网络舆情赋能金融科技股票收盘价预测研究

崔炎炎　刘立新

内容提要：金融科技发展进程中，网络舆情或许能给该行业指标数据的预测做出贡献，但相关研究尚不充分。本文将万得（wind）数据库中金融科技股票的交易数据作为金融科技行业的缩影，利用情感分类模型对爬取的11万余条微博文本中的投资者情绪进行挖掘。研究发现：负向投资者情绪占比对84只金融科技股票样本的平均收盘价存在负向影响，且具有长期稳定的均衡关系。进而，本文构建了以负向投资者情绪、工作日变量及其他金融科技股票量化指标数据为模型输入、预测金融科技股票平均收盘价指标数据的长短时间记忆神经网络模型（Long Short-Term Memory，LSTM）。结果表明：引入投资者负向情绪占比后，实验组LSTM模型比对照组的预测评价指标结果更加优秀，表明网络舆情对金融科技股票收盘价预测具有重要作用；实验组LSTM模型在不同预测期限上的预测效果评价指标均优于其他对照模型（随机森林、多层神经网络和支持向量回归模型），进一步证实了其良好的预测性能和模型稳健性。本文研究进一步丰富了自然语言处理和深度学习技术在金融科技领域的研究，为金融科技行业相关指标数据的预测提供了新的思路。

关键词：LSTM；投资者情绪；金融科技；股票收盘价

DOI: 10.19343/j.cnki.11–1302/c.2022.06.010

中图分类号：F222.3　　**文献标识码：**A　　**文章编号：**1002–4565(2022)06–0148–13

Research on the Impact of Internet Public Opinion on Fintech Stock Closing Price Forecast

Cui Yanyan & Liu Lixin

Abstract: In the development of Fintech, the Internet public opinion may contribute to the forecast of the industry's index data, but the relevant research is still insufficient. We use the Fintech stock transaction data in the Wind database as a microcosm of the financial technology industry, and use the sentiment classification model to mine the investor sentiment in the crawled more than 110000 Weibo texts. The study finds that the proportion of negative investor sentiment has a negative effect on the average closing price of the sample 84 Fintech stocks, and has a long-term stable equilibrium relationship. Furthermore, we construct a long short-term memory (LSTM) neural network model, which uses negative investor sentiment, weekday variables, and other quantitative index data of Fintech stocks as model inputs to predict the average closing price index data of Fintech stocks. The results show that after introducing the proportion of negative investor sentiment, the LSTM model of the experimental group has better results than the control group in forecast evaluation index, which illustrates the important role of internet public opinion in predicting the closing price of Fintech stocks. Then, the LSTM model in the experimental group also has better prediction results over different prediction periods than other control models (random forest, MLP

neural network and support vector regression model), which further confirms its good prediction performance and model robustness. The article further enriches the research of natural language processing and deep learning technology in the field of Fintech, and provides new ideas for the prediction of relevant index data in the Fintech industry.

Key words: LSTM; Investor Sentiment; Fintech; Stock Closing Price

一、引言

近年来，我国数字经济保持快速发展态势，政府相关部门不断推进数字产业化和产业数字化[①]发展进程。随着大数据、云计算、物联网、人工智能、区块链等技术群落的渗透，传统产业数字化、智能化水平明显提高，经济产业未来发展的边界也持续拓宽。在数字技术助力下，我国服务业的全面变革正在发生。其中，金融产业数字化日趋成熟，一系列新型信息通信技术在金融领域的创新应用也成为业界热议的话题。

一方面，我国金融科技的发展水平已经处于世界领先地位[②]，对促进我国经济高质量发展具有重要意义。在中小企业融资难问题上，金融科技运用网络数字资料弥补信用数据缺失的空白，缓解了中小企业囿于缺乏抵押品和公开信用记录导致金融机构难以评估其贷款信用风险而陷入贷款难的境况。在消费金融用户逾期风险加大问题上，金融科技助力消费金融从业机构动态评估消费金融用户在贷前、贷中、贷后三个不同环节中的还款意愿与还款能力，能够有效提升消费金融业务的风控效果。在解决民众理财业务个性化需求上，金融科技促进了智能营销的发展，使理财产品得到精准化推荐，为进一步挖掘理财用户潜力和发展智能投顾系统提供了支持。另外，在保险业务创新上，金融科技对保险业务的产品设计、销售和售后服务的全流程实现了智能化改造，一定程度上改善了保险业务的同质化问题。而在支付业务安全保障问题上，金融科技的发展也有助于监管方和支付机构利用相关技术提升用户数据的安全性，从而降低支付欺诈、洗钱、账户被盗用等支付犯罪行为的发生概率。

金融科技的发展对提升金融服务和人民生活水平起到了很好的推动作用，但另一方面，近年来金融科技风险事件的发生也为广大投资者带来了一定的风险隐患。以P2P网络借贷行业为例，由于行业准入门槛较低、平台风险监管不健全等原因，我国P2P网络借贷行业的发展从最开始的无人问津到井喷式生长，再到如今行业的全部清退，暴露了新金融业态带来的很多问题（李苍舒和沈艳，2019）。金融是现代经济的核心，防范金融领域系统性风险是关乎我国经济安全和社会稳定的大事。金融科技作为助力金融领域变革发展的驱动力有必要进行合理的发展规划和监管，否则，一旦金融科技领域的风险得不到妥善处理就可能传播到整个金融体系，从而引发系统性金融风险。

纵览我国金融科技行业，金融科技股票市场的交易数据能够直接反映整个金融科技行业的发展现状，揭示隐藏的风险，是投资者和监管方预判风险的参考依据。如果能较好地预测金融科技股票市场的交易数据，就能更清晰地掌握金融科技行业态势，预判金融科技市场风险，进而更好把握金融科技发展方向。

---

①据中国信息通信研究院《G20 国家数字经济发展研究报告（2018 年）》，数字产业化即信息通信产业，包括电子信息制造业、电信业、软件和信息技术服务业、互联网行业等；产业数字化即传统产业由于应用数字技术所带来的生产数量和生产效率的提升，其新增产出构成数字经济的重要组成部分。

②在 2018 年全球金融科技发明专利排行榜（TOP20）中，有 6 家企业属于我国，且这 6 家企业的专利申请量在上榜的 20 家企业中占比高达 42%。

一直以来，股票市场中的很多投资者会聚集于微博、论坛、贴吧等相互交流，发表自己对股票市场的看法。这些文字信息能够表达当下投资者的情绪，其是否可以对股票市场的量化指标数据进行预测是个值得关注的问题。因此，本文以我国金融科技行业为立足点，以金融科技股票市场交易的量化指标数据和金融科技领域的网络舆情资料为研究对象，提出以下两个假说。

假说1：投资者情绪与金融科技股票市场量化指标数据存在相关关系。

假说2：投资者情绪能够对预测金融科技股票市场量化指标数据起到重要作用。

为证明上述假说，本文首先从爬取的微博博文或微博博文评论信息中提取投资者发表的文字评论内容，运用所构建的情感分类模型对投资者的情绪表达进行分类，并按照时间顺序得到各类情绪表达的时间序列。其次，将投资者情绪的分类表达时间序列与wind数据库中84只金融科技股票市场的量化指标数据进行相关关系分析和协整关系检验，发现负向投资者情绪占比对样本中金融科技股票市场的平均收盘价指标数据具有反向影响作用，两者之间存在长期稳定的均衡关系。最后，为了预测金融科技股票市场的量化指标数据——平均收盘价，并探究网络舆情对预测金融科技股票市场量化指标的重要性，本文构建了长短期记忆网络（LSTM）模型，对重点关注变量设置实验组和对照组进行实验，并从模型预测效果和稳健性上与其他机器学习模型进行了对比。本文可能的特点和创新之处在于，首先，考虑到金融科技股票市场的数据变化较快，相较于传统的预测方式，本文引入随着时间更迭的投资者情绪进行预测，对提高模型的预测效果有很大帮助。此外，本文以我国金融科技行业为框架，将投资者情绪与金融科技股票市场的量化指标数据搭建桥梁构造预测模型，弥补了数字经济时代下投资者情绪在我国金融科技股票市场指标数据预测研究中的不足，为进一步把握当下我国金融科技行业现状提供了支持，也为展望我国金融科技行业未来的发展方向、制定市场策略提供了有效的数据佐证。

二、文献综述

（一）自然语言处理与深度学习技术

网络文本信息中投资者情绪的挖掘和分类涉及到自然语言处理与深度学习技术，作为自然语言处理领域的重要主题之一，文本分类由多项具体任务组成（Minaee等，2021），包括情感分析、新闻分类、主题分析、问答系统以及自然语言推理。随着自然语言技术的发展，训练机器读懂人类语言的模型发展得越来越丰富。在这一过程中，深度学习技术的作用功不可没。

至今，自然语言处理领域的发展通常可分为三个阶段。第一阶段主要包括分布式词表征（词向量）的实现（Bengio等，2003），通俗来讲就是通过数据量化的方式来对人类的词语进行表示。其中，Word2vec模型（Mikolov，2013）和Glove模型（Pennington等，2014）推动了词嵌入模式在自然语言处理领域的广泛发展，提高了自然语言建模中词语相似性等问题的模型适应能力。在理论上，基于预测的Word2vec模型与基于统计的Glove模型主要差异在于算法实现过程中计算损失方式的不同。在实际应用上，Glove模型较Word2vec模型更适用于大规模语料的训练，具有更强的并行化算法执行能力。第二阶段更多地包含以卷积神经网络（Convolutional Neural Networks，CNN）和循环神经网络（Recurrent Neural Network，RNN）为基础的模型对词序列信息提取过程的探索。以CNN为基础的模型更多地利用卷积结构从“空间”上训练模型，而以RNN为基础的模型则着重于时间长短上的“记忆”（Lecun和Bottou，1998），其中应用较为广泛的是基于LSTM的模型（Tai等，2015；Zhou等，2016）。相较于传统RNN模型而言，该类模型缓解了RNN的梯度消失和梯度爆炸问题，能够在

更长的序列中得到更好的表现。进而，以LSTM为框架进行特征提取并基于语言模型进行训练的ELMo模型（Peters等，2018）又将词语上下文的相关关系进行了很好的学习，且能够动态地学习出在不同上下文中词向量的表示含义，使得自然语言处理领域的发展更进一步。最后，第三阶段则以注意力机制作为划分。Transformer模型（Vaswani等，2017）利用注意力机制替代了RNN，在多项任务上表现优秀。而以Transformer为框架进行特征抽取，在预训练步骤中实现并行训练掩码语言模型（Masked Language Model，MLM）和下一句序列预测任务（Next Sentence Prediction，NSP）的BERT模型（Devlin等，2019）更是得到了广泛关注，该模型通过为语料中的部分词语进行掩码重建的方式来学习词序列中的语义信息，进一步提高了自然语言建模任务的模型处理能力。而后以BERT为原型进行改进的RoBERTa模型（Liu等，2019）和ALBERT模型（Lan等，2020）都为更好地应用BERT模型进行了不同参数结构调整的尝试，在自然语言处理的某些任务上也有较好的表现。

（二）投资者情绪

随着自然语言处理领域的发展，如何对经济金融领域中非结构化的不确定性信息，如投资者情绪等进行挖掘、把握，并将其对经济金融活动产生的影响构建相关模型进行分析，逐渐成为研究热点。现有文献中，欧阳资生和李虹宣（2019）立足于金融市场，将关于网络舆情对金融市场影响的研究进行了梳理。王靖一和黄益平（2018）利用所构建的金融科技情绪指数刻画了媒体对于金融科技行业的关注情况与正负情感值，进而分析了媒体情绪对于网络借贷市场活动的影响。Oliveira等（2017）从Twitter数据集包含的微博客中获取了网友们的投资者情绪、注意力指标和一些信心指标，对股票市场各种指数和投资组合的回报率、波动性和交易量进行了预测。结果发现Twitter微博客中的投资者情绪和微博客的发布量对标普500指数、低市值的投资组合和某些行业的收益预测具有相关影响。杨晓兰等（2016）从采集的90余万条股吧评论内容中提取了投资者情绪指标，与构建的本地关注指标一起对股票市场的一些量化指标进行了实证分析，结果发现若投资者情绪积极，本地关注指标就正向作用于股票收益率；但若投资者情绪为消极，该作用就为负。汪昌云和武佳薇（2015）选择样本首次公开募股（Initial Public Offering，IPO）公司在上市前被各主流媒体报道的内容为研究对象，设计了媒体语气指标来代表公司的投资者情绪，研究发现负面媒体语气能更好地解释IPO抑价率等变量的趋势。邵新建（2015）等发现拟上市公司可以通过加大正面广告的宣传力度提高投资者对其的乐观情绪，进而提高证券发行价格。张谊浩等（2014）验证了投资者网络搜索行为对股票市场资产定价的影响。顾文涛等（2020）发现，通过构建相应的情绪指数能有效提高股票收益率的预测效果。

上述文献包括了投资者情绪对股票市场量化指标数据预测研究，说明投资者情绪对于经济金融问题研究的重要性，但关于投资者情绪对金融科技股票市场的作用并没有引起更多关注。另外，上文提到的LSTM模型不仅为自然语言处理领域的发展提供了基础，也能为更好地预测股票市场的相关问题提供支撑。其中，杨青和王晨蔚（2019）利用LSTM神经网络模型对全球30个股票指数的三种不同预测期限进行了实证分析，结果表明该模型具有较强的泛化能力，预测效果也很好。Xiong等（2015）则应用LSTM神经网络模拟了标准普尔500指数的波动性，测试集实证结果也显示LSTM模型能够更好地预测股票指标。为了更好地对金融科技股票市场的状态进行描述，本文利用自然语言处理领域的文本分类模型对网络文本信息中的投资者情感进行分析，并对样本金融科技股票的平均收盘价指标进行预测，以进一步阐释金融科技行业中投资者情绪的作用以及LSTM模型对股票市场指标预测的效果。

三、样本数据处理与分析

（一）数据来源

互联网发展降低了人们获取信息的成本，也丰富了人们表达自我观点的途径，以智能手机为载体，微博依靠其简单的操作方式已经成为众多网民最常使用的搜索信息或表达自我情绪的手机软件之一。因此，本文以“金融科技”为关键词，运用网络爬虫技术从微博的博文或微博博文的评论中获取了2020年5月6日至12月15日每日1000条，共计115463条中文文本信息内容。

此外，对于本文样本量化指标的数据资料，囿于数据的可获得性，本文在wind数据库中共收集了我国84家金融科技成份类目公司股票[①]的数据资料，包括2020年5月6日至12月15日（除周六日和法定节假日以外）共计152天的每日收盘价（元）、每日开盘价（元）、每日成交量（股）和每日市盈率。

（二）文本信息的情感分类

1.文本信息标注。

为了提取本文所采集的文本信息的情感，首先对微博博文或微博博文评论文本信息中的情感含义进行预设分类。经过对样本文本信息的考量，本文将文本的情感共分成负向文本、中立文本和正向文本三个类别。其中，负向文本是指文本信息中包含对金融科技行业股票的消极看法、对金融科技行业股票过往表现失望、对金融科技公司高管团队的质疑以及对金融科技公司业务和服务等的批评。正向文本内容则包括对金融科技股票、金融科技公司发展情况的好评和夸奖。其他文本内容则归为中立文本一类。

为了应用情感分类模型，本文从采集的文本信息样本中随机选取了20000条文本内容进行人工的情感分类标注工作，并在数据集中用数字–1代表负向的文本内容、数字0代表中立的文本内容、数字1代表正向的文本内容。

2.情感分类模型。

自然语言处理领域的情感分类模型发展至今，已经能够较好地帮助人们用计算机来代替人工完成多项工作，此处的文本情感分类就是其中一项具体任务。在上文完成对随机抽取的20000条文本内容进行人工情感分类后，计算机学习人类进行此项分类工作的“模板”[②]已经具备，而进一步“教”会计算机开展此项任务的过程则要应用自然语言处理领域的文本分类模型。本文实证部分应用BERT情感分类模型，该模型由预训练模型和下游任务模型结合而成。

应用BERT模型进行文本分类的具体过程包括三步。第一步，输入样本文本内容。图1是BERT模型的文本输入结构示意图，共有标记词嵌入（Token Embeddings）、片段词嵌入（Segment Embeddings）以及位置词嵌入（Position Embeddings）三部分。其中，标记词嵌入是对某个词语的编码，目的是将词语向量化，并用特殊的标记词[CLS]和[SEP]分别表示一句话的开始和结束；片段词嵌入表示某个词语所在句子的编码，以区分一段话中的不同句子；位置词嵌入则是某个词语在句子中的位置编码。另外，文本输入部分还会对样本中每一个文本长度小于BERT模型最大文本长度参数的文本进行填充（padding）为0的处理。

第二步，选择预训练模型。BERT模型为不同种语言的任务提供了不同的预训练模型。结合中文文本应用场景的实际情况，本文选择BERT-Base，Chinese模型[③]作为预训练模型，并在该模型的基础上进行下游任务的微调过程。BERT模型文本情感分类任务的预训练和微调流程结构如图2所示。

---

①因篇幅所限，样本 84 只金融科技股票的说明以附表展示，见《统计研究》网站所列附件。

②“模板”是指人工手动分类结果的 20000 条文本内容及其数字标签。

③具体而言，该模型中 Transformer 编码器的隐藏网络层数 L 为 12 层，隐藏层神经元节点数 H（即 Feed Forward 输出向量的维数）为 768，多头注意力机制的头数为 12，共计有 110M 参数量。

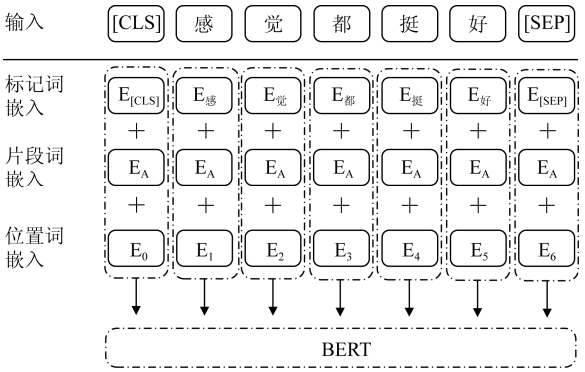

图1　BERT模型本文输入结构示意图

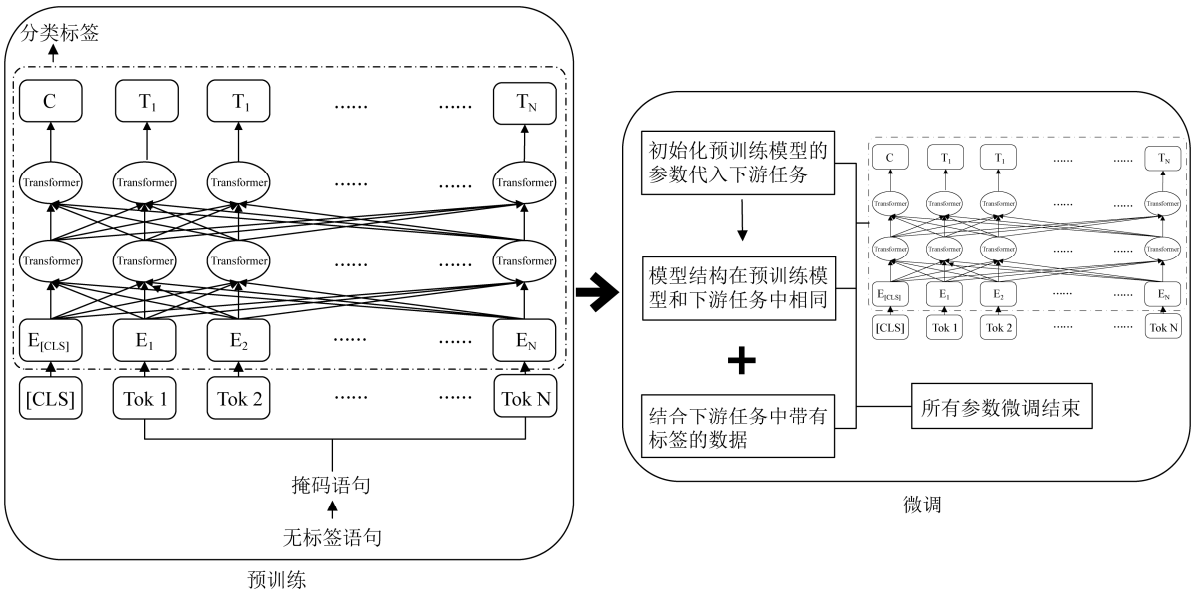

图2　BERT模型文本分类任务流程示意图

第三步，微调下游任务。图3是本文爬取的全部微博博文或微博博文评论内容的文本长度分布情况，可以看到样本中，文本字数大部分处于1~100个字之间，较少部分处于301~502个字之间，而文本内容字数最大长度为502。因此，本文将BERT模型最大文本长度参数设置为502。

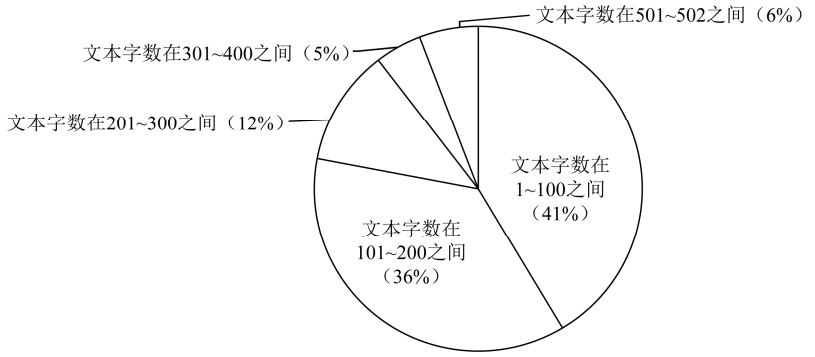

图3　文本内容长度分布图

为了评价BERT模型的效果，本文将前文分类后的20000条文本信息顺序随机打乱10次，每次取前16000条文本内容和数字标签结果作为训练集，其余的4000条样本作为验证集。应用正确率（accuracy）、精确率（precision）、召回率（recall）和F1得分（F1-score）4个指标对BERT模型和Kim于2014年提出的基于CNN的文本分类模型（TextCNN模型，Kim，2014）的预测效果进行对比评价。TP_{-1}表示模型将–1类样本正确地预测为–1类样本的个数（TP_0、TP_1同理），FP_{-1}表示模型将非–1类样本错误地预测为–1类样本的个数（FP_0、FP_1同理），TN_{-1}表示模型将非–1类样本正确地预测为非–1类样本的个数（TN_0、TN_1同理），FN_{-1}表示模型将–1类样本错误地预测为非–1类样本的个数（FN_0、FN_1同理）。因此，各评价标准的公式定义如下：

$$accuracy=\frac{TP_{-1}+TP_0+TP_1+TN_{-1}+TN_0+TN_1}{TP_{-1}+TP_0+TP_1+TN_{-1}+TN_0+TN_1+FP_{-1}+FP_0+FP_1+FN_{-1}+FN_0+FN_1} \tag{1}$$

$$precision=\frac{TP_{-1}+TP_0+TP_1}{\left(TP_{-1}+FP_{-1}\right)+\left(TP_0+FP_0\right)+\left(TP_1+FP_1\right)} \tag{2}$$

$$recall=\frac{TP_{-1}+TP_0+TP_1}{\left(TP_{-1}+FN_{-1}\right)+\left(TP_0+FN_0\right)+\left(TP_1+FN_1\right)} \tag{3}$$

$$\mathrm{F1}-score=\frac{2\times precision\times recall}{precision+recall} \tag{4}$$

其中，正确率是指模型正确预测样本的个数与所有带标签样本个数的比值；精确率是指所有被识别为某类别的样本中，确实为该类别的比例；召回率是指所有某类别样本中，被正确识别为该类别的比例；而F1得分是精确率和召回率的调和平均，该值最大值是1，最小值是0。两模型的10次评价结果如表1所示。

表1　　**评价指标结果**

实验次数	BERT模型各指标结果				TextCNN模型各指标结果			
	accuracy	*precision*	*recall*	F1-*score*	*accuracy*	*precision*	*recall*	F1-*score*
1	0.8670	0.8633	0.8670	0.8643	0.8300	0.8243	0.8300	0.8264
2	0.8615	0.8584	0.8615	0.8597	0.8380	0.8332	0.8380	0.8348
3	0.8660	0.8615	0.8660	0.8613	0.8335	0.8313	0.8335	0.8321
4	0.8565	0.8522	0.8565	0.8536	0.8305	0.8260	0.8305	0.8278
5	0.8547	0.8498	0.8548	0.8513	0.8357	0.8301	0.8357	0.8314
6	0.8562	0.8526	0.8562	0.8539	0.8210	0.8171	0.8210	0.8187
7	0.8570	0.8527	0.8570	0.8543	0.8245	0.8201	0.8245	0.8214
8	0.8610	0.8590	0.8610	0.8599	0.8303	0.8330	0.8303	0.8313
9	0.8545	0.8484	0.8545	0.8496	0.8325	0.8302	0.8325	0.8311
10	0.8645	0.8619	0.8645	0.8630	0.8353	0.8275	0.8353	0.8265

表1中BERT模型10次验证结果的正确率平均为0.8599，精确率平均为0.8560，召回率平均为0.8599，F1得分平均为0.8571。TextCNN模型10次验证结果的正确率平均为0.8311，精确率平均为0.8273，召回率平均为0.8311，F1得分平均为0.8282。上述结果显示出了BERT模型进行文本分类的优势和稳定性。本文BERT模型微调的其他具体参数结构及相关说明如表2所示，所有文本样本分类结果如表3所示。

3.投资者情绪与股票平均收盘价的关系。

由于我国股市收盘时间是在每个交易日的15点，为了进一步探究所挖掘的投资者情绪与84只样本金融科技股票平均收盘价指标数据之间的相关关系，本文首先对前文分类后的投资者情绪在时间

表2　　BERT模型具体参数

参数名称	参数描述	参数设置
max_seq_length	最大文本长度	502
train_batch_size	每批样本的大小	16
learning_rate	学习率	0.00002
num_train_epochs	全部训练集进行几次训练	4

表3　　样本文本最终分类结果　　（条）

负向评论	中立评论	正向评论
25103	22762	67598

上进行分离预处理，并与其他样本量化指标数据进行匹配，最终得到在数据采集期间内的152期负向投资者情绪占比的时间序列。接下来，本文对152期负向投资者情绪占比和84只样本金融科技股票平均收盘价指标数据两个时间序列之间的相关性进行了分析，结果表明Pearson相关系数为–0.1992，且相关性p值小于0.05。由此，本文以假设1为基础结合Pearson相关分析结果提出更加具体的假设，即：投资者负向情绪占比与金融科技股票平均收盘价之间可能存在负向作用关系，即随着投资者情绪负向评论占比的增多，金融科技股票市场的平均收盘价会降低。

为了更加严谨地对该假设进行科学论证，本文运用R软件对其进行协整关系检验。协整关系检验常被用来检验两个非平稳经济变量时间序列X和Y之间的关系，但协整关系检验的前提条件是时间序列X和Y必须是同阶单整的。鉴于本文样本时间序列的非平稳性，首先对样本时间序列进行一阶差分处理，再采用Engle-Granger（EG）两步法对样本时间序列进行协整关系检验。结果显示样本两个时间序列进行最小二乘（OLS）回归后的残差是平稳的，因此本文认为投资者负向情绪占比与金融科技股票平均收盘价之间具有长期稳定的均衡关系。至此，假说1得以证明。

四、LSTM预测模型的构建

（一）LSTM模型理论基础

为了更好地预测金融科技股票的平均收盘价，本文选择LSTM模型作为首选模型。理论上，LSTM模型适用于时间序列数据的预测，作为一种特殊的RNN结构，该模型能够在“长期记忆”中有很好的表现。基于Hochreiter和Schmidhuber（1997）对LSTM的提出，将该模型的训练过程示于图4，具体包括：对数据集进行LSTM输入结构的匹配；前向传播计算每个LSTM神经元的输出值；反向传播计算每个LSTM神经元的误差值；根据对应的误差值优化更新不同的权重值，得到不同的参数结构；对不同参数结构的模型进行评估，循环上述步骤直至得到最优化的LSTM结构。

在图4 LSTM神经元结构中可以看到每个LSTM神经元细胞的工作过程。具体到某一时刻t，对于输入序列$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$，首先，“遗忘门”$\boldsymbol{f}_t$决定了细胞中需要丢弃的信息，该门读取上一个LSTM神经元在$t-1$时刻的输出$\boldsymbol{h}_{t-1}$和此时的数据输入$\boldsymbol{X}_t$，通过*sigmoid*层输出一个0到1之间的数值，决定信息保留和舍弃的程度①，并将该数值传递给$t-1$时刻的细胞状态$\boldsymbol{c}_{t-1}$。“遗忘门”的计算公式：

$$\boldsymbol{f}_t=\sigma\left(\boldsymbol{w}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f\right) \tag{5}$$

①1 表示“完全保留”，0 表示“完全舍弃”。

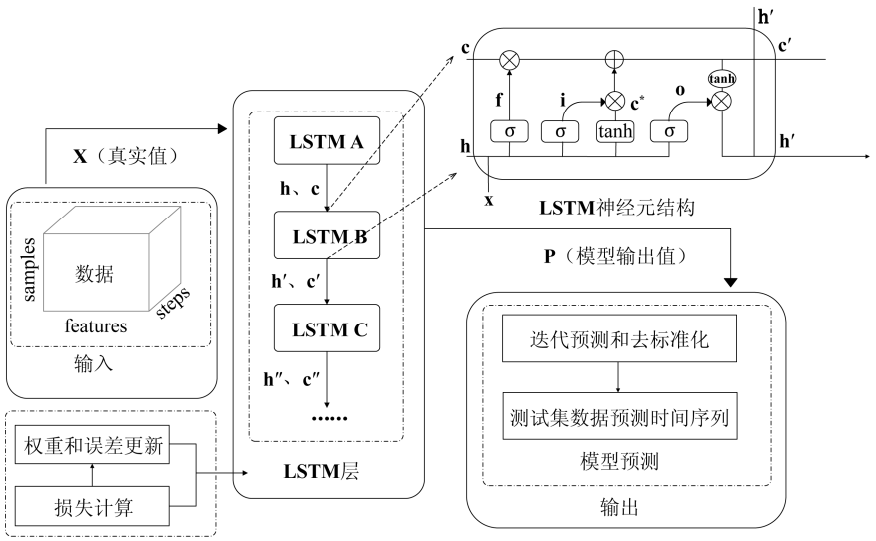

图4　LSTM模型训练流程

其次，“输入门”$\boldsymbol{i}_t$则要确定需要更新的信息。该步骤包含两部分：一是*sigmoid*层决定了哪些值需要进行更新；二是*tanh*层创建了一个新的候选向量$\boldsymbol{c}_t^*$，会成为新的细胞状态的组成部分。“输入门”部分的计算公式：

$$\boldsymbol{i}_t=\sigma\left(\boldsymbol{w}_i\cdot\left[\boldsymbol{h}_{t-1},\boldsymbol{x}_t\right]+\boldsymbol{b}_i\right) \tag{6}$$

$$\boldsymbol{c}_t^*=\tanh\left(\boldsymbol{w}_f\cdot\left[\boldsymbol{h}_{t-1},\boldsymbol{x}_t\right]+\boldsymbol{b}_f\right) \tag{7}$$

前两步共同完成细胞状态的更新，即将细胞在$t-1$时刻的状态$\boldsymbol{c}_{t-1}$更新为$\boldsymbol{c}_t$，公式如下所示。其中，旧状态$\boldsymbol{c}_{t-1}$与$\boldsymbol{f}_t$相乘完成了信息的丢弃与保留，$\boldsymbol{i}_t$与候选向量$\boldsymbol{c}_t$相乘决定了新状态的变化程度。

$$\boldsymbol{c}_t=\boldsymbol{f}_t\cdot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\cdot\boldsymbol{c}_t^* \tag{8}$$

最后，“输出门”$\boldsymbol{o}_t$决定了细胞状态的输出。*sigmoid*层决定当前细胞状态的哪些部分需要输出；同时，会把当前时刻的细胞状态通过*tanh*层进行处理①，并将其与*sigmoid*层的输出相乘。“输出门”结构的计算公式为：

$$\boldsymbol{o}_t=\sigma\left(\boldsymbol{w}_o\cdot\left[\boldsymbol{h}_{t-1},\boldsymbol{x}_t\right]+\boldsymbol{b}_o\right) \tag{9}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t+\tanh\left(\boldsymbol{c}_t\right) \tag{10}$$

其中，$\boldsymbol{w}_f$，$\boldsymbol{w}_i$，$\boldsymbol{w}_o$分别代表了“遗忘门”“输入门”和“输出门”的权重；$\boldsymbol{b}_f$，$\boldsymbol{b}_i$，$\boldsymbol{b}_o$分别代表了“遗忘门”“输入门”和“输出门”的偏置项。

（二）模型变量输入的确定

在对LSTM模型的输入进行考量时，为了更好地对金融科技股票的平均收盘价进行预测，本文参考Fu等（2019）的做法，将交易日变量作为一个5维的独热编码引进LSTM模型的输入中，以反映股票市场交易日的趋势性变化。并与样本金融科技股票市场的平均开盘价、样本金融科技股票市场平均成交量的对数、样本金融科技股票市场平均市盈率的对数以及负向投资者情绪占比一起引入LSTM模型进行实证分析。本文模型输入变量的具体说明如表4所示。

考虑到金融科技股票市场平均收盘价波动的周期性，本文将LSTM模型输入的步长设置为5，输

①得到一个[–1,1]之间的值。

出步长设置为1。此外，为了更清晰地看到LSTM模型输入变量对预测样本金融科技股票市场平均收盘价指标的不同作用，本文考虑对LSTM模型设置对照组和实验组进行对比讨论。具体如表5所示。

表4　　LSTM模型的输入

输入变量	解释说明
平均收盘价 CP_t	84只样本金融科技股票平均收盘价152期的时间序列数据
平均开盘价 OP_t	84只样本金融科技股票平均开盘价152期的时间序列数据
平均成交量的对数 $\ln(TV)_t$	84只样本金融科技股票平均成交量取对数处理后的152期的时间序列数据
平均市盈率的对数 $\ln(PER)_t$	84只样本金融科技股票平均市盈率取对数处理后的152期的时间序列数据
工作日变量 WD_t	152期时间序列日期所对应的工作日情况，该变量作为独热编码的取值为星期一、星期二、星期三、星期四和星期五中的任一个
负向投资者情绪 ZP_t	152期文本挖掘的负向投资者情绪占比时间序列

表5　　LSTM模型对照组及实验组说明

模型	模型作用描述	模型输入描述
对照组 M_1	验证假说2中的重点关注变量“负向投资者情绪”进入预测模型的重要性	CP_t、OP_t、$\ln(TV)_t$、$\ln(PER)_t$、WD_t
对照组 M_2	说明“工作日变量”进入模型的重要性	CP_t、OP_t、$\ln(TV)_t$、$\ln(PER)_t$
实验组 M_3	对所关注变量进行结果对比	表4中LSTM模型输入的全部变量。

五、实证结果分析

本部分LSTM模型的搭建在Python软件中以TensorFlow为框架完成，数据处理过程中用到的工具包包括numpy、pandas、sklearn等。其中，对于本文的时间序列类型数据，在进行数据集划分用于交叉验证时，应用了适用于时间序列类型数据的TimeSeriesSplit划分方式，以避免破坏时间序列类型数据本身的数据结构特征。此外，在调参过程中，首先采用50的迭代次数选择模型的优化器等其余参数，之后再改变迭代次数以确定拥有最好预测效果的模型，本文最终的LSTM模型训练参数设置如表6所示。

表6　　LSTM模型训练具体参数

参数名称	参数描述	参数设置
cell size	隐藏层的神经元个数	50
loss	损失函数的选择	均方误差（Mean Squared Error，MSE）
optimizer	优化器	Adam
batch size	每批样本的大小	16
learning rate	学习率	0.001
N epoch	全部训练集进行几次训练	300

（一）LSTM模型的实验对比

本文将数据采集期间152期时间序列中的前80%作为训练集，后20%作为测试集，对对照组和实验组模型 M_1、M_2 和 M_3 分别应用训练后的LSTM模型进行预测，并选择均方根误差（Root Mean Square Error，RMSE）、平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）和对称平均绝对百分比误差（Symmetric Mean Absolute Percentage Error，SMAPE）三个评价指标对模型的测试集预测效果进行评价。三个评价指标的具体公式如下所示：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{l=1}^{n}(\hat{y}_l-y_l)^2} \tag{11}$$

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t-y_t}{y_t}\right| \tag{12}$$

$$\mathrm{SMAPE}=\frac{100\%}{n}\sum_{t=1}^{n}\frac{\left|\hat{y}_t-y_t\right|}{\left(\left|\hat{y}_t\right|+\left|y_t\right|\right)/2} \tag{13}$$

其中，$\hat{y}_t$ 表示预测值，y_t 表示真实值，n 表示样本个数。

上述三个评价指标均衡量了预测值与真实值的接近程度，RMSE越接近于0，表示模型预测精度越好，MAPE和SMAPE越接近0%，说明模型效果越好。最终，对照组 M_1 、对照组 M_2 以及实验组 M_3 各评价指标的结果如表7所示。

表7　　对照组与实验组LSTM模型预测结果对比

模型	RMSE	MAPE（%）	SMAPE（%）
对照组M_1	0.0432	1.1996	1.1901
对照组M_2	0.0718	2.0213	1.9935
实验组M_3	0.0206	0.5236	0.5234

通过表7，将对照组 M_1 和对照组 M_2 的预测结果进行对比，发现引入工作日变量后，明显提高了模型 M_1 的预测效果，其中RMSE评价指标降低了39.83%，MAPE降低了40.65%，SMAPE降低了40.30%，说明工作日变量对预测金融科技股票平均收盘价的重要性。另外，通过对比对照组 M_1 和实验组 M_3，可以看到引入投资者负向情绪占比后，实验组 M_3 比对照组 M_1 的RMSE降低了52.31%，MAPE降低了56.35%，SMAPE降低了56.02%，说明投资者负向情绪占比对预测金融科技股票平均收盘价指标数据的重要性。由此，假说2得以证明。

（二）LSTM模型的预测效果与稳健性

通过对照组和实验组LSTM模型的对比分析讨论，已经证实了实验组 M_3 模型输入变量对预测样本金融科技股票市场平均收盘价的优秀表现，但为了进一步考量实验组 M_3 中LSTM模型本身的预测效果，本文选择随机森林（Random Forest）、多层神经网络（Multi-layer Perceptron Neural Network，MLP）和支持向量回归（Support Vector Regression，SVR）模型作为对照模型，并将模型预测的期限分为长期（52期）、中期（32期）和短期（12期）来检验LSTM模型的稳健性。本文对照模型的具体情况如表8所示。

表8　　对照模型概述

模型名称	模型训练情况
随机森林	随机森林算法（回归）是一种有监督的集成算法模型，该模型由众多回归树组成，其结果由众多回归树的结果汇总而来。本文随机森林（回归）模型数据训练后森林中树的数量设置为200。
多层神经网络	多层神经网络模型是一种具有多层感知器的前馈神经网络模型，该模型通过不断调整神经元之间的链接权值，能够更好地拟合训练数据之间的关系。本文多层神经网络模型训练后选择的优化器为Adam，激活函数为RELU，隐藏层数为3、隐藏神经元数分别为64、32和1。
支持向量回归	支持向量回归是支持向量机的回归版本，是一种基于结构风险最小化假设发展起来的非参数回归技术，能够对数据实现良好的泛化。本文支持向量回归模型训练后确定的核函数为高斯核函数（gamma=0.01），惩罚系数 $C=1\times10^7$ 。

表9报告了样本金融科技股票平均收盘价指标在长期、中期和短期三个不同预测期限内4个模型的预测效果。从整体上可以直观看到，LSTM模型在不同预测期限下的三个评价指标结果均低于其他三个对照模型。进一步通过具体计算可以得到，LSTM模型在三个预测期限上的RMSE平均值为0.022，相比于随机森林模型的0.028、支持向量回归模型的0.087以及多层神经网络模型的0.268分别降低了

21.428%、74.713%和91.791%；其次，LSTM模型在三个预测期限上的MAPE平均值为0.594%，相比于随机森林模型的1.004%、支持向量回归模型的2.738%以及多层神经网络模型的7.702%分别降低了40.837%、78.305%和92.288%；最后，LSTM模型在三个预测期限上的SMAPE平均值为0.593%，相比于随机森林模型的1.000%、支持向量回归模型的2.696%以及多层神经网络模型的8.036%分别降低了40.700%、78.004%和92.621%。因此，可以判断，LSTM模型在三个不同预测期限上，对样本金融科技股票平均收盘价的预测效果均优于其他对照模型，具有很好的预测能力。

表9　　样本金融科技股票平均收盘价预测结果

	长期（52期）			中期（32期）			短期（12期）		
	RMSE	MAPE（%）	SMAPE（%）	RMSE	MAPE（%）	SMAPE（%）	RMSE	MAPE（%）	SMAPE（%）
LSTM	0.028	0.722	0.721	0.021	0.585	0.584	0.018	0.475	0.474
Random Forest	0.036	1.560	1.553	0.023	0.821	0.818	0.024	0.632	0.628
SVR	0.072	2.060	2.033	0.086	2.734	2.693	0.104	3.420	3.361
MLP	0.451	13.026	14.147	0.267	7.528	7.450	0.087	2.552	2.512

在模型的稳健性方面，可以看到LSTM模型在三个不同预测期限上，对样本金融科技股票平均收盘价预测的三个结果评价指标均没有明显波动。鉴于随机森林模型的预测效果与LSTM模型最为接近，故本文将LSTM模型与随机森林模型的稳健性进行对比。经过具体计算可知，LSTM模型在三个不同预测时期上的RMSE评价指标预测结果之间相差的范围在35.714%之内，相比于随机森林模型的36.111%低了1.099%；LSTM模型在三个不同预测时期上的MAPE评价指标预测结果之间相差的范围在34.211%之内，相比于随机森林模型的59.487%低了42.490%；LSTM模型在三个不同预测时期上的SMAPE评价指标预测结果之间相差的范围在34.258%之内，相比于随机森林模型的59.562%低了42.483%。因此，在模型稳健性方面，LSTM模型优于随机森林模型。综上所述，基于本文实证数据，LSTM模型在预测效果和模型稳健性上均为最佳模型。

六、研究结论

本文利用爬取的网页版微博博文和微博博文评论的文本信息中提取的负向投资者情绪，反映股票市场交易日趋势性变化的交易日变量，金融科技股票市场样本的平均收盘价、平均开盘价、平均成交量的对数以及平均市盈率的对数作为LSTM预测模型的输入变量，对样本金融科技股票的平均收盘价进行了预测。从实证结果中可知，投资者情绪对预测金融科技股票样本的平均收盘价具有重要作用，与随机森林模型、多层神经网络模型和支持向量回归模型相比，本文构造的LSTM模型在预测效果和模型稳健性上来看均为最佳模型。

综上，本文研究对金融科技股票市场投资者制定投资策略以及监管金融科技公司信息披露具有一定意义。一方面，投资者负向情绪占比与金融科技股票平均收盘价指标之间存在负相关关系，投资者可以利用网络舆情信息预测金融科技股票平均收盘价的指标变化，从而制定相应的投资策略。另一方面，本文的研究结论也表明，金融科技股票市场信息的披露和监管十分重要。数字经济时代，各类数字信息已经渗透到人们生活的方方面面，促进网络舆情信息的真实化、及时性、透明化，对完善我国现代金融监管体系、提高我国金融科技水平具有重要作用，也是维护我国金融科技股票市场投资者利益的重要举措。另外，如何更加细化地对网络舆情中投资者的情绪进行分类，从而更好地对金融科技股票市场量化指标数据进行预测是未来值得探索的方向。

参考文献

[1] 顾文涛, 王儒, 郑肃豪, 等. 金融市场收益率方向预测模型研究: 基于文本大数据方法[J]. 统计研究, 2020, 37(11): 68–79.

[2] 李苍舒, 沈艳. 数字经济时代下新金融业态风险的识别、测度及防控[J]. 管理世界, 2019, 35(12): 53–69.

[3] 欧阳资生, 李虹宣. 网络舆情对金融市场的影响研究: 一个文献综述[J]. 统计与信息论坛, 2019, 34(11): 122–128.

[4] 邵新建, 何明燕, 江萍, 等. 媒体公关、投资者情绪与证券发行定价[J]. 金融研究, 2015(9): 190–206.

[5] 汪昌云, 武佳薇. 媒体语气、投资者情绪与 IPO 定价[J]. 金融研究, 2015(9): 174–189.

[6] 王靖一, 黄益平. 金融科技媒体情绪的刻画与对网贷市场的影响[J]. 经济学（季刊）, 2018, 17(4): 1623–1650.

[7] 杨青, 王晨蔚. 基于深度学习 LSTM 神经网络的全球股票指数预测研究[J]. 统计研究, 2019, 36(3): 65–77.

[8] 杨晓兰, 沈翰彬, 祝宇. 本地偏好、投资者情绪与股票收益率: 来自网络论坛的经验证据[J]. 金融研究, 2016(12): 143–158.

[9] 张谊浩, 李元, 苏中锋, 等. 网络搜索能预测股票市场吗?[J]. 金融研究, 2014(2): 193–206.

[10] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3(3): 1137–1155.

[11] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J/OL]. [1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arxiv.org), 2019-5-24.

[12] Fu X L, Zhang S, Chen J, et al. A Sentiment-aware Trading Volume Prediction Model for P2P Market using LSTM[J]. IEEE Access, 2019(7): 81934–81944.

[13] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997(8): 1735–1780.

[14] Kim Y. Convolutional Neural Networks for Sentence Classification[J/OL]. https://arxiv.org/abs/1408.5882, 2014-9-3.

[15] Lecun Y, Bottou L. Gradient-based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[16] Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J/OL]. [1907.11692v1] RoBERTa: A Robustly Optimized BERT Pretraining Approach (arxiv.org), 2019-7-26.

[17] Lan Z, Chen M, Goodman S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[J/OL]. [1909.11942v5] ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (arxiv.org), 2020-2-9.

[18] Minaee S, Kalchbrenner N, Cambria E, et al. Deep Learning Based Text Classification: A Comprehensive Review[J/OL]. [2004.03705v3] Deep Learning Based Text Classification: A Comprehensive Review (arxiv.org), 2021-1-4.

[19] Mikolov T. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013(26): 3111–3119.

[20] Oliveira N, Cortez P, Areal N. The Impact of Microblogging Data for Stock Market Prediction: Using Twitter to Predict Returns, Volatility, Trading Volume and Survey Sentiment Indices[J]. Expert Systems with Applications, 2017(73): 125–144.

[21] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[A]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)[C]. 2014: 1532–1543.

[22] Peters M, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations[A]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies[C]. 2018: 2227–2237.

[23] Tai K S, Socher R, Manning C D. Improved Semantic Representations From Tree-structured Long Short-term Memory Networks[J]. Computer Science, 2015, 5(1): 1–11.

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[A]. 31st Conference on Neural Information Processing Systems (NIPS 2017)[C]. 2017: 1–15.

[25] Xiong R, Nichols E P, Shen Y. Deep Learning Stock Volatility with Google Domestic Trends[J/OL]. [1512.04916] Deep Learning Stock Volatility with Google Domestic Trends (arxiv.org), 2015-12-15.

[26] Zhou P, Qi Z, Zheng S, et al. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling[A]. The 26th International Conference on Computational Linguistics: Technical Papers[C]. 2016: 3485–3495.

作者简介

崔炎炎（通讯作者），对外经济贸易大学统计学院博士研究生。研究方向为金融统计。电子邮箱: 201801410140@uibe.edu.cn。

刘立新，对外经济贸易大学统计学院教授、博士生导师。研究方向为金融统计。

(责任编辑: 张晓梅)