

基于混合深度学习方法的地网络舆情分析

李帅文, 刘 继

(新疆财经大学统计与数据科学学院, 新疆 乌鲁木齐 830012)

✉1691133906@qq.com; Liuji5000@126.com



摘 要: 为了研究地网评论情感对地网价格波动的影响及其造成的市场现象, 提出了基于CNN(卷积神经网络)和BiLSTM(双向长短期记忆网络)混合方法的地网络舆情分析模型。构建地域专属情感词典, 通过CNN和BiLSTM获得地网情感极性, 利用量化情感指数计算得到情感得分, 将其与百度指数拟合计算互信息数。结果表明: 经过量化的地网情感值为67.84%与百度指数的相关系数值大致相同, 二者走势基本吻合, 利用该方法可实时测度地网评论舆情走势, 为数字经济背景下地网企业决策者提供智力支持。

关键词: 地网; 网络舆情; 情感指数; 最大互信息

中图分类号: TP183 **文献标识码:** A

Public Opinion Analysis of Real Estate Network based on Hybrid Deep Learning Method

LI Shuaiwen, LIU Ji

(School of Statistics and Data Science, Xinjiang University of Finance and Economics, Urumchi 830012, China)

✉1691133906@qq.com; Liuji5000@126.com

Abstract: In order to study the influence of real estate text comment sentiment on real estate price fluctuation and its market phenomenon, this paper proposes a real estate network public opinion analysis model based on the hybrid method of CNN (Convolutional Neural Network) and BiLSTM (Bi-directional Long Short-Term Memory). A region-specific sentiment dictionary is constructed and the text sentiment polarity is obtained through CNN and BiLSTM. The sentiment score is calculated by using the quantitative sentiment index, and the number of mutual information is calculated by fitting it to the Baidu index. The results show that the quantified emotional value of real estate text is 67.84%, which is roughly the same as the correlation coefficient value of Baidu Index. The trend of the two is basically consistent. This method can be used to measure the trend of public opinion of real estate reviews in real time, and provide intellectual support for decision-makers of real estate enterprises in the context of digital economy.

Keywords: real estate; network public opinion; sentiment index; maximum mutual information

1 引言(Introduction)

随着互联网的发展, 大量的网评论信息走进研究者的视野, 成为反映地网关注者情感表达的重要方式。现有的研究将网评论的情感极性(积极或消极)作为反映市场状态的指标, 而本文通过研究发现, 网评论所蕴含的情绪得分不能直接用于衡量百度指数的变化, 此外网评论信息属于非结构化的数据, 具备非线性、非平稳性的特征^[1], 所以以往的

量化投资方法应用于地网价格研究并不适用。

近年来, 越来越多的研究者尝试使用深度学习解决基于时序数据的地网波动问题。但是, 每天会产生很多的地网评论信息, 利用深度学习只能得到单个网的情感极性, 因此单一使用深度学习得到网情感用于衡量百度地网指数走势是有一定缺陷的。

针对上述问题, 本文提出在深度学习模型的基础上构建

情感指数,通过情感指数将深度学习得到的文本情感进行每日量化,从而估计其对于百度地产指数的影响。

2 相关研究工作(Related research work)

现有的对房地产市场舆情的研究可以分为两类:一类是基于数值分析;另一类是基于传统数学方法从多角度解读影响房地产价格波动的因素。

为了能直接利用数值、数据,尽可能地了解数据背后的规则,张爱琳等^[2]收集包头市2015—2019年的年平均房价数据,用其构建预测房地产价格波动的GM(1,1)模型,从而预测包头市房地产未来3年的价格变化趋势。刘洋等^[3]发现应用几何布朗运动模型研究房地产价格变化有一定的优势。几何布朗运动模型仅单一考虑房地产价格数值信息,未考虑影响房地产价格变化的直接因素和间接因素。柳冬等^[4]在进行房地产价格预测时,首先对影响房地产价格的因素进行分析,然后选取房地产行业的热点问题,利用多种方法分析我国房地产市场价格的变化趋势。赵怡爽^[5]通过多元因子分析,从12个影响指标中挑选出重要性排名前两位的影响指标,运用层次分析方法将上述指标进行加权赋值,得到最终综合分数,进而预测未来房地产价格变动。邵为爽等^[6]利用数据挖掘理论,将属性约简算法使用到房地产价格影响因素提取上,从而将降维后的数据放入网络训练并预测房地产价格走势。这些方法都是基于影响房地产价格的因素,利用经济模型对房地产价格进行预测。多元因子分析和层次分析方法皆在基础模型上加入了其他影响因素。

随着对影响因素的研究逐渐深入,研究者发现文本情感信息对房地产价格的预测有非常重要的影响。因此,融合文本信息的房地产价格预测模型应运而生。

何平等^[7]通过构建投资者情绪指数,提取股民文本评论情绪用于研究股民的情感变化是否会影响到股票市场价格波动。姜富伟等^[8]通过专属情感词典计算媒体文本情绪指数,发现媒体文本情绪指数具有代表性作用,可以作为预测股票价格变动的参考依据。由上述研究可知,文本情绪表达会间接干预经济变化。由于传统情感分析方法在文本情感分类任务中耗时多且效率较低,缪亚林等^[9]利用CNN和双向GRU(门控循环单元)提取文本内在特征信息,然后通过单层GRU降维,使用激活函数进行情感分类。BEHERA等^[10]提出将循环神经网络(LSTM)与CNN结合,发现CNN-BiLSTM混合模型兼具LSTM和CNN的优点,可以使文本分类任务预测效果更好。上述研究验证了深度学习在文本情感计算中的优越性,随后研究人员提出深度学习文本情感分类和金融预测联合模型。例如, JIN等^[11]提出了一种基于LSTM的股票市场预测模型,该模型考虑了投资者的情绪倾向。

这些现有的基于房地产的研究和利用文本信息预测股票走势及提供股票投资建议,让我们有了基于深度学习提取文本情感,将其运用于地产行业的想法。

3 投资者情感指数与地产走势相关性分析模型(Correlation analysis model of investor sentiment index and real estate trend)

为了高效提取特征,提高预测精度,本文将CNN和BiLSTM网络融入一个统一的框架之中,提出了一个名为CNN-BiLSTM的情感分析模型。该模型可以充分利用文本信息,自动学习和提取其内存特征。CNN-BiLSTM网络结构模型如图1所示。

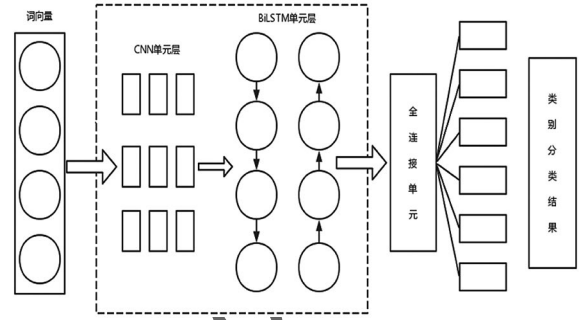


图1 CNN-BiLSTM网络模型结构图

Fig.1 Diagram of CNN-BiLSTM network model structure

在此模型中,首先通过Word2Vec(词向量)模型训练好词向量,其次使用CNN提取输入文本的局部特征,再次利用双向循环网络提取上下文总体文本特征,在通过Softmax分类得到情感极性后,利用自定义情绪指数计算每天的情感得分,最后将其与地产百度指数做相关性分析。

3.1 卷积神经网络(CNN)

CNN模型主要是由卷积层和池化层构成,卷积之前需要利用文本输入层提取文本特征,最后通过全连接层得到文本分类结果。在文本分类任务中,将文本训练得到的词向量传递给输入层,然后通过卷积层提取词语内部特征。卷积核的选取可以根据研究者的需求进行设置,然后通过池化进行运算,对文本特征做进一步处理,提取主要特征信息,将获得的信息全部传输到全连接层,通过激活函数得到文本分类结果^[12]。

设输入词向量为 X ,卷积神经网络使用卷积核提取文本信息,从而更好地获得字词之间的语义内涵,其运算过程如下:

$$C = f(X \otimes W + b) \quad (1)$$

式(1)中, $\otimes$ 表示卷积运算, W 和 b 表示权重和偏置, $f(\cdot)$ 表示激活函数。把提取到的信息放入池化层,通过运算得到主要信息,运算过程如下:

$$\tilde{C} = \max(C) \quad (2)$$

最后将结果进行全连接运算。

3.2 双向前长短期记忆网络(BiLSTM)

循环神经网络是将获得信息经过多次训练的网络模型,LSTM是基于循环神经网络模型的优化模型,其优点在于不会出现梯度消失和爆炸问题。运用LSTM模型的第一步是通过遗忘门选择上一时间有多少文本信息可以进入当前时刻,该步骤输入为 h_{t-1} 和 x_t ,输出为 f_t ;第二步是使用激活函数(Sigmoid)

构成的输入门控制信息更新,此步骤输入为 h_{t-1} 和 x_t ,经过激活函数后输出 a_t 、 w_t 和状态信息;第三步是输出门控制状态信息的输出,以 h_{t-1} 、 x_t 和上一时刻的状态信息为输入,在运算之后得到最终结果。模型在 t 时刻的运算过程如下:

$$f_t = \text{Sigmoid}(W_f \cdot \text{cat}[h_{t-1}, x_t] + b_f) \quad (3)$$

$$a_t = \text{Sigmoid}(W_a \cdot \text{cat}[h_{t-1}, x_t] + b_a) \quad (4)$$

$$w_t = \tanh(W_c \cdot \text{cat}[h_{t-1}, x_t] + b_c) \quad (5)$$

$$c_t = f_t \cdot c_{t-1} + a_t \cdot w_t \quad (6)$$

$$\text{Out}_t = \text{Sigmoid}(W_{\text{out}} \cdot \text{cat}[h_{t-1}, x_t] + b_{\text{out}}) \quad (7)$$

$$H_t = \text{Out}_t \cdot \tanh(c_t) \quad (8)$$

式(3)中, $\text{Sigmoid}(\cdot)$ 表示激活函数, W_f 和 b_f 分别表示 t 时刻的权重和偏置。式(7)和式(8)中, Out_t 表示 t 时刻输出的内容, H_t 表示 t 时刻输出的隐藏状态内容, cat 表示向量之间具有拼接运算。

LSTM模型在处理梯度消失以及梯度爆炸问题时有较好的测试效果,计算时间也进一步减少,但是模型只考虑了文本单向信息,未考虑文本的双向信息。在针对文本情感分析问题上,如果在只考虑句子前向信息的基础上,将文本后向信息也加入进去,就可以更好地让模型获取文本内在规则,进而提高对句子的情感极性预测效果。在多分类文本任务中, BiLSTM模型可以充分提取文本信息,让分类任务具有更好的效果。

BiLSTM模型是由正向LSTM模型和反向LSTM模型构成,每层LSTM网络分别对应输出的一个隐藏状态信息,模型参数由反向传播进行更新。BiLSTM模型的结构如图2所示。

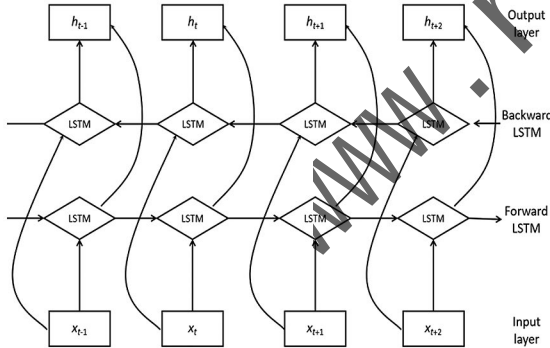


图2 BiLSTM模型的结构

Fig.2 Structure of the BiLSTM model

3.3 情感指数构建

由于每天产生的地产评论信息较多,所以需要构建情感指数衡量每日综合情感。情感指数构建步骤如下。

第一步,根据CNN-BiLSTM情感分类器,情感得分大于0.6的文本表示具有积极情绪,情感标签为1;情感得分小于0.4的文本表示具有消极情绪,情感标签为0。

第二步,设置变量 neg 和 pos ,用来计算某一天综合文本情感得分,以 neg 为例介绍两个变量的计算方法:某天的综合文本情感得分初始化为0,假设某天的某条评论情感极性为0, $neg = neg + 1 + \log(\text{本评论点赞数量} + \text{本评论回复数量} + 1, 2)$ 。

其中, $\log(x, 2)$ 表示以2为底的对数,情感指数构建添加了点赞和评论数量,是考虑了文本评论中点赞数和评论数对情感值的影响程度。取对数是为了让数据更加平滑,防止极值过大。+1是为了防止有的评论点赞和回复都为0。

第三步:计算某一日评论的总体情绪得分 $score$ 。模型计算公式如下:

$$score = \log(pos / (pos + neg + 0.0001) \times (\text{该日评论总数} + 1)) \quad (9)$$

式(9)中, $pos / (pos + neg + 0.0001)$ 表示的是计算市场的情绪极性, (该日评论总数+1)表示的是考虑评论数量越多市场情绪变化越大。

3.4 最大信息系数(MIC)

最大信息系数是用于寻找数据集中具有潜在相关关系的变量对,是用来确定两个变量相关度的一种标准,具有广泛性和公平性。例如, MIC 获得的有序数对 (X, Y) 对应的数据集是 D ,数据样本大小为 n ,分别将 X 和 Y 按照 x 值和 y 值划分成 x 和 y 维,这种划分称之为 $x \times y$ 网格划分,记为 G 。

数据集 D 中分别有两个节点变量 X 和 Y 。 X 和 Y 的特征矩阵是一个无限矩阵,给定信息系数公式如下:

$$M(X, Y|D)_{i,j} = \frac{I^*(X, Y, D, i, j)}{\log \min(i, j)} \quad (10)$$

式(10)中, $I^*(X, Y, D, i, j) = \max I(X, Y, D|_G, i, j)$ 是 G 被划分成 $i \times j$ 个网格中的最大互信息。

4 实证分析(Empirical analysis)

4.1 实验过程

房地产的价格走势是一个二分类问题,市场参与主体如果持积极客观的态度,看好市场未来发展,那么就可以判定房地产价格走势极大概率会上升;反之则判定为下降。

本文实验主要由三个部分构成,过程示意图如图3所示。

步骤一,分别爬取百度指数和房地产评论数据,对百度指数和房地产评论数据进行预处理,得到百度指数和房地产评论数据库。

步骤二,对评论数据进行人工标注后,利用混合CNN-BiLSTM建立评论情感分类模型。

步骤三,将预测部分评论数据输入“步骤二”得到评论情感分值,利用自定义情绪指数计算每日最终情感得分,分析其与百度指数的相关性。

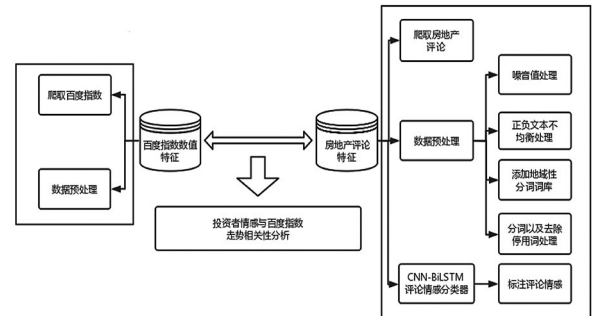


图3 实验流程图

Fig.3 Experimental flowchart

4.2 实验数据与预处理

本实验文本数据是相关房地产交易网页中2020年1月初至2021年11月末有关新疆房地产评论文本数据, 首先经过人工筛选剔除文本噪声, 其次对文本进行标签标注, 可以获得11 961条有效数据用来对模型进行训练和测试。在本实验中, 研究人员将2020年1月初至2021年5月末共计9 468条数据用来训练模型, 将2021年6月初至2021年11月末共计2 492条数据用来做预测分析。由于正负样本不均衡, 因此在此处采用下采样方法做特殊出口。在构建词向量阶段, 由于缺乏新疆地域性房地产词典, 所以对文本进行分词处理时, 将搜狗网房地产11个词库做合并处理, 构建基础分词词典, 在词基础上利用结巴分词, 将人工筛选出的108条词汇和根据TF-IDF(词频-逆文档频率)提取到的前2 000条词汇添加到分词词典, 从而完成新疆地区房地产词典的构建。

4.3 实验参数设置

本实验中, Keras(人工神经网络库)被用作神经网络的框架, 并使用Python编程语言实现网络结构。实验实现过程需要环境配置, 实验性能也需要预定义相关超参数, 环境配置与具体参数分别如表1和表2所示, 其中Null表示该参数不需要设置。

表1 模型实验环境设置

Tab.1 Model experiment environment settings

库名称	版本号
Python	3.7
tensorflow	1.14.0
keras	2.3.1
pandas	0.24.2
matplotlib	3.0.2
gensim	3.8.1
numpy	1.19.4
wordcloud	1.8.1

表2 模型超参数设置

Tab.2 Model hyperparameter settings

参数	CNN	BiLSTM
词向量维度	100	100
卷积核个数/个	64	Null
Padding	same	Null
Dropout	0.5	0.5
迭代次数/次	10	10
Batch_size	256	256
Activation	relu	tanh

4.4 实验评价指标

为了直观、定量地分析所提模型和情感指数性能, 本实验中采用准确率(Acc)、精准率(P)、召回率(R)和 $F1$ 四类评价指标测度分类模型。下面给出了4类指标对应的计算公式, 其中

Acc 衡量了模型分类准确性, R 代表召回率, P 代表精准率, F 是精确率和召回率的调和平均数, 当参数 $\alpha=1$ 时 F 就是 $F1$ 测度。

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

$$R = \frac{TP}{TP + FN} \quad (12)$$

$$P = \frac{TP}{TP + FP} \quad (13)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (14)$$

上式中: TP 表示预测的正向文本语料实际为正例; FP 表示预测的正向文本语料实际为负例; TN 表示预测的负向文本语料实际为负例; FN 表示预测的负向文本语料实际为正例。

4.5 情感分类器实验结果

为了验证CNN-BiLSTM模型在文本情感分类的有效性, 本文选择五个对比模型(SVM、LSTM、TextCNN、BiLSTM、BiLSTM+Attention)与本文所提模型进行测试效果对比, 其对比结果如表3所示。

表3 文本情感分类性能对比表

Tab.3 Comparison table of text sentiment classification performance

模型	Acc	P	R	$F1$
SVM	0.795 6	0.852 3	0.738 6	0.779 5
LSTM	0.886 1	0.892 5	0.878 6	0.884 6
TextCNN	0.843 9	0.853 6	0.845 6	0.856 3
BiLSTM	0.916 2	0.894 9	0.908 4	0.893 6
BiLSTM+Attention	0.901 2	0.935 7	0.812 9	0.825 6
CNN-BiLSTM	0.933 8	0.926 8	0.924 9	0.928 7

从表3中可以看出, 基于CNN-BiLSTM模型对房地产文本分类的准确率为93.38%, 基于BiLSTM和BiLSTM+Attention模型的性能次之, 基于SVM模型的性能最差。因此, 本文选择CNN-BiLSTM模型计算地产评论情感极性。

4.6 相关性分析

在CNN-BiLSTM模型得到情感极性后, 通过本文设计的情感指数求得地产评论每日的情感得分($score$)。首先为了验证本文提出指标是符合实际的, 则需要找出能够代表新疆乌鲁木齐房地产市场发展以及市场态度的指标测试指数。百度指数是由搜索引擎提出的, 是基于互联网大数据统计搜索量的一个指标。搜索量的数目可以体现投资者对市场的关注度, 从而反映市场情绪变化。因此, 选取百度指数作为检验指标, 接着将计算出的情感得分和百度地产指数通过MIC分析相关性, 得到MIC结果是23.08%, 相关性较弱, 绘制出的折线图如图4所示。

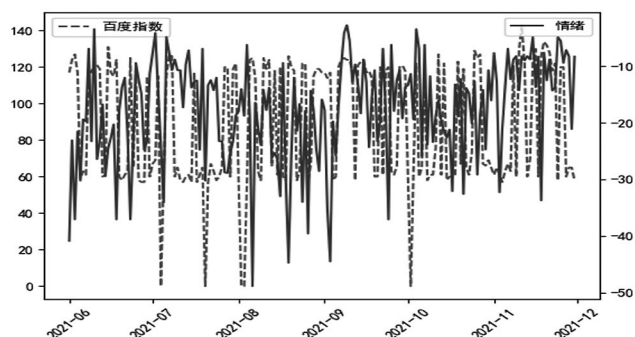


图4 情绪相关分析图

Fig.4 Analysis chart of sentiment correlation

从图4可以看出,情绪指数和百度指数走势大致相同,但是MIC的结果和预想目标不太一致。通过移动平均值对数据做平滑处理,消除一部分噪声。在对情绪指数和百度指数分别做30日的日均线,将情绪指数滞后5天,重新计算最大信息系数和绘制折线图,绘制的折线图如图5所示。

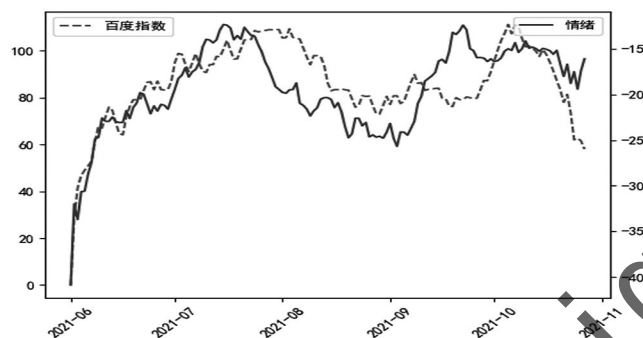


图5 情绪滞后折线图

Fig.5 Line chart of sentiment lag

计算得到MIC系数为67.84%,表现出较强相关性,从图5中也可以看出情绪指数和百度指数趋势基本一致,说明研究人员利用深度学习模型分析情感指数与百度指数具有较好的性能。根据市场情绪走势,2021年7月,新疆乌鲁木齐市投资者情绪高涨,楼市话题度较热,其中2021年9月市场情绪指数和百度指数出现反常情况,这与市场大环境低迷等情况有关。2021年,住房和城乡建设部提出把发展保障性租赁住房作为“十四五”住房建设重点任务,并且提出“三孩”配套支持政策落地:实施差异化租赁和购买房屋优惠政策,城市价值不断凸显导致情绪高涨,然而近年受某些因素的影响,几乎所有行业受到冲击,导致市场投资者情绪趋于悲观趋势,反映出投资者对房地产市场的担忧。现阶段国家和房地产企业拿出各种促销活动,吸引投资者眼球,使投资者对房地产关注度有所上升。

5 结论(Conclusion)

本文在构建情绪指数的基础上,引入深度学习提取文本情感极性分析房地产价格走势,采用新疆乌鲁木齐市房地产

行业文本语料和百度地产指数对模型进行验证,结果表明,通过CNN-BiLSTM模型提取的文本极性,通过构建量化情感指数计算得到某一日情感得分,将其与百度指数拟合,最大互信息值为67.84%,相关程度非常高,并且对市场效应进行分析,通过移动平滑可以让二者走势基本吻合。此模型既可以保证文本情感和房地产价格有较强的相关性,又可以保障二者走势的可视化趋于一致,市场决策者可以根据此模型研究结论,通过实时测度房地产评论舆情走势,为数字经济下市场决策者提供智力支持。

参考文献(References)

- [1] 李杰,李欢.基于深度学习的短文本评论产品特征提取及情感分类研究[J].情报理论与实践,2018,41(02):143-148.
- [2] 张爱琳,白丽娜.基于GM(1,1)模型的包头市房地产价格预测[J].建筑经济,2020,41(S2):266-269.
- [3] 刘洋,张悦,朱丽芳.基于几何布朗运动的房地产价格模拟预测[J].统计与决策,2018,34(09):86-89.
- [4] 柳冬,王雯璐,谢海滨,等.我国房地产价格影响要素分析与趋势预测[J].管理评论,2010,22(05):3-10.
- [5] 赵怡爽.房地产价格影响因素分析及预测[J].统计与决策,2014,(13):107-110.
- [6] 郎为爽,李晓红,张天抒,等.数据挖掘在房地产价格预测中的应用研究[J].数学的实践与认识,2020,50(05):306-311.
- [7] 何平,吴添,姜磊,等.投资者情绪与个股波动关系的微观检验[J].清华大学学报(自然科学版),2014,54(05):655-663.
- [8] 姜富伟,孟令超,唐国豪.媒体文本情绪与股票回报预测[J].经济学(季刊),2021,21(04):1323-1344.
- [9] 缪亚林,姬怡纯,张顺,等.CNN-BiGRU模型在中文短文本情感分析的应用[J].情报科学,2021,39(04):85-91.
- [10] BEHERA R K, JENA M, RATH S K, et al. Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data[J]. Information Processing & Management, 2021, 58(1):102435.
- [11] JIN Z, YANG Y, LIU Y. Stock closing price prediction based on sentiment Analysis and LSTM[J]. Neural Computing and Applications, 2019, 32(13):9713-9729.
- [12] 杨兴锐,赵寿为,张如学,等.结合自注意力和残差的BiLSTM-CNN文本分类模型[J].计算机工程与应用,2022,58(03):172-180.

作者简介:

李帅文(1997-),男,硕士生.研究领域:数据智能分析,文本挖掘。

刘继(1974-),男,博士,教授.研究领域:数据智能分析,文本挖掘。