, 40(1): 41-45.
PENG D L, GULI Z, SUN B. Sentiment analysis of Chinese product reviews based on models of SVM and LSTM[J]. Computer Engineering & Software, 2019, 40(1): 41-45.

[20] 景楠, 胡怡, 韩喜双. 基于ARIMA与LSTM的新冠肺炎网络关注度趋势研究[J]. 中国安全科学学报, 2020, 30(12): 37-42.
JING N, HU Y, HAN X S. Trend of COVID-19 network attention based on ARIMA and LSTM[J]. China Safety Science Journal, 2020, 30(12): 37-42.

[21] 张陶, 于炯, 廖彬, 等. 基于图嵌入与支持向量机的社交网络节点分类方法[J]. 计算机应用研究, 2021, 38(9): 2646-2650, 2661.
ZHANG T, YU J, LIAO B, et al. Node classification method in social network based on graph embedding and support vector machine[J]. Application Research of Computers, 2021, 38(9): 2646-2650, 2661.

[22] 宋婷, 陈战伟, 杨海峰. 基于分层注意力网络的方面情感分析[J]. 大数据, 2020, 6(5): 82-91.
SONG T, CHEN Z W, YANG H F. Aspect sentiment analysis based on a hierarchical attention network[J]. Big Data Research, 2020, 6(5): 82-91.

作者简介

易杰(1998-), 男, 青海大学计算机技术与应用系硕士生, 主要研究方向为深度学习、数据挖掘。

曹腾飞(1987-), 男, 博士, 青海大学计算机技术与应用系副教授, 中国计算机学会会员, 主要研究方向为舆情分析、服务推荐与管理。

黄明峰（1977- ），男，清华大学创新领军工程博士生，云上贵州大数据产业发展有限公司研究员级高级工程师，主要研究方向数据治理、大数据应用、数据可信流通。

黄肖翰（1999- ），男，青海大学计算机技术与应用系硕士生，主要研究方向为深度学习、数据挖掘。

张子震（1994- ），男，青海大学计算机技术与应用系硕士生，主要研究方向为强化学习、舆情分析。

收稿日期： 2021-11-02

通信作者： 曹腾飞，caotf@qhu.edu.cn

基金项目： 国家自然科学基金资助项目（No.62101299）；青海省自然科学基金资助项目（No.2020-ZJ-943Q）；青海大学党建与思想政治教育研究项目（No.szzx2012）

Foundation Items: The National Natural Science Foundation of China (No.62101299), The Natural Science Foundation of Qinghai Province (No.2020-ZJ-943Q), Qinghai University Party Building and Ideological and Political Education Research Project (No.szzx2012)