数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：一种基于 CNN-BiLSTM 多特征融合的股票走势预测模型
作者：徐月梅，王子厚，吴子歆
网络首发日期：2021-04-15
引用格式：徐月梅，王子厚，吴子歆．一种基于 CNN-BiLSTM 多特征融合的股票走势预测模型．数据分析与知识发现.
https://kns.cnki.net/kcms/detail/10.1478.g2.20210415.1407.002.html

网络首发时间：2021-04-15 16:14:05
网络首发地址：https://kns.cnki.net/kcms/detail/10.1478.g2.20210415.1407.002.html

一种基于 CNN-BiLSTM 多特征融合的股票走势预测模型

徐月梅[1]，王子厚[2]，吴子歆[1]
（1. 北京外国语大学 信息科学与技术学院，北京 100089；2. 国家计算机网络应急技术处理协调中心，北京 100029）

摘要 **[目的]**股票价格波动的本质是对新信息的反应，在传统基于股市数值分析的基础上，研究新闻对股票市场的影响，有助提高股票走势预测的准确率。**[方法]**引入卷积神经网络和双向长短时记忆模型挖掘财经新闻中的新闻事件类型和新闻情感倾向，提出一种深度融合股市财务数据、新闻事件特征及新闻情感特征的股票预测模型。为了验证所提模型对不同行业个股走势的可行性，分别选取家用电器行业和通信行业的两支股票作为实验对象。**[结果]**引入新闻事件和情感特征后，模型的预测准确率进一步提升，家用电器行业准确率提高了 11%，通信行业准确率提高了 22%。**[局限]**模型未考虑不同预测周期对股票预测的影响。**[结论]**引入新闻事件类型和情感倾向能够提高股票走势预测的性能。本文评估影响股票走势的因素，并对影响股票走势预测的特征重要性进行排序。

关键词 深度学习;特征融合;情感倾向;股票走势
中图分类号 **TP393**
DOI: 10.11925/infotech.2096-3467.2020.0907

A CNN-BiLSTM-based Multi-feature Integration Model for Stock Trend Prediction

XU Yuemei[1], WANG Zihou[2], WU Zixin[1]
（1. School of Information Science and Technology, Beijing Foreign Studies of University, Beijing 100089, China;
2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China）

Abstract **[Objective]**The essence of stock price fluctuation is a response to all available information. Based on the traditional financial numerical data analysis, this paper exploits the impact of online news on stock market to further improve the accuracy of stock trends prediction. **[Methods]** Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM) are introduced to extract news events and sentiment orientations of financial news from news data. A stock trends prediction model is then proposed to deeply combine the stock financial numerical data feature and the news event-sentiment feature to help improve the accuracy of stock trends prediction. In order to verify the feasibility of this model for predicting the trend of individual stocks in different industries, two individual stocks are selected as the

experimental objects, e.g., GREE Electric Appliance in the household appliance industry and ZTE in the electronic appliance industry. **[Results]** Experiment results show that the proposed model improves the prediction accuracy by 11% and 22% in the household appliance industry and the electronic appliance industry, respectively, compared with the exiting algorithm. **[Limitations]** The impact of prediction period on the performance of the proposed model has not yet been considered. **[Conclusions]** The news event and sentiment orientations of news play a great impact on the fluctuation of stock trends. This paper ranks the importance of every feature and finds the most important one that influences the stock trends prediction.

Key words Deep Learning; Feature Combination; Sentiment Analysis; Stock Trends

1 引言

股票走势预测是根据与股票相关的数据对股票价格的波动进行预判。现有研究表明，股票价格的走势并不遵循随机游走，在一定程度上可以被预测[1]。传统的量化投资属于早期的股票预测方法，通过结构化的、线性的历史交易数据对股票走势进行预测，主要采用线性回归、参数估计方法 [2] 。然而，股票价格的走势除了与历史交易数据相关，还容易受到非线性因素的影响，例如政策因素、投资心理、以及突发事件等。如何统筹考虑线性的历史交易数据以及非线性的因素对股票价格的影响，进一步提高股票预测的准确率是近年来的研究热点。

随着信息技术的发展，媒体新闻信息作为反映与股票走势非线性因素相关的重要信号，被用于提高股票走势预测的准确率。大量研究表明媒体新闻信息会对股票价格产生影响[3]。现有研究将媒体新闻信息中的情感极性（正面或者负面）作为反映市场状态的指标[4-6]。例如，Bollen 等人利用道琼斯指数分析推特的用户情感来进行股价预测[7]。现有研究确证了新闻媒体所蕴涵的情绪信号对于股票市场价格的预测能力，然而仍存在一些问题亟待解决。此外，新闻文本信息属于非结构化的数据，具备非线性、非平稳的特征，使得传统的量化投资分析方法并不适用。

股票走势预测是一个经典的基于时序数据的预测问题，近年来由于深度学习在文本处理方面的突出表现，越来越多学者尝试将深度学习用于解决基于时序数据的股票预测问题。循环神经网络（Recurrent Neural Network，RNN）将时间概念融入到网络结构中，天然地适用于处理时序数据。然而当输入的时序数据的序列太长时，RNN 会面临梯度消失问题。为了解决这一问题，长短时记忆模型（Long Short-Term Memory，LSTM）作为 RNN 的改进版本被提出。研究表明，LSTM 在处理基于时序数据的股票预测问题时性能优于 RNN 和传统的机器学习算法，例如随机森林、支持向量机（Support Vector Machine，SVM）等[8-10]。大多现有基于深度学习的股票预测问题采用 LSTM 或者改进的 LSTM，例如 PSO-LSTM [9]、Stacked LSTM[10]、time-weighted LSTM[11]搭建股票预测模型。

现有基于 LSTM 的股票预测模型主要通过分析文本信息的情感极性特征，将媒体新闻的情感极性与历史交易数据作为股票预测算法的输入，需要考虑以下 3 个方面的问题。首先，财经新闻中的情感极性并不明显，大多是对客观事件的总结和报道，使得财经新闻的情感极性分析准确率并不高，影响股票走势的预测准确率[12]。其次，如何将非结构化的文本信息特征与结构化的股票交易数据融合在一起。现有大多数研究将新闻文本的情感值与高维度的历史交易数据、公司财务数据直接拼接在一起，作为股票预测的输入[12-14]。这种方法很容易将情感信

息淹没在高维度的结构化信息中。添加了情感维度的股票预测准确率甚至低于不添加情感维度的股票预测方法[15]。最后，新闻文本的情感极性并不总是与股票走势的涨跌正相关。例如，新闻“中兴通讯高层大变动，少壮派受重用”的情感属性为正面，但并没有对中兴股票的上涨有积极影响。新闻文本情感极性可能与股票的涨跌负相关。与情感极性相比，新闻事件本身更能够代表媒体新闻对股票走势的影响。

为了解决上述 3 个问题，本文提出一种基于 CNN-BiLSTM 多特征融合的股票走势预测模型，通过融入新闻事件类型和情感极性提高股票走势预测的准确率。卷积神经网络（Convolutional Neural Network，CNN）在新闻文本的事件特征提取上的性能更为突出[16]，而双向 Bi-LSTM（Bi-LSTM）采用两个 LSTM 网络获取文本前向和后向的语境信息，更适合用于处理考虑上下文语境的情感极性判别，在情感分析上较单个 LSTM 能提升 3%的性能[17]。因此，所提模型一方面采用从新闻报道中提取出客观的财经事件，例如中标事件、上市事件、停牌事件等。另一方面，采用 Bi-LSTM 对新闻报道的情感极性进行分析，计算新闻文本的情感分值。股票新闻特征包括新闻的事件类型和情感分值，与股票的财务数值特征一起作为 LSTM 网络的输入，利用历史的股票信息预测股票未来的涨跌情况。此外，所提模型通过对比实验研究影响股票走势预测模型的特征重要性。

2 相关研究工作

现有对股票走势的预测研究可以分为两类：一类是基于数值分析，另一类是融合数值和文本信息的股票预测模型。

早期量化投资分析属于传统的数值分析方法，主要基于历史交易数据、公司财务数据和宏观数据对股票走势进行预测[18]。Kai Chen 等人验证个股的历史股票数据可以用来预测个股的未来走势[19]。这些方法主要根据历史交易数据发现并描述数据随时间变化的规律。为了挖掘大量数值信息的规律，传统的机器学习算法，例如 SVM、人工神经网络（Artificial Neural Networks，ANN）、朴素贝叶斯、随机森林等，被用于对大量的历史股票数据分析[20]。张玉川等人利用 SVM 对个股涨跌进行预测，通过实证分析验证 SVM 对个股涨跌分类的有效性[21]。Karen 等人对比 ANN 和 SVM 在股票走势预测上的性能，发现 ANN 在预测准确率上优于 SVM，前馈 ANN 由于能够同时预测股票走势的涨跌以及股票价格而被广泛采用[22]。然而，由于股票价格本质上是随机变量的噪声观测值，只采用历史交易数据进行分析具有一定局限而无法进一步提升预测效果。行为经济学理论指出，投资者面对复杂和不确定的决策问题时，很容易受到个人和社会环境情感状态的影响[23]。股票价格变化的根源是对新信息的反应，媒体新闻文章作为外生变量的信息，对短期价格预测有帮助[24]。

另一类融合股票数值信息和新闻信息的股票预测模型应运而生。Vanstone 等人研究新闻及新闻的情感极性是否会对股票价格的预测起作用[24]。研究表明通过统计与股票相关的新闻文本数量以及推特条数，并将其作为股票价格预测的输入可以进一步提高预测准确率。Chen 等人利用带有门循环单元的 RNN 模型研究新浪微博上的财经新闻的情感极性，并融合股价数值特征一起预测股票走势[25]。Manuel R.等人利用 CNN 和 RNN 研究融合新闻标题和技术指标的股票走势模型，证明新闻标题比新闻内容更有利于提高预测准确率[26]。岑咏华等人考察新闻网站、股吧、博客等媒介信息所蕴含的情感信号对于股票市场的影响效应，发现投资者对于积极情感的反应更及时更强烈[27]。

这些现有的融合股票数值信息和新闻信息的股票预测方法主要考虑提取新闻的情感极性作为股票数值信息的补充。然而，新闻是对客观事实的描述，情感极性大多比较隐晦，使得情感特征对于预测准确率的提升并不明显。考虑到这一局限，现有研究[24-26]大多采用情感表达较为明显的推特、微博文本作为新闻信息的来源。Zhao 等人提出使用隐含隐含狄利克雷分布（Latent Dirichlet Allocation，LDA)主题模型提取出微博文本的关键词，再基于关键词分析微博文本的情感特征作为股票预测的输入[28]。区别于现有工作，考虑到新闻事件比新闻情感更能代表新闻媒体信息对股票走势的影响，本文采用融合新闻事件和情感特征的多特征融合方法，对股票的数值特征进行补充，进一步提升股票预测的准确率。

3 基于新闻事件和情感特征的股票预测模型

股票走势预测是一个二分类问题，当涨跌幅高于阈值时，判为上涨样本；反之则为下跌样本。设采样间隔为 a 天，根据过去 $t-a$ 至 $t-1$ 天的数据，预测第 t 天的股票涨跌。

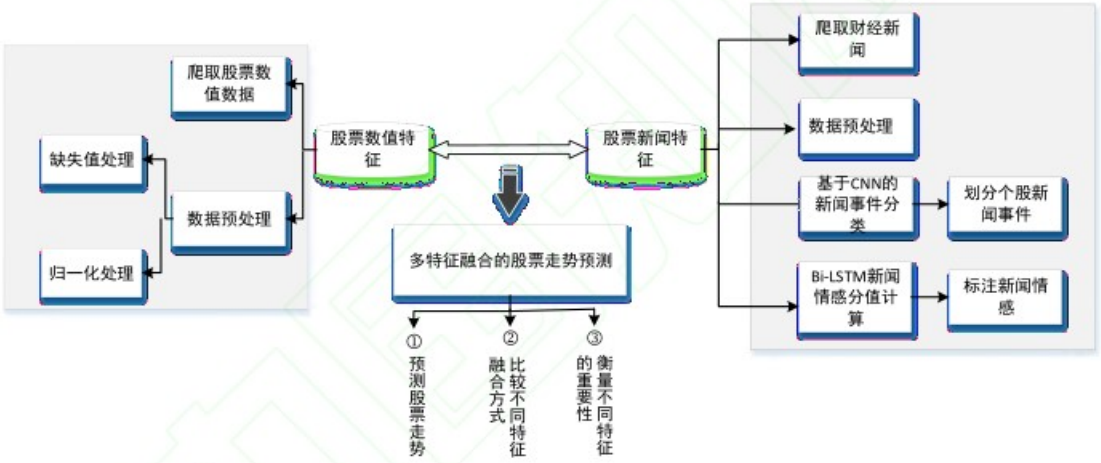


图 1 融合新闻事件和情感特征的股票走势预测流程图

Fig.1 The flowchart of stock trends prediction model based on combination of news event and news sentiment orientation

所提模型的架构如图 1 所示。基本思路是：

（1）分别爬取股票相关的财务数据和新闻标题数据，对财务数据和新闻标题数据进行预处理，得到股票数据库和新闻数据库。

（2）划分个股新闻事件、对新闻进行事件类型标注，利用 CNN 训练标注数据得到新闻事件分类器。

（3）对新闻文本进行情感标注，利用 Bi-LSTM 建立新闻情感分类器。

（4）将（2）和（3）训练得到的股票新闻特征和股票数值特征作为 LSTM 网络的输入，比较不同特征的融合方式、并预测股票走势的涨跌。

3.1 股票财务数值特征提取

相关研究表明公司的市盈率、市净率、主力资金净流入等指标跟个股的走势相关[19,29-30]。因此，股票财务数值特征选取公司的财务数据（如：市盈率、市净率），公司的资金流向数据（如：主力资金净流入、换手率）和股票数据（如：开盘价、收盘价）。考虑到大盘指数和板块指数对股票走势的影响，也选取了个股的大盘指数和板块指数作为股票财务数值指标。

对财务数值数据首先进行数据缺失值处理：如果某一天的数据缺少某个指标（停牌等原因），则将该天的数据去掉。此外，考虑到不同财务数值指标的性质不同，数量级也不同，直接使用指标的原始数值可能导致数值较高的指标在训练中有更为主导的作用，削弱数值较低的指标的影响，因此对财务数值数据进行z-score[31]标准化处理，保证不同指标数据之间的可比性。

设 $\boldsymbol{D_j}=[d_{1j},d_{2j},...,d_{ij},...,d_{nj}]$ 表示第 j 个财务指标在 T_n 天内的数值组成的向量，d_{ij} 表示第 j 个财务指标在第 i 天的数值，将 $\boldsymbol{D_j}$ 中的每一个数值进行z-score标准化处理如下：

$$d_{ij}=\frac{d_{ij}-\mu_j}{\sigma_j} \qquad (1)$$

式中：μ_j 和 σ_j 表示财务指标 $\boldsymbol{D_j}$ 中所有数值的均值和标准差。表1描述了由 p 个股票指标在 T_n 天内构成的财务特征矩阵，D_1 到 D_p 表示 p 个财务特征。

表 1 股票财务特征矩阵

Table1 Feature matrix of stock finance

	D_1	D_2	···	D_j	···	D_p
T_1	d_{11}	d_{12}	···	d_{1j}	···	d_{1p}
···	···	···	···	···	···	···
T_i	d_{i1}	d_{i2}	···	d_{ij}	···	d_{ip}
···	···	···	···	···	···	···
T_n	d_{n1}	d_{n2}	···	d_{nj}	···	d_{np}

3.2 基于 CNN 模型的新闻事件分类及特征提取

新闻事件特征提取从新闻标题中提取出客观的金融事件。首先对新闻标题进行数据预处理：利用结巴分词器进行分词和去停用词，同时引入自定义的停用词和金融词典提高分词准确率。自定义金融词典包括常见金融词汇、A 股上市公司代码及简称、A 股上市公司实际控制人及高管姓名等。

根据客观事件划分金融新闻事件。参考国泰安（CSMAR）经济数据库中新闻词条的关键字字段作为事件划分的分类依据，一共划分了 82 个新闻事件类型。表 2 列举了部分新闻事件。

表 2 部分新闻事件类型（部分）

Table2 Part of categories of news events

事件类别	事件名称
交易类	停牌 复牌 资金流入 资金流出 大宗交易 股价倒挂 创新高
股权类	挂牌 借壳 举牌 收购并购 资产重组 资产冻结 股权转让
投融资类	投资 投建 中标 发行债券 发行股票 可转债 募资 质押 分红
公司事务类	注册资本变更 快速发展 战略合作 拓展业务 高管减持或离职
外部事件类	登上龙虎榜 交易所处罚 评级利好 评级下调 政策利好

基于 CNN 搭建新闻事件分类器。新闻事件分类器的训练过程与基于 CNN 的文本分类过程类似，详细过程可参考笔者前期研究工作[32] ，其主要思路是：根据划分的 82 个新闻事件对股票新闻进行标注，训练 CNN 模型，模型包括输入层、卷积层、池化层和全连接层操作，每一层的输出是下一层的输入。首先采用 Word2Vec 训练新闻标题，将得到的词向量矩阵作为卷积层的输入，卷积层利用滤波器对新闻标题的词向量矩阵进行卷积操作，产生特征图。池化层对特征图进行采样，抽取每个特征图中最重要的特征传入全连接层。最后，全连接层通过 Softmax 函数获得新闻标题最终分类结果，输出新闻标题的事件类型。

表3 新闻事件特征矩阵

Table3 Feature matrix of news event

	S_1	S_2	…	S_j	…	S_q
T_1	s_{11}	s_{12}	…	s_{1j}	…	s_{1q}
…	…	…	…	…	…	…
T_i	s_{i1}	s_{i2}	…	s_{ij}	…	s_{iq}
…	…	…	…	…	…	…
T_n	s_{n1}	s_{n2}	…	s_{nj}	…	s_{nq}

通过CNN新闻分类器对特定股票每天的新闻进行统计。每条新闻输入到新闻分类器将输出一个事件类型，再统计每天各个事件出现的频率，得到表3的新闻特征矩阵。S_1到S_q表示q个新闻事件，这里q=82，s_{ij}表示在日期T_i出现事件特征S_j的频次。

3.3 新闻文本情感特征提取

新闻事件是对客观事件的描述，而新闻文本的情感表征新闻事件的情感极性，例如消极或者积极。新闻文本的情感极性判断需要考虑上下文的语境信息，Bi-LSTM采用2个LSTM网络进行训练，一个训练序列从文本前面开始，一个训练序列从文本后面开始，这两个训练序列连接到同一个输出层。Bi-LSTM能够整

合每个点的过去和未来信息，相比于单个LSTM在文本情感极性判断上更有优势。图2为利用Bi-LSTM计算新闻标题情感分值的过程。

对数据预处理后的每一条新闻标题 d，最长裁剪 N 个词语，并设置 LSTM 的处理步长为 N。对于长度不足 N 的新闻标题，进行左置零补齐。针对每一个采样时刻（$t \le N$），将词语 w_t 通过 Word2Vec 训练得到的词向量 $\boldsymbol{x}_t$ 输入到一个包含 L 个神经元的 LSTM 神经网络层。该神经网络层输出一个维度为 L 的隐含状态向量 $\boldsymbol{h}_t$。每一个神经元设置三种门限结构，即遗忘门、输入门和输出门，基于过去的隐含状态向量 $\boldsymbol{h}_{t-1}$ 和当前输入 $\boldsymbol{x}_t$，决定需要遗忘哪些信息、输入哪些新的信息以及对哪些新的记忆信息编码输出得到 $\boldsymbol{h}_t$。具体地，在时刻 t，LSTM 层的计算如公式(2)-(6)所示[33]：

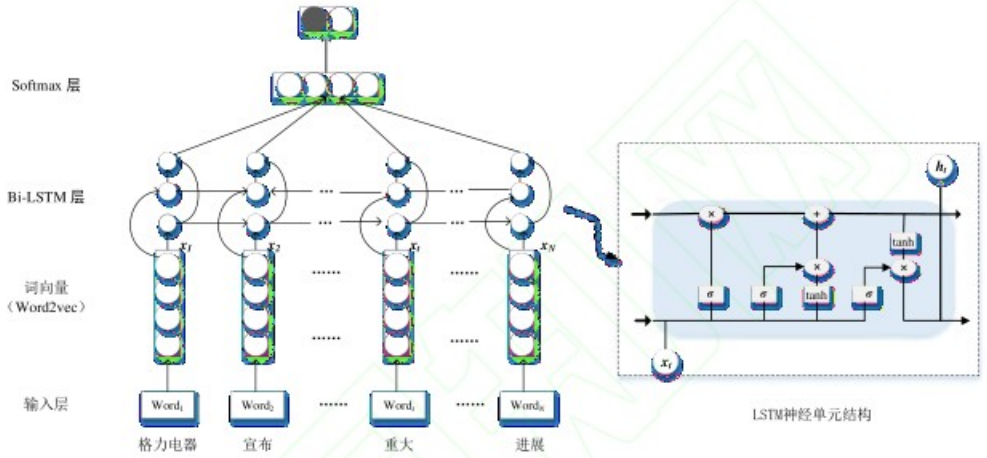


图 2 基于 Bi-LSTM 的新闻情感分析模型

Fig.2 Analysis model of news sentiment orientation based on Bi-LSTM

$$f_t = \sigma(W_f \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \quad (3)$$

$$\boldsymbol{C}_t = f_t \Theta \boldsymbol{C}_{t-1} + i_t \Theta \tanh(W_c \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \quad (4)$$

$$o_t = \sigma(\boldsymbol{W}_o \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b} \quad (5)$$

$$\boldsymbol{h}_t = o_t \Theta \tanh(\boldsymbol{C}_t) \quad (6)$$

首先，遗忘门基于 Sigmoid 函数 σ 判断过去的记忆对于当前的记忆状态下有多大意义值得保留，通过公式(2)计算生成系数 f_t。接着，输入门判断当前的词输入向量 $\boldsymbol{x}_t$ 在多大意义上值得保留，根据公式(3)并生成系数 i_t。然后，神经元根据公式(4)更新生成当前时刻的状态 $\boldsymbol{C}_t$。最后，输出门基于 $\boldsymbol{C}_t$ 判断新的记忆在多大意义上可以输出，输出的隐含状态 $\boldsymbol{h}_t$ 由公式(6)表示。公式(2)-(6)中，$\boldsymbol{W}_o$ 和 $\boldsymbol{b}_o$ 分别表示权重矩阵和偏置向量，Θ 为乘操作。

经过上述计算，得到新闻标题 d 在 $t = N$ 时刻上的隐含状态编码 $\boldsymbol{h}_N$，基于 $\boldsymbol{h}_N$ 通过 Softmax 函数得到 d 在不同情感类别{1-1}，即{正向，负向}上的概率分布向量：

$$\boldsymbol{y}_d = Soft\max(\boldsymbol{W}_o \cdot \boldsymbol{h}_N + \boldsymbol{b}_o) \quad (7)$$

基于 $\boldsymbol{y}_d$，计算新闻标题 d 的情感倾向（Sentiment Orientation）：

$$so_d = (1,-1) \cdot \boldsymbol{y}_d \quad (8)$$

式中：$so \in [-1,1]$，当 $so_d > 0$，情感倾向为积极（正向），否则情感消极（负向）。

设在日期 T_i 一共有 m_i 条新闻，每条新闻的情感极性用 so_j 来表示，$j = 1,...,m_i$，$so_j \in [-1,1]$。则在 T_i 的新闻情感总分值如下计算：

$$S_i = \frac{1}{m_i} \sum_{j=1}^{m_i} so_j \quad (9)$$

将新闻情感向量 $\boldsymbol{S}=[S_1, S_2, ..., S_i, ...S_n]^{'}$ 作为一列特征添加到表3所示的新闻事件特征矩阵中，得到股票新闻特征矩阵。

3.4 股票走势预测

股票预测将财务特征矩阵、股票新闻特征矩阵按照日期合并，作为股票走势预测模型的输入。股票走势预测作为前向时序建模预测，不用考虑后向序列的影响，因此采用 LSTM 而非 Bi-LSTM 作为股票涨跌预测。

将财务特征矩阵、新闻特征矩阵直接拼接在一起作为LSTM模型的输入可能会导致梯度消失问题[12]，因此本文采用2个LSTM神经元，将财务特征矩阵和新闻特征矩阵分别输入到两个LSTM中，再将结果进行向量合并，输入到全连接神经网络中，最后输出涨跌结果。当采样间隔为 a 天，将根据过去 $t-a$ 至 $t-1$ 天的数据，预测第 t 天的股票涨跌。

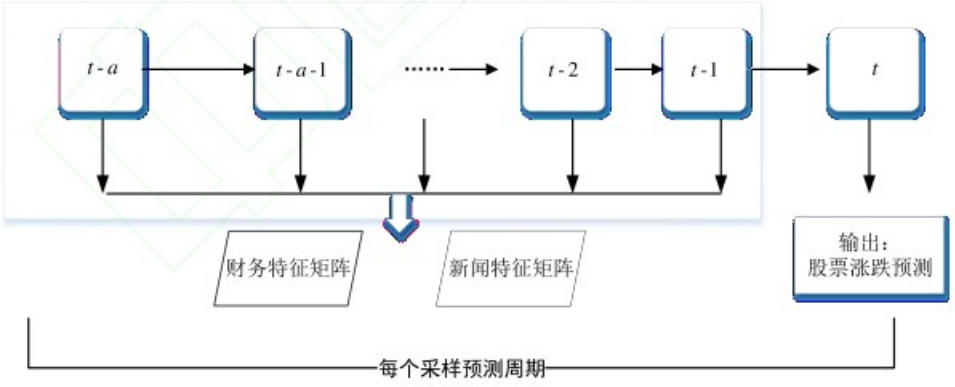


图3 股票预测模型采样周期示意图

Fig.3 Sampling period of stock trends prediction model

4 实验分析

4.1 实验数据集

为了评估所提模型对不同行业和板块的适用性，实验分别选取家用电器行业和通信行业的两支股票：格力电器（000651.SZ）和中兴通讯（000063.SZ）作为实验对象。家用电器行业属于周期性行业，受到国内或国际经济周期性波动的影

响较大。与此相反，通信行业属于防御性行业，受到经济周期性衰退和繁荣的影响较小，在股价表现上较为稳定。

实验一共选取了12个股票指标构成财务特征矩阵，分别是：开盘价、最低价、最高价、收盘价、主力资金净流入量、换手率、涨跌、涨跌幅、市盈率、市净率、深证A指、以及板块指数。

财务数值数据主要来源于 Wind 金融数据库。Wind 是国内以金融证券数据为核心的大型金融工程和财经数据仓库，以科学的核查手段和先进的管理方法确保数据的准确性在 99.95%以上。在 Wind 数据库中抓取 2010-02-02 到 2020-02-28 十年间，格力电器共 18766 条新闻数据，29352 条财务数据；中兴通讯共 18796 条新闻数据，29364 条财务数据。新闻事件分类器的标注数据来源为国泰安经济金融数据库中的新闻，一共对 27800 条新闻数据进行事件标注，用于事件分类器模型的训练。

4.2 新闻事件和情感分类器实验结果

（1）参数设置和模型输入

新闻事件分类器和情感分析模型的性能受到词向量维度、卷积窗口大小、迭代次数、过滤器数量等因素影响。在实验中通过十折交叉验证评估模型表现，选定最合适的参数组合。在本文数据集上表现最好的参数组合如表4所示。表中Null表示不需要这个参数设置。

表 4 模型参数设置

Table4 Parameter settings of the prediction model

参数	CNN	Bi-LSTM
词向量维度	300	300
卷积核个数	96	Null
卷积核大小	3,4,5	Null
Dropout	0.5	0.5
Batch_size	128	128
迭代次数	10	20
标题截取长度	Null	15
单层 LSTM 神经元个数	Null	[256,256]

（2）评价指标

计算精确率（Precision）、召回率（Recall）以及 $F1$ 值($F1$-measure)作为实验评价指标。第 i 类新闻事件分类的精确率、召回率和值分别用 Pre_i 、R_i 和 $F1_i$ 表示，计算如下：

$$Pre_i = \frac{TP_i}{(TP_i+FP_i)} \qquad (10)$$

$$R_i = \frac{TP_i}{(TP_i + \mathrm{FN}_i)} \qquad (11)$$

$$F1_i = \frac{2 \times Pre_i \times R_i}{Pre_i + R_i} \qquad (12)$$

式中：TP_i为被正确分类到第i类的样本数量，FP_i为被错误的分到第i类的样本数量，FN_i为原本属于第i类，但是被分到其他类别的样本数量。

（3）实验结果

表5列举了新闻事件分类的实验结果，为了验证基于CNN的新闻事件分类器的性能，将其与SVM算法以及基于最大熵（Maximum Entropy，Maxent）算法的新闻事件分类性能做比较。SVM在传统的机器学习算法中表现优异，而Maxent是线性对数模型，具有较好的分类能力。实验将新闻事件数据集90%的数据共25020条新闻作为训练集合，10%的数据共2780条新闻用于测试集。

表5 新闻事件分类精确率对比

Table5 Comparisons of experiment results on news event classification

	SVM	Maxent	CNN
训练集	90.78%	72.02%	93%
测试集	85.23%	69.42%	87.7%

基于CNN新闻事件分类器在训练集精确率达93%，测试集精确率达87.7%，优于基于SVM算法和基于Maxent算法的新闻事件分类精确率。为了分析基于CNN的新闻事件分类器对不同新闻事件的分类效果。表6列举了各类新闻事件的准确率、召回率和$F1$值，由于篇幅关系仅列出部分的事件类型。高召回率和高$F1$值表示新闻事件预测的覆盖率和准确率都比较好。从表6可以看出高召回率和高$F1$值的事件类型大多属于公司公告类，公司公告的内容在不同公司之间的差异性不大，公告具有一定的模板性，分类效果较好。分类性能较差的新闻事件大部分是对价格走势的预测，走势好或者走势不好在不同行业、不同公司之间新闻的内容差别比较大，因此识别起来准确率较低。

表6 新闻事件分类的性能统计表

Table6 Performance statistic of news event classification

分类性能较好的新闻事件类型示例				分类性能较差的新闻事件类型示例			
新闻事件	精确率	召回率	$F1$值	新闻事件	精确率	召回率	$F1$值
登上龙虎榜	1.00	1.00	1.00	业绩下降	0.64	0.58	0.61
停牌	0.98	1.00	0.99	政策利好	0.81	0.65	0.72
工商变更	1.00	1.00	1.00	资本变更	1	0.22	0.36
中标	1.00	1.00	1.00	聘请高管	0.50	0.40	0.44
可转债	0.97	0.97	0.97	业绩增长	0.68	0.73	0.71

质押	1.00	1.00	1.00	预计下滑	0.67	0.61	0.64
交易所问询	0.94	1.00	0.97	利差消息	0.42	0.47	0.44
退市	1.00	1.00	1.00	利好消息	0.46	0.65	0.54

表7为新闻情感分类的实验结果。基于Bi-LSTM的新闻情感分类在训练集上的精确率为99%，在测试集上精确率为91%。基于SVM的新闻情感分类性能次之，在训练集和测试集上的精确率分别为86.64%和81.13%；基于Maxnet的新闻情感分类效果最差。利用训练得到的新闻情感分类模型分别对格力电器和中兴通讯的新闻数据集进行情感分类、计算新闻情感分值，生成得到新闻的情感向量矩阵。

表7 新闻情感分类精确率对比

Table7 Comparisons of experiment results on news sentiment classification

	SVM	Maxent	Bi-LSTM
训练集	86..6%	82.8%	99.0%
测试集	81.1%	76.1%	91.0%

4.3股票走势预测结果

为了验证引入新闻信息后对股票预测模型性能的提高，实验将对比测试以下算法的性能：（1）未引入新闻特征的LSTM：仅利用股票财务特征进行股票走势预测；（2）引入新闻事件的LSTM：利用股票财务特征和表3所示的新闻事件特征进行股票走势预测；（3）新闻事件/情感融合的LSTM：利用股票财务特征、新闻事件特征、以及情感特征行股票走势预测。（4）新闻事件/情感融合的GBDT：基于梯度提升决策树（Gradient Boosting Decision Tree, GBDT）模型[34]进行股票预测，输入与所提模型（3）相同，对比LSTM和GBDT在相同输入条件下的股票预测性能。GBDT是一种泛化性很强的决策树算法，通过多次迭代训练多个回归树弱分类器，能有效避免过拟合、抵抗噪音。

设涨跌幅阈值为1%，即：当涨跌幅高于阈值0.01时，判定为上涨样本，用数字1表示，当涨跌幅低于阈值0.01时，判定为下跌样本，用数字0表示。模型的采样间隔a=14，输入第$t-13$天至第$t-1$天的数据，输出第t天的涨跌结果。不同模型的股票走势预测准确率如表8所示。

表8 不同模型的股票走势预测精确率对比

Table8 Comparisons of experiment results on stock trends prediction

股票	采用财务特征的 LSTM	引入新闻事件的 LSTM	新闻事件/情感融合的 GBDT	新闻事件/情感融合的 LSTM
格力电器	0.6998	0.7545	0.6257	**0.7812**
中兴通讯	0.6467	0.7851	0.6545	**0.8127**

由表8可知，相对比仅采用财务特征的LSTM模型，引入新闻事件的LSTM以及新闻事件/情感融合的LSTM模型在股票预测走势准确率上都有提高，格力电器个股的预测准确率从0.6998，提高到了0.7545和0.7812，分别提高了7%和11%；中兴通讯个股的预测准确率从0.6467提高到了0.7851和0.8127，分别提高了21%和22%。新闻短文本数据所包含的定性信息对股价走势预测有正面影响，其中事件特征相比于情感特征，影响更大。进一步印证了金融新闻文本中情感较为隐晦

的特点。相对于GBDT模型，LSTM模型在股价走势预测上更具优势，准确率得到一定提升。

同时，观察模型对于行业和模块的适用性，在仅采用财务特征矩阵的LSTM模型上，家电行业相对于通信行业表现更好，然而在融合新闻和财务特征的LSTM模型上，通信行业相对于家电行业表现更好。因此，所提模型对通信行业的适用性更高一些，也符合防御性行业相对于周期性强的行业股价更为稳定的预期。

图4 格力电器（000651.SZ）走势预测示例图

Fig.4 Example of individual stock trends prediction on GREE electric appliance (000651.SZ)

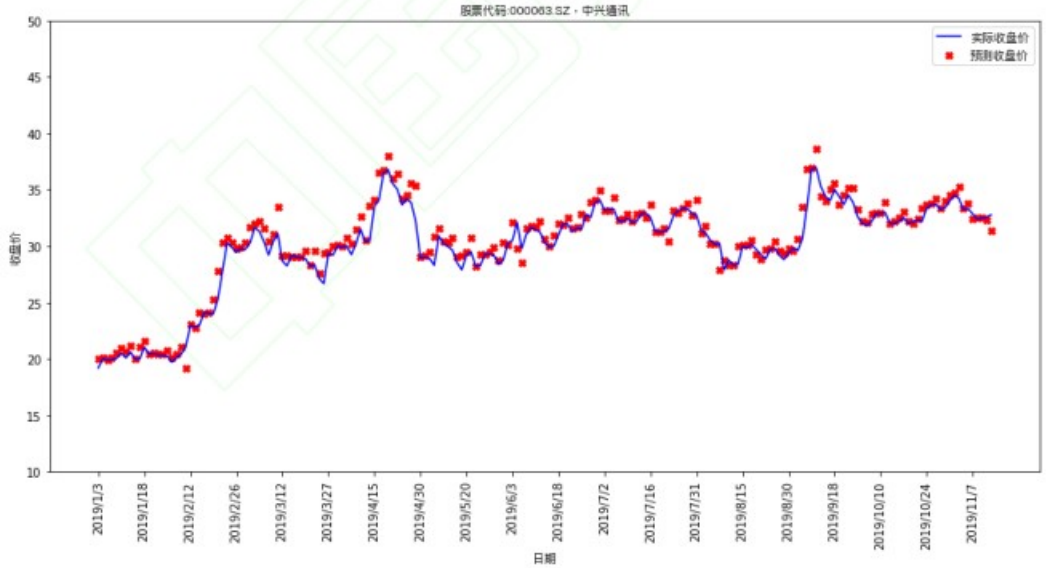


图5 中兴通讯（000063.SZ）走势预测示例图

Fig.5 Example of individual stock trends prediction on ZTE communication (000063.SZ)

为了直观展示所提模型在预测股价走势的效果，图4和图5分别描绘了所提模型在格力电器和中兴通讯两支股票在2019年的预测结果。图中，蓝色实线为实际收盘价，红色符号为预测收盘价。由于股价预测为二分类问题，输出结果为上涨或者下跌。为了能呈现股票走势预测效果，假设采样间隔为13天，若第14天预测

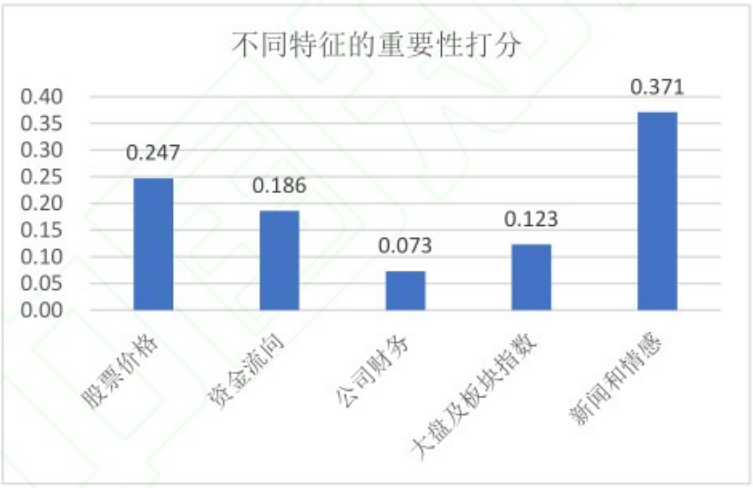
结果为上涨，则预测收盘价设置为第13天实际收盘价与第13天和第14天的实际收盘价差值之和，反之，设置为第13天实际收盘价与第13天和第14天的实际收盘价差值之差。从图4看出对个股走势的预测结果和实际情况的走势基本一致，说明该模型在预测个股走势方面具有较好的性能。

4.4 股票走势的影响因素分析

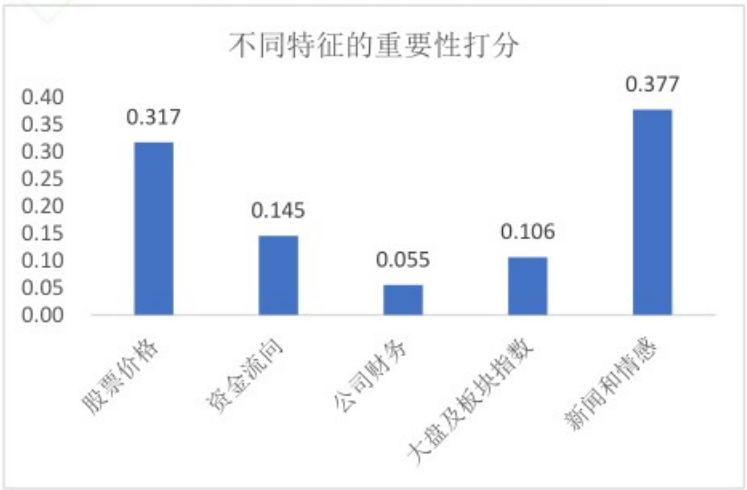
（1）影响股票走势的特征重要性分析

本小节讨论不同特征对股票走势预测模型的影响。GBDT 模型在个股走势预测上的效果次于所提模型，但是其自带的特征排序功能可对输入的特征组合进行重要性打分。

图 6 展示了 GBDT 对格力电器和中兴通讯两个股票数据的不同特征进行重要性打分的结果。由于输入特征的维度较多，这里将 12 个股票的财务特征归纳为 5 类：股票价格特征的分数等于开盘价、收盘价、最高价、最低价、涨跌、涨跌幅的重要性分数之和；资金流向特征的分数等于主力资金净流入额、换手率的重要性分数之和；公司财务特征的分数等于市净率和市盈率的重要性分数之和；大盘及板块指数特征分数等于深证指数和板块指数的重要性分数之和，新闻和情感特征分数等于新闻事件以及新闻情感特征的重要性分数之和。



（a）格力电器不同特征重要性曲线



（b）中兴通讯不同特征重要性曲线

图6 GBDT对不同特征的重要性排序结果

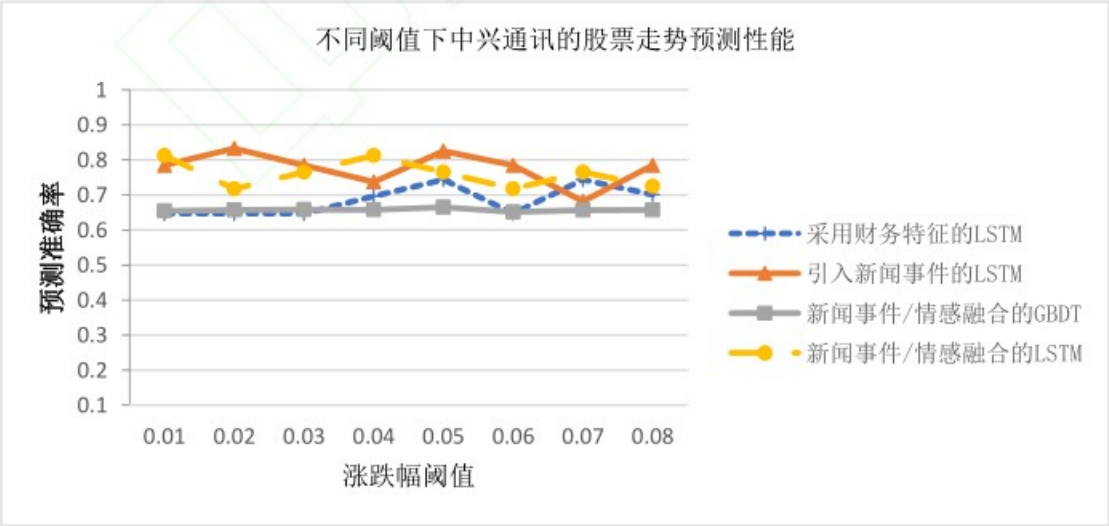
Fig.6 Importance ranking of different features based on GBDT

可以看出，格力电器的新闻事件和情感特征的重要性为 0.371，财务数值特征的重要性为 0.639，在影响股票走势预测上历史财务数据仍是主要的参考量，然而媒体新闻的重要性也不可小觑。在所有的财务数值特征中，股票价格的重要性最大，为 0.247；公司财务特征的重要性排名最后，为 0.073。通过分析股票财务数值的源数据发现，股票的市净率和市盈率在季度期间没有太大起伏，不会对股价波动产生明显影响，因此财务特征的重要性最低，与实际吻合。

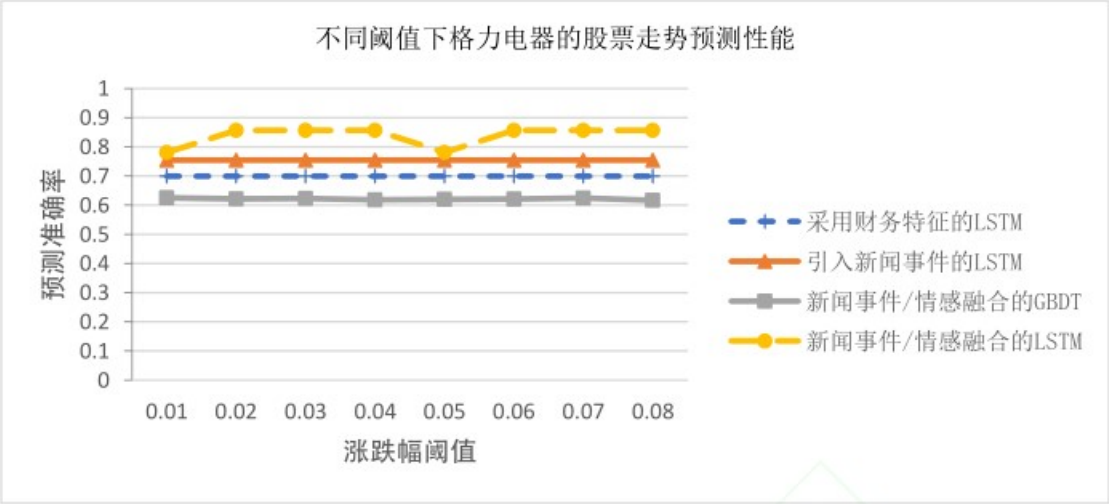
（2）涨跌幅阈值分析

这里讨论涨跌幅阈值对预测模型的影响。涨跌幅阈值的取值影响对股票上涨或者下跌的标注，从而影响预测模型的性能。实验中分别设置涨跌幅阈值从 1%到 8%，以 0.01 为步长区间，测试不同模型在格力电器和中兴通讯两支股票上的预测效果。实验结果如图 7 所示。

实验结果有几点发现：首先，随着阈值取值从 0.01 增加到 0.08，模型的预测准确率并没有线性的增加或者降低，而是随机的波动。阈值的取值会影响股票价格涨跌幅在 0.01-0.08 区间数据的标注，进而影响预测模型的准确率。阈值的设置是为了消除这部分数据的抖动对模型的影响，具体的取值可以结合经验以及实验测试得到。其次，在所有阈值取值的情况下，引入新闻事件的 LSTM 和新闻事件/情感融合的 LSTM 的性能均优于采用财务特征的 LSTM 算法，验证了在股票走势预测模型中引入媒体新闻特征能够提高股票预测准确率。最后，对比图 7(a)和图 7(b)发现，随着涨跌幅阈值的变化，不同模型对格力电器股票走势预测的性能基本保持不变。通过检查股票财务数值的源数据发现，格力电器的股票涨跌幅在 0.01 至 0.08 之间的数据仅有 1.5%，占比较低，这部分数据的标注改变对模型的预测性能差别不大；而中兴通讯的股票涨跌幅在 0.01 至 0.08 之间的数据有 5%，因此阈值的取值变化对模型的预测性能影响更大一些。



（a）中兴通讯股票走势预测性能随着阈值变化曲线



（b）格力电器股票走势预测性能随着阈值变化曲线

图7 涨跌幅阈值对股票走势预测模型的影响

Fig.7 Impact of threshold value on stock trend prediction

5 结论

本文提出一种基于 CNN-BiLSTM 多特征融合的股票走势预测模型，通过挖掘媒体新闻信息中的新闻事件和情感倾向，作为影响股票走势的市场信号融入到股票预测模型中提高预测准确率。实验选取来自家用电器行业和通信行业的两支股票作为对象验证所提模型的效果。结果表明，与现有算法比较，所提模型在家用电器和通信行业这两支股票的预测准确率相比现有算法分别提高了 11%和22%。所提模型对通信行业的适用性更高一些，也符合防御性行业（如通信行业）与周期性强的行业（如家用电器行业）相比，股价更为稳定的预期。本文还研究了影响股票走势预测模型的特征重要性。分析表明，股票价格特征的重要性最大，新闻特征次之，公司财务特征重要性最小。

本文通过媒体新闻事件和情感倾向的非数值股票信号挖掘，优化了现有的股票预测模型。通过融合股票更多的相关控制变量提高股票预测准确率作为研究的发展方向，未来工作可以考虑在以下两个方面开展：一是研究提高预测精度对模型的影响，现有模型主要是进行涨跌二元的分类，考虑到股票投资中涨幅大小对投资获利影响差别较大，未来可以研究预测精准度细分为四类时（小涨、大涨、小跌和大跌）对模型的影响。二是在预测模型中融入更多股票相关的控制变量，例如收益率、崩盘风险等，并比较这些控制变量对股票预测性能的影响。

参考文献

[1] Sun X Q, Shen H W, Cheng X Q. Trading network predicts stock price[J]. Scientific Reports, 2015,4:1-6.

[2] Adebiyi A A, Adewumi A O, Ayo C K. Comparison of ARIMA and acritical neural networks models for stock price prediction[J]. Journal of Applied Mathematics,2014:1-7.

[3] Birz G. Stale Economic News, Media and the Stock Market[J]. Journal of Economic Psychology, 2017,61(3):384-412.

[4] Chong E, Han C, Park F C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies[J]. Expert Systems with Applications, 2017, 83:187-205.

[5] Akita R, Yoshihara A, Matsubara T. Deep learning for stock prediction using numerical and textual information[C]// Proceeding of IEEE/ACIS International Conference on Computer and Information Science, Okayama, IEEE, 2016:978-984.
[6] Usmani M, Adil S H, Raza K. Stock market prediction using machine learning classifiers and social media, news[C]// Proceeding of International Conference on Computer and Information Sciences, Kuala Lumpur IEEE, 2016:322-327.
[7] Hajek P. Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns[J]. Neural Computing and Applications, 2018, 29(7):343-358.
[8] Li J, Lin Y F. Time series data prediction based on multi-time scale RNN. Comput. Appl. Software Shanghai, 2018:35(62): 33-37.
[9] Zhang Y S, Yang S T. Prediction on the highest price of the stock based on PSO-LSTM Neural Network[C]. International Conference on Electronic Information Technology and Computer Engineering (EITCE). Xiamen, China: IEEE. 2019:1565-1569.
[10] Althelaya K A, El-Alfy E, Mohammed S. Stock Market Forecast Using Multivariate Analysis with Bidirectional and Stacked (LSTM, GRU) [C]. Computer Society National Computer Conference (NCC). 2018. Saudi Arabia: IEEE. 2018:1301-1307.
[11] Zhao Z, Rao R, Tu S, et al. Time-Weighted LSTM Model with Redefined Labeling for Stock Trend Prediction[C]. International Conference on Tools with Artificial Intelligence. Boston, MA, USA : IEEE. 2017:1210-1217.
[12]孔翔宇,毕秀春,张曙光,财经新闻与股市预测—基于数据挖掘技术的实证分析,财经新闻与股市预测[J].数理统计与管理,2016,35(2):215-224.
Kong X Y, Bi X C, Zhang S G, Financial news and prediction for stock market: an empirical analysis based on data mining techniques[J].Journal of applied statistics and management. ,2016,35(2):215-224 (In Chinese).
[13] Oncharoen P, Vateekul P. Deep Learning for Stock Market Prediction Using Event Embedding and Technical Indicators[C]// Proceeding of International Conference on Advanced Informatics: Concept Theory and Applications, Krabi, IEEE, 2018:19-24.
[14] Tsai C F, Lin Y C, Yen D C, et al. Predicting stock returns by classifier ensembles[J]. Applied Soft Computing, 2011, 11(2):2452-2459.
[15] 张梦吉,杜婉钰,郑楠,引入新闻短文本的个股走势预测模型[J]. 数据分析与知识发现,2019,3(5):11-17.
Zhang M J, Du W Y, Zheng N, Predicting stock trends based on news event[J].Data Analysis and Knowledge discover,2019, 3(5):11-17(in Chinese).
[16] Luan Y D, Lin S F. Research on Text Classification Based on CNN and LSTM[C]// International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian: IEEE. 2019: 435-442.
[17] Ashima Y, Dinesh K V. Sentiment analysis using deep learning architectures: a review[J]. Artificial Intelligence Review (2020) 53:4335–4385.
[18] Rapach D, Zhou G. Forecasting stock returns[J]. Handbook of Economic Forecasting. 2013:328–383.
[19] Chen K, Zhou Y, Dai F. A LSTM-based method for stock returns prediction: A case study of China stock market[C] // Proceeding of IEEE International Conference on Big Data, Washington DC, IEEE,2015:2823-2824.
[20] Gandhmal D P, Kumar K. Systematic analysis and review of stock market prediction techniques[J]. Computer Science Review, 2019: 34-47.
[21]张玉川,张作泉,黄珍,支持向量机在选择优质股票中的应用[J].统计与决策, 2008, (4):163-165.
Zhang Y C, Zhang Z C,Huang Z, Application of support vector machine in selecting high quality stock[J].Statistics and decision, , 2008, (4):163-165 (In Chinese).

[22] Kara Y, Boyacioglu M A, Bayken K. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange[J]. Expert Systems with Applications, 2011, 38(5):5311-5319.
[23]威尔金森. 行为经济学[M]. 中国人民大学出版社, 2012.
[24] Bruce J V, Adrian G, Geoff H. Do news and sentiment play a role in stock price prediction? [J]. Applied Intelligence, 2019, 49: 3815–3820.
[25] Chen W, Zhang Y, Yeo C K, et al. Stock Market Prediction Using Neural Networks through News on Online Social Networks[C]// Proceeding of International Smart Cities Conference, Wuxi: IEEE.2017:23-29.
[26] Vargas M R, Dos Anjos C E M, Bichara G L G , et al. Deep learning for stock market prediction using technical indicators and financial news articles[C]// Proceeding of International Joint Conference on Neural Networks, Brazil: IEEE. 2018:1-8.
[27]岑咏华,谭志浩,吴承尧.财经媒介信息对股票市场的影响研究：基于情感分析的实证[J]. 数据分析与知识发现, 2019, 3(9):98-114.
Cen Y H, Tan Z H, Wu C R. Impact of financial media information on stock market: an empirical study of sentiment analysis[J].Data Analysis and Knowledge Discover,2019:3(9):98-114 (In Chinese).
[28] Zhao W T, Wu F, Fu Z Q, et al. Sentiment Analysis on Weibo Platform for Stock Prediction[C]//International Conference on Artificial Intelligence and Security (ICAIS), Hohhot, China: IEEE. 2020: 323-333.
[29] Basu S. The investment performance of common stocks in relation to their price/earnings ratio: a test of the efficient market hypothesis[J]. The Journal of Finance, 1977, 32(3):663-682.
[30] French K R, Fama E F. Size and book-to-market factors in earnings and returns[J]. The Journal of Finance, 1995, 50(1):131-155.
[31] Cheadle C, Vawter M P, Freed W J et al. Analysis of microarray data using Z score transformation[J]. The Journal of Molecular Diagnostics, 2003, 5(2):73-81.
[32]徐月梅,刘韫文,蔡连侨.基于深度融合特征的政务微博转发规模预测模型[J].数据分析与知识发现.2020,4(2/3):18-28.
Xu Y M, Liu Y W, Cai L Q. A Model of Retweeting Scale Prediction of Government Microblogs Based on Deep-combined Features[J]. Data Analysis and Knowledge Discover,2019:4(2/3):18-28 (In Chinese).
[33] Staudemeyer R C, Morris E R. Understanding LSTM-a tutorial into long short-term memory recurrent neural networks[J]. arXiv, 2019:1-42.
[34] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics, 2001:1189-1232.

基金项目：本文系北京外国语大学一流学科建设项目“基于语义神经网络的文本话题和情感分析研究与实现（编号：No.YY19ZZA012）研究成果之一。

作者贡献声明：

徐月梅：提出研究思路，设计研究方案，进行实验；
王子厚：设计研究方案，论文修订；
吴子歆：采集、清洗和分析数据。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储：https://github.com/Oscaria/stock-prediction.git

[1] 徐月梅，王子厚，吴子歆. Alldata_finance.csv.模型财务数据

[2] 徐月梅，王子厚，吴子歆. Alldata_news.csv.模型新闻数据

[3] 徐月梅，王子厚，吴子歆.Preprocess.py.数据预处理、分析代码.

[4] 徐月梅，王子厚，吴子歆.GetCategory.py. 新闻事件分类器代码

[5] 徐月梅，王子厚，吴子歆.GetSentiment.py. 新闻情感分析代码

[6] 徐月梅，王子厚，吴子歆. Word2vec.py. word2vec 模型训练程序

[7] 徐月梅，王子厚，吴子歆. Stock_prediction.py. 股票走势预测程序.

[8] 徐月梅，王子厚，吴子歆. GBDT.py. GBDT 对比实验测试程序.

[9] 徐月梅，王子厚，吴子歆. SVM.py. SVM 对比实验测试程序.

[10] 徐月梅，王子厚，吴子歆. Maxnet.py. 最大熵对比实验测试程序.