

基于LSTM网络和投资者情绪的A股大盘指数可预测性研究

曹雨辰, 林靖澄

贵州大学经济学院, 贵州 贵阳

收稿日期: 2025年4月8日; 录用日期: 2025年4月21日; 发布日期: 2025年5月21日

摘要

资本市场的股票价格走势深刻反映并影响着国家的宏观经济政策。随着国家对深度学习和机器学习在经济学问题中应用的推动, 将这些技术与经济学结合已成为当前的研究热点。近年来, 提取文本中的有效信息并研究其对股票价格波动的影响, 已引起学术界广泛关注。本文利用2023年12月4日至2024年12月2日的A股大盘指数数据, 基于财经新闻情感分析构建的投资者情绪指数, 提出了一种融合情感特征的股指预测模型。研究表明, 加入投资者情绪指标的预测模型在预测精度和预测值与真实值的接近度等方面显著优于其他机器学习模型, 投资者情绪指数呈现出比其他特征变量更强的特征重要性。将预测的股指进行替换后, 结论仍保持一致。

关键词

投资者情绪, 机器学习, LSTM, 大盘指数预测

Predictability of A-Share Market Index Based on LSTM Network and Investor Sentiment

Yuchen Cao, Jingcheng Lin

School of Economics, Guizhou University, Guiyang Guizhou

Received: Apr. 8th, 2025; accepted: Apr. 21st, 2025; published: May 21st, 2025

Abstract

The stock price trends in capital markets profoundly reflect and influence national macroeconomic policies. With the growing emphasis on the application of deep learning and machine learning in

文章引用: 曹雨辰, 林靖澄. 基于 LSTM 网络和投资者情绪的 A 股大盘指数可预测性研究[J]. 电子商务评论, 2025, 14(5): 1433-1445. DOI: 10.12677/ec.2025.1451423

economics, integrating these technologies with economic research has become a current focus. In recent years, extracting valuable information from texts and studying its impact on stock price fluctuations have attracted widespread academic attention. This paper utilizes A-share market index data from December 4, 2023, to December 2, 2024, and an investor sentiment index constructed based on financial news sentiment analysis. It proposes a stock index prediction model that incorporates sentiment features. The study shows that the model with the investor sentiment index significantly outperforms other machine learning models in terms of prediction accuracy and the closeness between predicted and actual values. Additionally, the investor sentiment index demonstrates stronger feature importance than other variables. The results remain consistent when the predicted stock index is replaced.

Keywords

Investor Sentiment, Machine Learning, LSTM, Market Index Prediction

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

国务院于 2017 年 7 月发布的《新一代人工智能发展规划》指出,“开展跨学科探索性研究,推动人工智能与神经科学、认知科学、量子科学、心理学、数学、经济学、社会学等相关基础学科的交叉融合”。作为人工智能的核心技术之一,机器学习与深度学习在近年来已经成为至关重要的技术。

股票价格的波动在金融领域中不仅影响投资者的决策,还能反映和影响国家宏观经济政策,因此受到了社会各界的关注。过去三十余年,中国资本市场发展迅速,现已成为全球第二大市场。尽管如此,依然面临着许多结构性问题,尤其是投资者结构的不均衡。在股票市场中,个人投资者占比相对较大,相比机构投资者,他们通常偏好交易小市值股票、价格较低的股票以及业绩不佳的股票,并且持股周期较短、交易频繁,容易受到市场波动的影响。因此,市场参与者和监管机构迫切需要及时、全面地掌握市场信息,以便做出准确的股价预测,并进行有效的逆周期调节。

然而,股票市场数据作为金融时间序列,通常表现出非线性、非平稳以及高噪声等特点,这使得股票指数的预测难度较大,效果常常不理想。此外,按照半强势有效市场假说(Fama, 1970) [1],大量关键数据可能存在于其他文本数据中,非结构化文本的分析进一步增加了金融时间序列预测的复杂性。计量经济学模型通常假设较强且参数设置不灵活,这限制了其在金融时间序列中的泛化能力,传统的经济学方法难以捕捉这些数据中的非线性特征。相较之下,机器学习方法能够从数据中自我学习,无需依赖预先设定的模型,具备自动提取特征的能力,因此更适合用于解决股票价格预测中的挑战。

深度学习技术在过去几年里持续发展,已成为最前沿的机器学习方法。它能够根据数据的内在特征进行自动学习,无需人工特征设计,且具有更强的表达能力,因此预测效果更为优越。这些技术进展为构建更有效的训练框架提供了基础,而长短期记忆(LSTM)网络正是在这种环境下被提出的(Hochreiter and Schmidhuber, 1997) [2]。

2. 文献回顾

2.1. 利用机器学习进行金融市场预测

在股票市场预测研究中,国内外学者大量采用机器学习技术,并取得了显著的进展。本文概述了深

度学习在金融时间序列分析中的三个主要优势: 首先, 它不受维度限制, 能够处理不同类型的相关数据; 其次, 深度学习具有强大的非线性拟合能力, 能够有效适应金融数据的特性; 此外, 它还能够减轻过拟合问题, 避免陷入局部最优解(Heaton *et al.*, 2017) [3]。因此, 近年来, 股票预测的研究逐渐从传统机器学习技术, 如支持向量机(SVM)和随机森林(RF), 转向深度学习方法, 特别是卷积神经网络(CNN)和循环神经网络(RNN)。Wei and Chaudhary (2018)提出了一种集成人工选择、分割算法和训练误差反馈的综合方法(TST 方法), 通过 RNN 识别股票价格时间序列的趋势, 从而显著提高了预测效果[4]。然而, RNN 模型存在梯度消失和梯度爆炸的问题, 导致其难以有效捕捉长期依赖关系。为了应对这一挑战, LSTM 通过引入门控机制, 显著缓解了梯度消失问题, 从而更好地学习金融时间序列中的长期规律。

LSTM 在股价预测中已获得广泛应用, 许多学者进行过相关研究。例如, Akita *et al.* (2016)分析了深度学习模型、段落向量及 LSTM 在金融时间序列预测中的效果, 并通过东京证券交易所 50 家上市公司数据进行验证, 研究结果表明, 基于文本的分布式表示优于传统的数值数据处理方法[5]。Wei *et al.* (2017)提出了一种创新的深度学习框架, 利用层次化结构方程首次应用于股票价格预测, 结合 6 个市场指数及期货数据进行测试, 结果显示该方法的预测精度和盈利性优于其他模型[6]。杨青与王晨蔚(2019)采用 LSTM、SVR、MLP、ARIMA 等多种模型对全球 30 个股票指数进行了预测, 结果表明, LSTM 模型在不同指数及各预测期限的精度和稳定性方面优于其他方法[7]。为了增强模型的泛化能力, 陈卫华与徐国祥(2018)通过 EEMD 方法将价格序列分解为本征模态分量(IMF)和趋势项, 并结合 Adaboost 与 LSTM 对各分量逐一预测, 最后通过加权平均得到股票价格的预测结果[8]。

2.2. 文本分析应用于股指预测

传统的金融实证分析多聚焦于财务报表和股票市场等结构化数据。随着自然语言处理(NLP)技术的不断发展, 文本数据的收集与分析效率得到了显著提升。文本分析是通过对文本进行深入挖掘与数据分析, 识别作者的立场、观点、价值观及其背后潜在的意图与目的。该技术在计算语言学、信息检索、内容分析以及文体学等多个领域得到了广泛应用[9]。在股价预测的研究中, 分析上市公司发布的文本信息及财经新闻中的金融市场内容, 能够帮助研究者更好地理解 and 揭示资产价格波动的规律与机制。

近年来, 情绪分析在经济学领域的应用越来越广泛。德温特资本市场公司通过分析全球 3.4 亿微博用户的情绪数据, 帮助指导股票交易, 并最终获得了 7%的收益率。Bing *et al.* (2014)采用数据挖掘方法, 研究了如何利用 1500 万条推特数据预测纳斯达克和纽约证券交易所 30 家公司的股票价格, 结果显示预测准确率为 76.12% [10]。Martin (2013)对信息进行了情绪和主观性分析, 并使用法国推特数据训练了一个基本的神经网络模型, 最终实现了 80%的方向性准确率和 2.97%的平均绝对百分比误差[11]。Khatri and Srivastava (2016)则通过推特数据进行情绪分析, 将情绪划分为四类, 并结合极性指数及市场数据, 利用人工神经网络(ANN)进行预测[12]。龙文等(2019)采用线性判别分析方法研究了 A 股市场中的“媒体效应”, 结果表明他们结合财经新闻建立的模型能够实现超额收益[13]。

相关文献表明, LSTM 模型在处理时序数据方面表现优异, 而财经新闻作为信息传播媒介, 具有较强的时效性和真实性, 能够及时反映与股票相关公司的实际动态, 因而非常适合用于股价预测。因此, 本文将在财经新闻情感分析的基础上构建投资者情绪指数, 并结合 LSTM 模型提出一种融合情感分析特征的股指预测模型。

3. 变量说明及数据来源

3.1. 变量说明

本文所使用的变量如表 1 所示。要预测的目标变量是中国 A 股上证指数(股票代码 000001.SH)的收

盘价, 替代的目标变量为 A 股指数, 参考杨青(2019)的做法[7], 所选特征变量包括两大指数的开盘价、最高价、最低价、昨收价、涨跌幅、成交额(按百万计)和成交量(按百万股计)。A 股指数与上证指数的主要区别在于, A 股指数不仅包括上海证券市场的上证综合指数, 还涵盖了深圳证券市场的深证成分股指数, 因而其成分股既包括沪市股票, 也包括深市股票。相较之下, 上证指数仅代表上海市场的上证综合指数, 成分股仅限于上海市场。本文主要使用的是上证指数相关数据, A 股指数的数据则用于验证。本文使用的数据范围是 2023 年 12 月 4 日到 2024 年 12 月 2 日的日度数据。

此外, 本文主要研究的目标变量是投资者情绪指数。开盘价、最高价等股指基础数据对股指指数收盘价的预测作用更大, 还是投资者情绪指数的预测作用更优, 是本文主要考究的问题。

Table 1. Variable declaration
表 1. 变量说明

变量名称	
主要特征变量	投资者情绪指数
特征变量	开盘价
	最高价
	最低价
	昨收
	涨跌幅
	成交额/百万
预测变量	成交量/百万股
	上证指数(收盘价)
替代预测变量	A 股指数(收盘价)

3.2. 数据来源

本文使用的投资者情绪指数来源于由数库科技构建的“中国 A 股市场情绪指数”(Chinascope, 2024) [14]。该指数是数据库基于多年积累的知识库和前沿 NLP 算法, 对每日超过数十万篇的财经新闻进行情绪识别, 基于这些海量的新闻分析数据, 联合 J.P. Morgan 亚太量化研究团队, 构建的投资者情绪指数。其余数据来源于 WIND 数据库。

3.3. 描述性统计

变量的描述性统计结果如表 2 所示。

Table 2. Descriptive statistics
表 2. 描述性统计

变量		数量	均值	标准差	最小值
主要特征变量	投资者情绪指数	239	1.013	0.043	0.913
上证指数特征变量	上证指数开盘价	239	3013.899	168.189	2680.48
	上证指数最高价	239	3036.356	172.543	2724.11
	上证指数最低价	239	2995.438	166.372	2635.09
	上证指数昨收	239	3016.444	167.894	2702.19

续表

	上证指数涨跌幅	239	0.001	0.012	-0.0306
	上证指数成交额/百万	239	411831.72	184282.42	207457.55
	上证指数成交量/百万股	239	38586.931	15652.202	21788.518
预测变量	上证指数收盘价	239	3017.834	169.39	2702.19
A 股指数特征变量	A 股指数开盘价	239	3159.391	176.251	2810.07
	A 股指数最高价	239	3182.944	180.815	2855.03
	A 股指数最低价	239	3140.031	174.35	2762.5
	A 股指数昨收	239	3161.722	175.368	2832.82
	A 股指数涨跌幅	239	0.001	0.012	-0.031
	A 股指数成交额/百万	239	411329.96	184107.99	207070.47
	A 股指数成交量/百万股	239	38538.321	15631.133	21753.053
替代预测变量	A 股指数收盘价	239	3163.518	177.511	2832.82

如图 1 所示，投资者情绪指数与大盘指数的变化趋势较为相似，且情绪指数对指数价格的影响具有一定的时间延迟。因此，将市场情绪指标融入指数预测模型，有可能提升模型的预测精度。

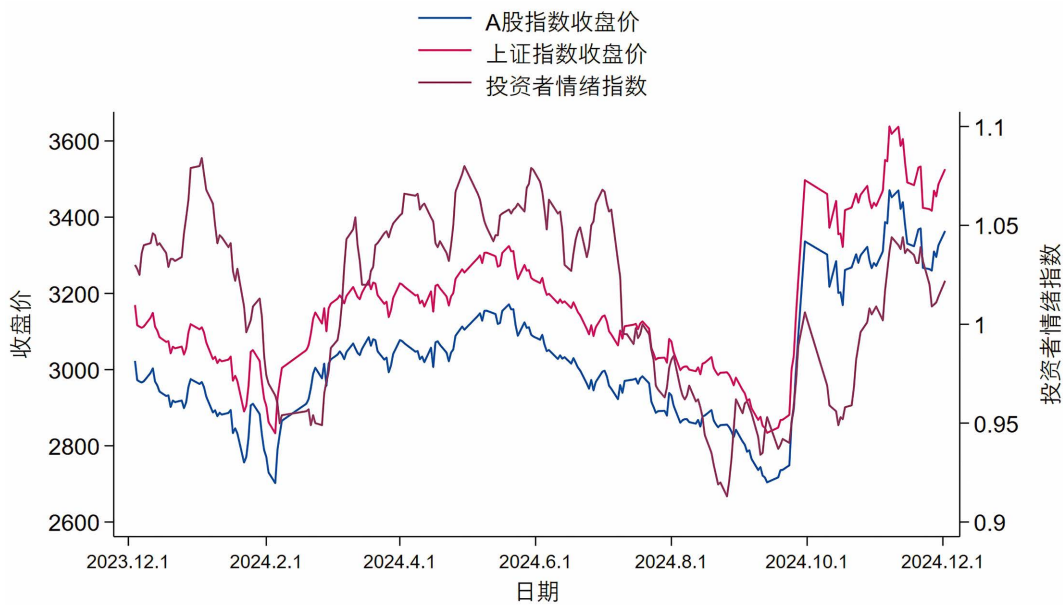


Figure 1. The trend of the investor sentiment index and the market index (closing price)
图 1. 投资者情绪指数与大盘指数(收盘价)走势

4. 实证分析

4.1. LSTM 模型参数选择

调整深度学习模型的超参数是提高预测准确度的重要途径，不同的超参数配置对模型性能具有显著影响。

在 LSTM 模型的优化过程中，本文主要调节了以下几个参数：首先，隐藏层神经元的数量，它对模

型的表达能力有重要作用。其次，Dropout 正则化中的丢弃率，用于减少过拟合现象，这一参数控制在训练过程中随机丢弃神经元的比例。第三，激活函数，它决定了隐藏层神经元的输出形式，并且对模型的非线性表达能力产生影响。本模型在进行非线性变换时采用了 tanh 激活函数，即双曲正切函数。第四，优化器，它负责更新模型权重以最小化损失函数。本文选用的 Adam 优化器，与其他自适应学习率算法相比，Adam 算法收敛速度更快、学习效果更为有效。最后，学习率是控制优化器更新权重时步长的关键参数，较大的学习率可能导致收敛不稳定，而较小的学习率可能使训练过程过长。

通过调参，本文最终选出了给定超参数搜索空间内最佳的超参数组合，验证集上的损失值为 0.00333，说明该参数下的 LSTM 模型性能较好。实验表明，第一，模型在不同的隐藏层配置下进行训练，以找到最优的隐藏层神经元。当隐藏层神经元设为 100 时，模型的预测效果最佳。第二，为了防止过拟合，训练过程中随机丢弃 10% 的神经元，丢弃率设为 0.1。第三，优化器的学习率为 0.01，优化器更新权重时步长适中，能平衡收敛速度与稳定性。相关参数见表 3。

Table 3. Correlation parameter
表 3. 相关参数

参数	搜索空间	最优值
神经元数量	[50, 100, 150]	100
Dropout 率	[0.1, 0.2, 0.3]	0.1
优化器的学习率	[0.001, 0.01, 0.1]	0.01

4.2. 未加入投资者情绪指标的预测结果

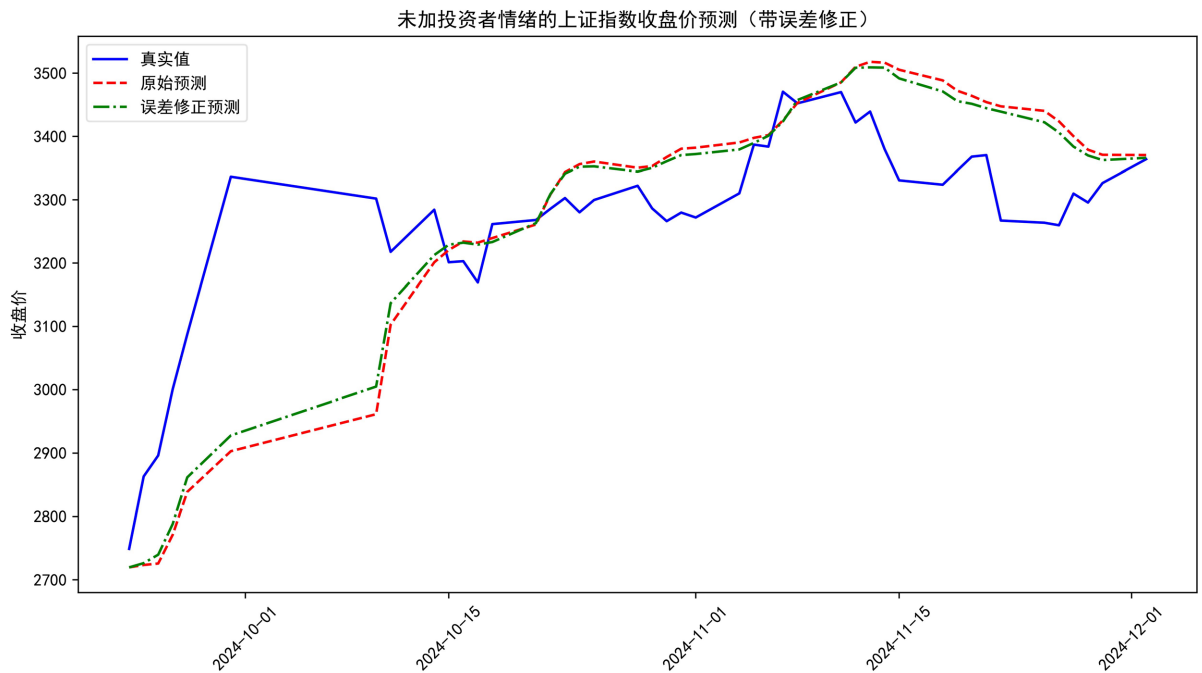


Figure 2. Shanghai composite index closing price forecast without investor sentiment (correction for error)
图 2. 未加投资者情绪的上证指数收盘价预测(带误差修正)

本部分重点分析在不加入投资者情绪指标的情况下，利用 LSTM 模型对 A 股大盘指数收盘价进行预

测的有效性,以及对比不同模型与 LSTM 模型的预测效果。预测数据为上证指数收盘价,时间跨度为 2023 年 12 月 4 日至 2024 年 12 月 2 日。输入的特征变量为表 1 所示的上证指数部分的特征变量。如前文所述,本文使用的 LSTM 模型训练集和测试集划分比例为 8:2,通过测试集评估各种模型对于样本外序列的预测能力。

图 2 展示了在不考虑投资者情绪指数的情况下对上证指数收盘价的预测。蓝色曲线代表实际收盘价,红色曲线为 LSTM 模型的初步预测,绿色曲线则表示在误差修正后得到的优化预测。从图中可见,红色曲线与蓝色曲线的趋势高度吻合,有效地捕捉了市场的变化趋势。例如 10 月上旬指数上升阶段,红线能基本跟随蓝线的增长趋势;11 月中旬之后的下降阶段,红线同样捕捉到了下行的趋势。同时红色曲线相较真实值更平滑,表明 LSTM 在处理高频波动时存在一定滞后性,难以拟合短周期波动特征。

修正后的绿色线条在趋势上显著靠近真实值。例如:在 10 月中旬的快速上升阶段,绿色线条对涨幅的拟合优于红色线条;在 11 月初的高位震荡期,绿色线条较好地捕捉了震荡幅度,与蓝色线条更为一致。下降趋势的后半段,绿色线条能及时调整,贴近真实值,一定程度上避免了红色线条的滞后问题。可见,原始预测值中可能存在逐步累积的偏差,而引入历史误差修正后,绿色线条有效抑制了这一偏差累积,呈现出更稳定的预测结果。

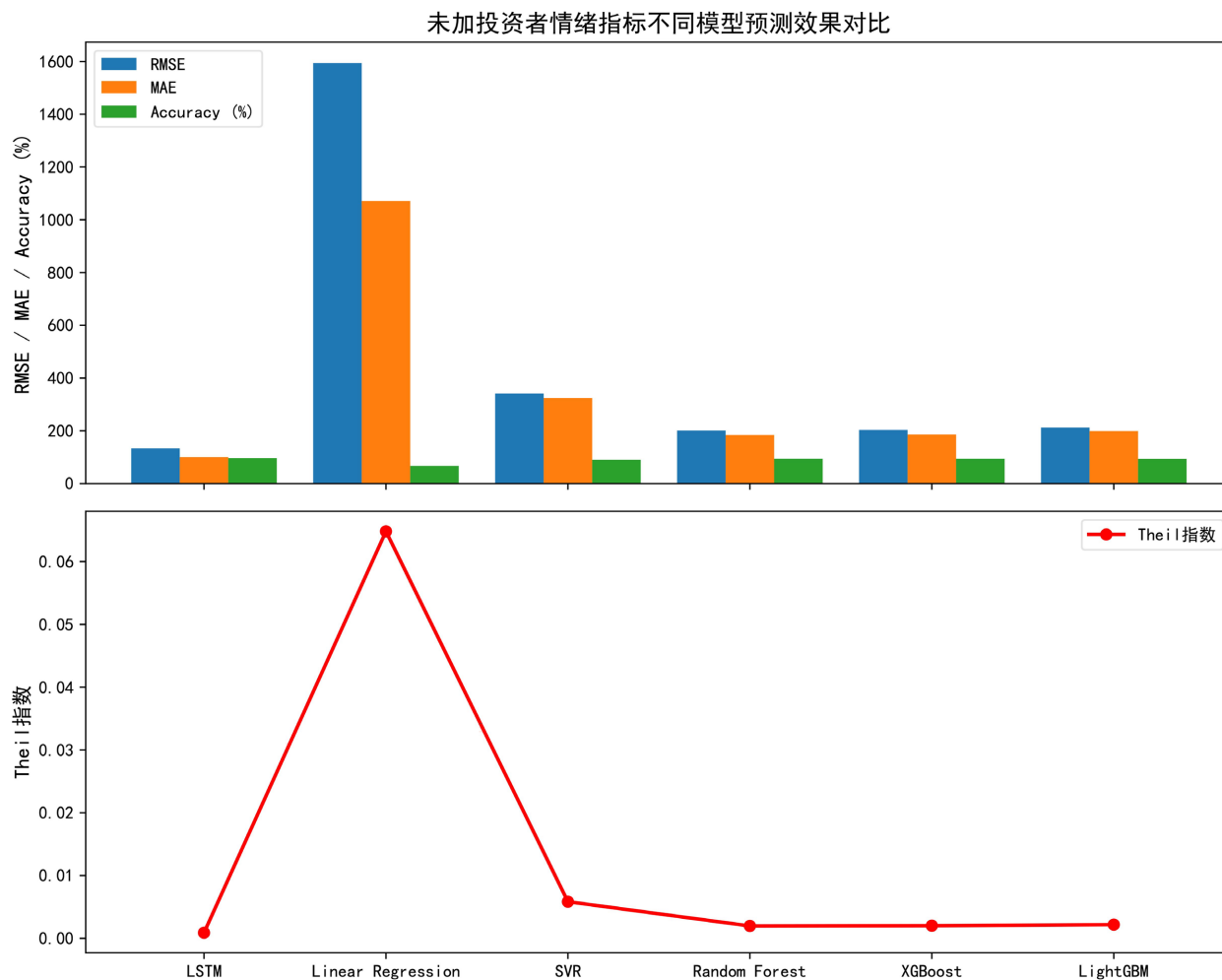


Figure 3. The comparison of prediction effect of different models which the investor sentiment index was not included

图 3. 未加入投资者情绪指标不同模型预测效果对比

Table 4. The prediction effect evaluation of LSTM model and other models before adding investor sentiment
表 4. 加入投资者情绪前 LSTM 模型与其他模型的预测效果评价

模型	RMSE	MAE	泰尔指数	皮尔森相关系数	涨跌预测准确率
LSTM (误差修正前)	130.3528	92.2627	0.0009	0.8364	97.15%
LSTM (误差修正后)	119.0156	83.8191	0.0007	0.8546	97.41%
线性回归	1593.828	1071.172	0.0648	0.1395	67.43%
SVR	341.0417	323.8644	0.0058	0.8552	90.23%
随机森林	201.4503	184.4403	0.0020	0.8380	94.42%
XGBoost	203.1623	186.1962	0.0021	0.8364	94.36%
LightGBM	212.5168	199.3136	0.0022	0.8929	93.98%

表 4 和图 3 展示了未加入投资者情绪指标的 LSTM 模型在误差修正前后的评价结果，以及使用线性回归、SVR 模型、随机森林模型、XGBoost 模型、LightGMB 模型进行预测的预测效果评价。评价指标方面，皮尔森相关系数反映了预测结果与真实值之间的变化相似程度，涨跌预测准确率反映了模型对序列变化趋势的预测能力，RMSE (均方根误差)和 MAE (平均绝对误差)用于衡量模型预测结果与实际值之间的差异，数值越小表明预测的准确性越高。泰尔指数衡量预测误差的相对大小，值越小说明误差相较实际值较小。

可以看出，在误差修正后的 LSTM 模型中，RMSE (119.02)和 MAE (83.82)均优于误差修正前的结果 (RMSE 为 130.35, MAE 为 92.26)。同时，误差修正后的模型泰尔指数得到下降，相关系数和涨跌预测准确率均得到提升。这一变化表明，误差修正方法显著提高了 LSTM 模型的预测精度，缩小了预测结果与实际值之间的偏差。虽然误差修正对涨跌预测准确率的影响较小，但对数值预测精度有显著改进，说明误差修正主要是针对数值拟合而非趋势预测。LSTM 模型展现了较强的时间序列建模能力，尤其适用于捕捉复杂的非线性关系和波动模式。再次证明引入历史误差修正后的 LSTM 模型呈现出更优的预测结果。

相比之下，LSTM 模型特别是误差修正后的 LSTM 模型在预测精度和趋势捕捉方面优于其他传统机器学习模型。线性回归模型的表现最差，在所有评估指标上均显著落后于其他模型，无法有效拟合复杂的非线性关系。SVR 模型在捕捉趋势方面表现较好，虽然精度不及 LSTM，但仍优于传统的回归模型。随机森林、XGBoost 和 LightGBM 这三种基于树的模型在精度和趋势预测方面表现较为接近，尤其在皮尔森相关系数和涨跌预测准确率上，显示出它们在捕捉市场波动的能力上具有竞争力。

4.3. 加入投资者情绪指标的预测结果

4.3.1. 预测结果分析

本部分研究在加入投资者情绪指标的情况下，利用 LSTM 模型对 A 股大盘指数收盘价进行预测的有效性，比较多种模型与 LSTM 模型在预测中的表现，并分析各个特征变量的重要性。数据集的划分比例及时间范围与前一部分一致。

图 4 展示了在考虑投资者情绪指数的情况下，上证指数收盘价的预测情况。与前文一致，蓝色曲线代表实际的上证指数收盘价，红色曲线显示 LSTM 模型的初步预测，绿色曲线则为通过误差修正方法优化后的结果。可见，在误差修正后，绿色线条明显更接近蓝色线条(真实值)，尤其是在局部波动较大的时段如 10 月中旬的转折点和 11 月初的调整期。具体而言，误差修正通过引入前期预测误差进行调整，使得修正后的绿色曲线在震荡阶段的拟合能力明显增强，较好地跟踪了真实值的波动。此外，在趋势转折点处，绿色线条对趋势反转的响应也更加灵敏，避免了红线在这些时刻的滞后效应。

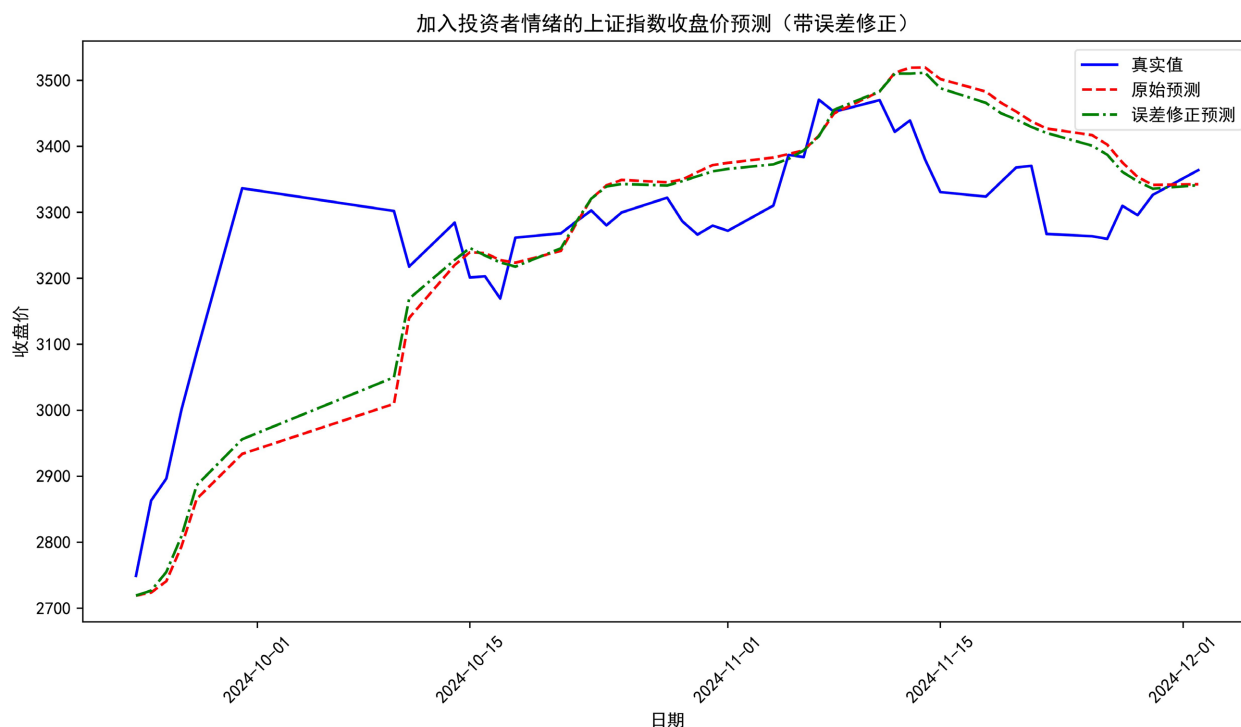


Figure 4. Shanghai composite closing price forecast with added investor sentiment (correction for error)

图 4. 加入投资者情绪的上证指数收盘价预测(带误差修正)

表 5 和图 5 展示了加入投资者情绪指标后的 LSTM 模型在误差修正前后的评价结果, 以及与其他不同模型的预测效果对比。

可以看出, 从预测精度(RMSE、MAE 和 泰尔指数)来看, LSTM 模型在误差修正后表现最优, 其 RMSE 和 MAE 分别下降至 112.6206 和 79.7373, 泰尔指数也降至 0.0006, 表明误差修正显著提高了 LSTM 模型的整体预测精度。随机森林模型的精度次之(RMSE 为 202.7929, MAE 为 185.6092), 相比传统线性回归(RMSE 为 393.0368, MAE 为 301.6126)和 SVR 模型(RMSE 为 346.6338, MAE 为 329.7739), 非线性模型在复杂数据拟合上表现出明显优势。XGBoost 和 LightGBM 的预测精度接近(RMSE 分别为 207.7789 和 215.0602), 但仍次于 LSTM。

在变化趋势捕捉能力方面, 误差修正后的 LSTM 模型相关系数达到 0.8622, 优于误差修正前(0.8455), 说明修正后的预测结果与真实值的变化相似性更高。LightGBM 模型的皮尔森相关系数达到 0.8941, 表现出较强的趋势捕捉能力, 略高于 LSTM; 随机森林和 XGBoost 的相关系数分别为 0.8350 和 0.8599, 表现良好。线性回归相关系数最低, 仅为 0.4890, 表明线性假设难以充分捕捉股指复杂的非线性变化特征。

在方向预测能力方面, LSTM 模型表现优异, 误差修正前和修正后的准确率分别为 97.29%和 97.53%, 说明其在股指涨跌方向的预测中几乎无误。随机森林与 XGBoost 的准确度分别为 94.38%和 94.17%, 其性能表现也较为优异。线性回归和 SVR 则表现较差, 分别仅为 90.80%和 90.05%。

对比加入投资者情绪指标前后的 LSTM 模型来看(表 4、表 5), 无论是误差修正前还是修正后的 LSTM 模型, 引入投资者情绪指数后, LSTM 模型的 RMSE、MAE、泰尔指数及相关系数明显优于未加入该指标时的模型, 虽然在涨跌预测准确率方面有所提高, 但变化较为有限。总体而言, 包含投资者情绪指标的 LSTM 模型在应对股指数据的复杂性及非线性特征时展现出显著的优势, 成为本任务中的最佳模型, 进一步验证了本文预测结果的有效性。

Table 5. The LSTM model with investor sentiment and other models to evaluate the forecasting effect
表 5. 加入投资者情绪 LSTM 模型与其他模型的预测效果评价

模型	RMSE	MAE	泰尔指数	皮尔森相关系数	涨跌预测准确率
LSTM (误差修正前)	123.0246	87.7234	0.0008	0.8455	97.29%
LSTM (误差修正后)	112.6206	79.7373	0.0006	0.8622	97.53%
线性回归	393.0368	301.6126	0.0069	0.4890	90.80%
SVR	346.6338	329.7739	0.0060	0.8299	90.05%
随机森林	202.7929	185.6092	0.0020	0.8350	94.38%
XGBoost	207.7789	192.9428	0.0021	0.8599	94.17%
LightGBM	215.0602	202.1202	0.0022	0.8941	93.89%

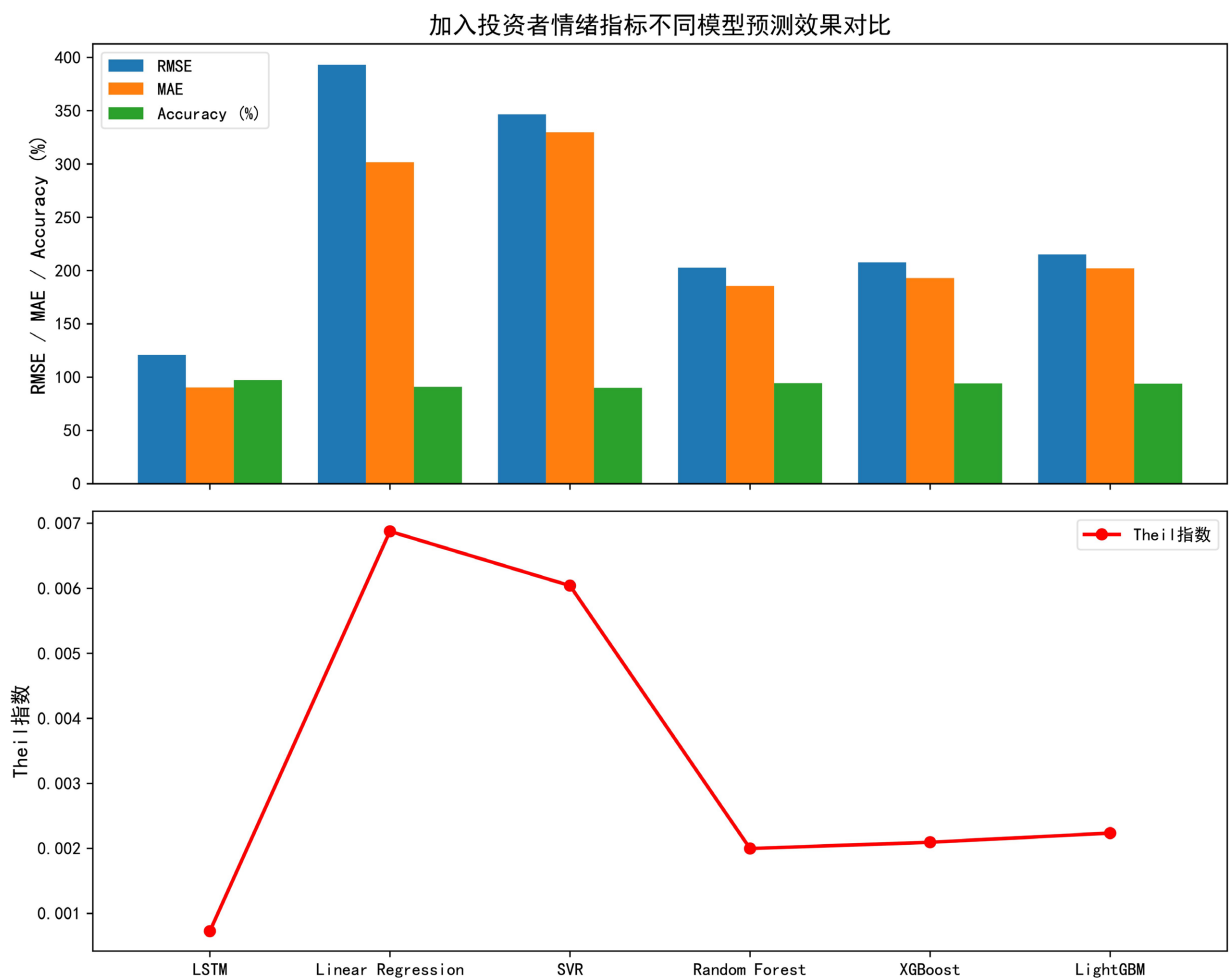


Figure 5. The prediction effect comparison of different models with investor sentiment index
图 5. 加入投资者情绪指标不同模型预测效果对比

4.3.2. 特征重要性分析

图6展示了加入投资者情绪指标后的LSTM模型在预测上证指数收盘价时,各个特征变量的贡献值,单个特征变量的特征重要性越高,说明该变量在预测时的贡献度越高,该特征变量越重要。

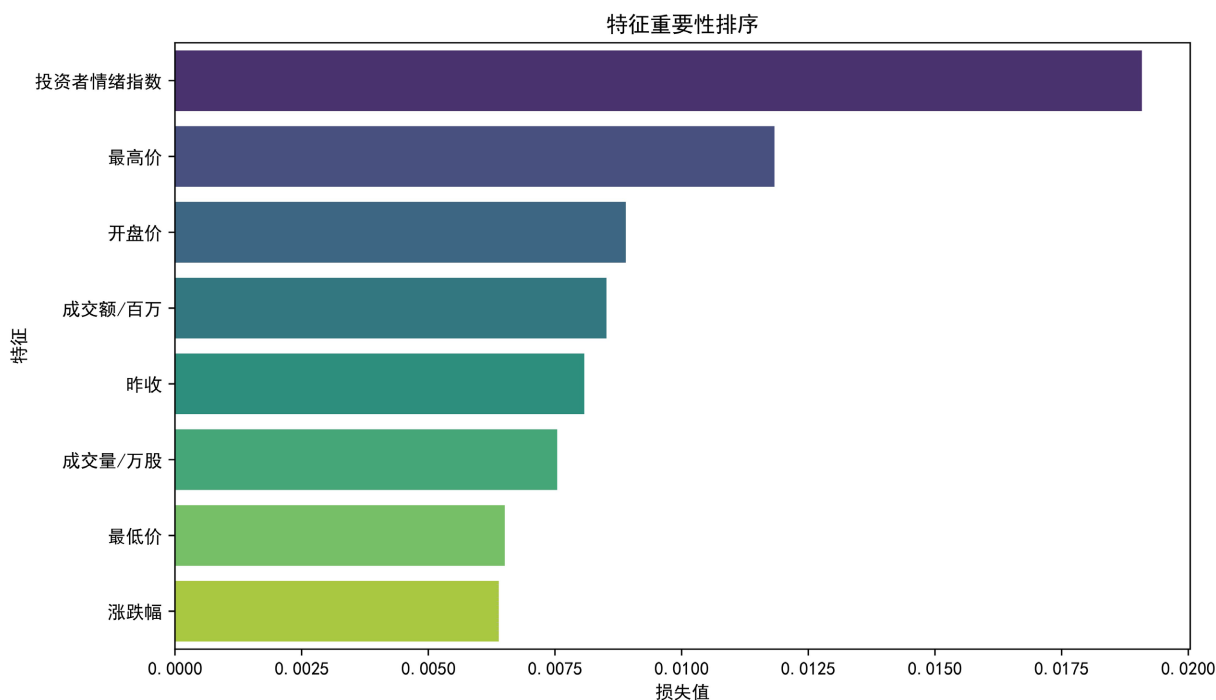


Figure 6. Ranking of feature importance

图 6. 特征重要性排序

本文在计算 LSTM 模型各特征变量的特征重要性时采取的思想为：第一，打乱特征列并计算损失，即每次取一个特征列，随机打乱它的顺序，重新输入到模型中进行预测，然后计算其损失。第二，记录每个特征列的损失，单个特征对应的损失就是该特征的“重要性”，损失值越大说明该特征对模型的重要性越大。

根据图 6 中特征重要性排序的结果，可以看到“投资者情绪指数”在模型预测中的贡献度最高，其损失值为 0.0134，明显高于其他特征。这表明，投资者情绪指数在预测上证指数收盘价时发挥了显著的作用。相较之下，其他特征如“昨收”和“最低价”的损失值分别为 0.0065 和 0.0055，贡献度次之，均远低于“投资者情绪指数”。此外，诸如“成交量/万股”和“成交额/百万”等特征的损失值则相对较小，表明它们对模型预测的影响较为有限。

综上，证明了投资者情绪这一特征变量在预测 A 股上证指数中的重要作用。

4.4. 稳健性检验

本文在上方的分析中，证明了加入投资者情绪作为特征变量的 LSTM 模型在众多机器学习、深度学习模型中具有优质的预测效果，为了检验该结论的稳健性，本文将预测的目标变量替换为同样可以衡量 A 股大盘指数的“A 股指数”，股票代码 000002.SH，对应的开盘价、最高价、最低价、昨收、涨跌幅、成交额/百万和成交量/百万股等基本数据替换为 A 股指数相关数据，投资者情绪指标保持不变。

图 7 呈现了引入投资者情绪指标后的 LSTM 模型对 A 股指数收盘价预测结果，表 6 则列出了该模型的预测效果评估。

可以看出，即便更换了预测目标变量，引入投资者情绪指标的 LSTM 模型依然保持了较好的预测精度，无论是误差修正前还是修正后，对应的 RMSE、MAE、泰尔指数、相关系数和涨跌预测准确率都与原模型预测效果差距不大。这再次证明了投资者情绪这一特征变量在预测 A 股大盘指数方面的重要作用。

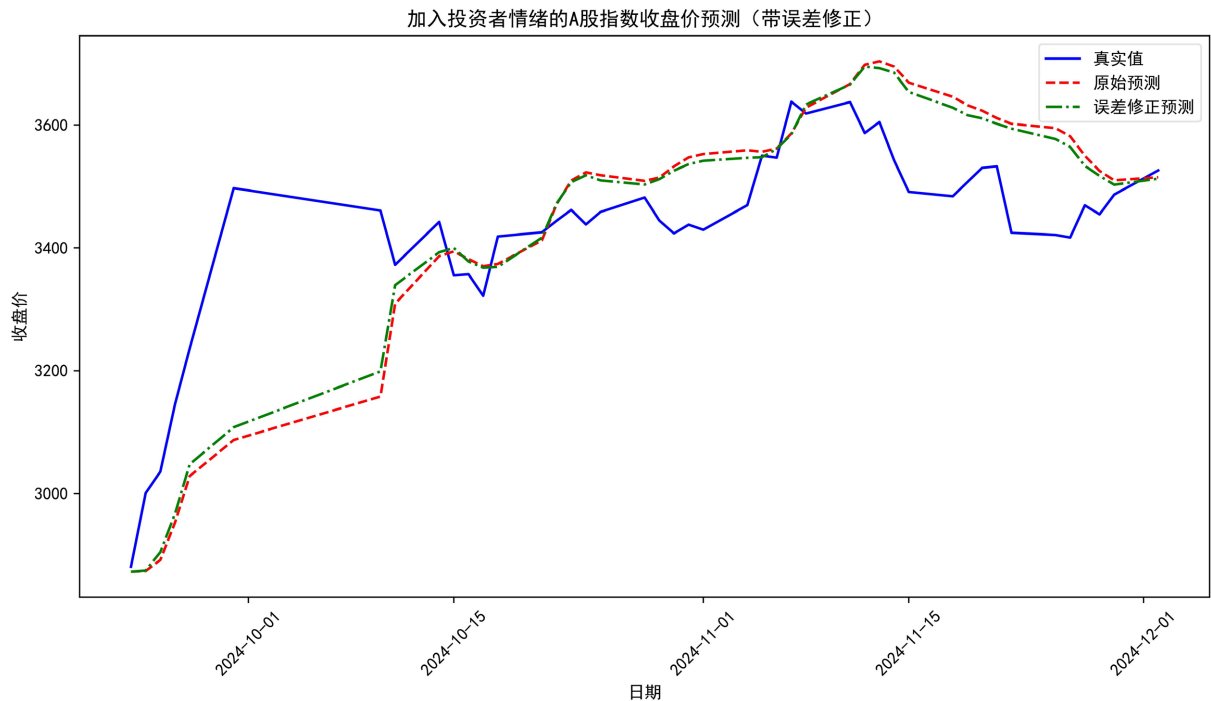


Figure 7. A-share index closing price forecast with investor sentiment (with error correction)
图 7. 加入投资者情绪的 A 股指数收盘价预测(带误差修正)

Table 6. Evaluation of the effect of predicting A-share index closing price
表 6. 预测 A 股指数收盘价效果评价

模型	RMSE	MAE	泰尔指数	皮尔森相关系数	涨跌预测准确率
LSTM (误差修正前)	117.1863	83.1403	0.0006	0.8584	97.55%
LSTM (误差修正后)	107.4088	75.9631	0.0005	0.8743	97.76%

5. 结论与政策建议

本文利用机器学习算法，研究了投资者情绪这一重要特征对 A 股大盘指数的可预测性问题。通过构建加入和不加入投资者情绪特征的 LSTM 模型，对比不同机器学习模型的预测性能，并研究各特征变量的特征重要性，本文研究发现：

第一，基于财经新闻进行情绪识别构建出的投资者情绪指数与 A 股大盘指数存在相似的变化趋势，同时，投资者情绪指数的波动具有一定的前瞻性，这表明股吧、财经新闻等平台上的投资者评论，某种程度上会对 A 股市场的股票表现产生影响。

第二，在与线性回归、SVR、随机森林、XGBoost 和 LightGBM 等预测模型进行比较时，本文提出的 LSTM 模型在预测精度、预测值与真实值的接近度以及整体准确性等方面展现了更强的表现。

第三，将投资者情绪指数纳入本文设计的 LSTM 模型后，模型的预测精度得到了显著提升，这为研究投资者情绪和资本市场的内在关系提供了一定的研究思路 and 方向，也对今后的投资策略构建和市场预警系统构建具有一定的参考价值。此外，本文提出的基于历史的误差修正方法，让 LSTM 模型的预测效果得到了有效改进，提高了其预测性能和预测精度。

基于此，本文提出了可行的政策建议：

第一，构建智能化投资者情绪监测体系。基于机器学习算法和情绪数据，建立投资者情绪监测平台，

通过集成自然语言处理技术与社交网络分析算法,实时抓取股吧、财经新闻、微博等文本数据,精准识别市场情绪波动。当前,金融股票市场的预测通常依赖于日度和月度等低频数据。如果能够借助深度学习技术提取股票市场日内高频数据的特征并构建情绪过热阈值或恐慌指数等分钟级情绪预警指标,将对日内高频投资策略的优化、风险控制和高频交易的执行产生重要影响。

第二,完善投资者教育生态体系。投资者情绪的波动是导致市场出现非理性波动的重要原因,为提高市场的理性运行,政府和金融监管机构应加强投资者教育。针对新入市投资者,开发开户强制学习的“情绪管理模拟器”,利用VR技术还原极端市场场景;针对高频交易者,在交易界面嵌入实时情绪提示弹窗,关联账户操作与情绪指数波动;针对机构投资者,将情绪分析能力纳入从业人员资格考核。通过普及理性投资理念、风险教育等手段,减少情绪驱动的投机行为。同时,可以定期开展投资者心理干预,帮助投资者理性应对市场波动,避免情绪化决策对市场产生负面影响。

第三,推进监管科技基础设施建设。建议设立国家级“金融情绪数据中心”,攻关隐私计算、知识图谱等关键技术。联合大型科技企业研发分布式情绪计算框架,实现券商、基金、互联网平台数据的合规化融合分析;建设监管级情绪知识图谱,标注10万+实体关系,开发情绪传染路径模拟系统。在高校开设监管科技课程,培养“金融+AI”复合人才,定向输送至中证信息、上交所技术公司等机构,形成监管科技能力闭环

参考文献

- [1] Malkiel, B.G. and Fama, E.F. (1970) Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25, 383-417. <https://doi.org/10.1111/j.1540-6261.1970.tb00518.x>
- [2] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, 9, 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [3] Heaton, J.B., Polson, N.G. and Witte, J.H. (2016) Deep Learning for Finance: Deep Portfolios. *Applied Stochastic Models in Business and Industry*, 33, 3-12. <https://doi.org/10.1002/asmb.2209>
- [4] Wei, Y. and Chaudhary, V. (2018) TST: An Effective Approach to Extract Trend Feature in Stock Time Series. 2018 *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Bangalore, 19-22 September 2018, 120-125. <https://doi.org/10.1109/icacsi.2018.8554383>
- [5] Akita, R., Yoshihara, A., Matsubara, T. and Uehara, K. (2016) Deep Learning for Stock Prediction Using Numerical and Textual Information. 2016 *IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Okayama, 26-29 June 2016, 1-6. <https://doi.org/10.1109/icis.2016.7550882>
- [6] Bao, W., Yue, J. and Rao, Y. (2017) A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-Short Term Memory. *PLOS ONE*, 12, e0180944. <https://doi.org/10.1371/journal.pone.0180944>
- [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] 陈卫华, 徐国祥. 基于深度学习和股票论坛数据的股市波动率预测精度研究[J]. 管理世界, 2018, 34(1): 180-181.
- [9] 唐国豪, 姜富伟, 张定胜. 金融市场文本情绪研究进展[J]. 经济学动态, 2016(11): 137-147.
- [10] Bing, L., Chan, K.C.C. and Ou, C. (2014) Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements. 2014 *IEEE 11th International Conference on e-Business Engineering*, Guangzhou, 5-7 November 2014, 232-239. <https://doi.org/10.1109/icebe.2014.47>
- [11] Martin, V. (2013) Predicting the French Stock Market Using Social Media Analysis. 2013 *8th International Workshop on Semantic and Social Media Adaptation and Personalization*, Bayonne, 12-13 December 2013, 3-7. <https://doi.org/10.1109/smap.2013.22>
- [12] Khatri, S.K. and Srivastava, A. (2016) Using Sentimental Analysis in Prediction of Stock Market Investment. 2016 *5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, 7-9 September 2016, 566-569. <https://doi.org/10.1109/icrito.2016.7785019>
- [13] 龙文, 毛元丰, 管利静, 崔凌逍. 财经新闻的话题会影响股票收益率吗?——基于行业板块的研究[J]. 管理评论, 2019, 31(5): 18-27.
- [14] Chinascope. 中国 A 股市场情绪指数[EB/OL]. <https://sentiment.chinascope.com>, 2024-12-05.