投资者情绪波动对股票收益率的分析和预测

王 峣，周上皓，胡亨丰，曾 钐

暨南大学 广东广州

【摘要】本研究通过分析中国股市投资者情绪波动，构建了一个包含宏观经济指标和情感分析的综合情绪指数，并利用 LSTM 和 BiLSTM 等机器学习模型预测股票收益率。结果显示，该模型在市场波动时能快速反映情绪变化，为投资者提供准确预测，增强了市场情绪与股票收益关系的理解和投资决策的科学性。

【关键词】投资者情绪指数；机器学习；股票收益率

【收稿日期】2024 年 11 月 9 日 【出刊日期】2024 年 12 月 28 日 【DOI】10.12208/j.aif.20240004

Analysis and forecast of stock returns based on investor sentiment fluctuations

Yao Wang, Shanghao Zhou, Hengfeng Hu, Shan Zeng

Jinan University, Guangzhou, Guangdong

【**Abstract**】This study analyzes the fluctuations in investor sentiment in the Chinese stock market and constructs a comprehensive investor sentiment index that includes macroeconomic indicators and sentiment analysis. Machine learning models such as LSTM and BiLSTM are utilized to predict stock returns. The results indicate that the model can quickly reflect sentiment changes during market volatility, providing accurate forecasts for investors and enhancing the understanding of the relationship between market sentiment and stock returns, as well as the scientific nature of investment decisions.

【**Keywords**】Investor sentiment index; Machine learning; Stock returns

1 导言

在金融市场的波动中，股市预测对于投资者至关重要，它有助于评估风险并指导科学决策。尽管有效市场假说认为价格已反映所有信息，但中国股市的特殊性为预测提供了可能性。本研究将探讨投资者情绪波动对股票收益率的影响，旨在构建一个包含情绪分析的全面股市预测模型，为投资者提供新的分析工具，以更好地理解市场动态。

2 文献综述

2.1 投资者情绪与股票市场的基础研究

在金融行为学中，投资者情绪对股票市场的影响是一个重要研究领域。研究表明，情绪波动对股票收益率有显著影响。早期研究如叶皓珏和金辉（2020）发现，市场成熟度影响情绪对收益的影响。Baker 和 Wurgler（2006）指出，在某些股票类别中，情绪影响依然显著。张丹和廖士光（2017）通过封闭式基金折价率和认购权证隐含波动率来衡量情绪，揭示了其对市场收益和波动性的影响。雄庆举和吕鹏博（2010）通过主成分分析提取情绪因素，研究了情绪与首日收益的关系。姚远等人（2019）和文凤华（2014）进一步探讨了不同情绪状态对股票收益的影响。王乾（2019）采用 LSTM 网络结合网络上投资者情绪，建立了股票趋势预测模型。

2.2 情绪分析在股票预测中的应用

在情绪分析的应用方面，林亮（2023）构建了高频情绪指标，分析了其对股票收益率的预测能力。李小晗（2009）探讨了月相变化对情绪周期性波动的影响。昌小明（2022）通过恐慌指数和风险厌恶指数，研究了情绪对收益率和波动率的预测能力。

综上所述，投资者情绪与股票市场的关系复杂且多维，情绪波动对市场动态有深远影响。未来的研究应深化对情绪因素的理解，探索其在不同市场条件下的作用机制，并利用网络上投资者情绪等情绪指标为投资者提供决策依据。

3 研究的设计与实施

3.1 研究对象

本研究聚焦于中国股市中投资者情绪波动对股票收益率的影响，特别是通过社交媒体网络所反映的情绪变化。研究构建了一个综合情绪指数，该指数融合了宏观经济指标和利用情感分析技术从网络上投资者评论留言中提取的情绪数据。通过应用机器学习模型，尤其是 LSTM 和 BiLSTM，研究旨在预测股票收益率，并分析市场情绪与股票收益之间的关系。结果显示，综合情绪指数能显著提高股市预测的准确性，尤其是在市场波动时期。这一发现强调了情绪指标在金融市场分析中的重要性，并为投资者提供了一种新的市场情绪量化工具，以支持他们的投资决策。

3.2 研究方式

本研究结合了金融行为学、情感分析和机器学习技术来分析和预测股市情绪与股票收益率之间的关系。研究构建了三个不同的预测模型，每个模型都逐步集成了更多的情绪指标和特征，以提高预测的准确性。以下是模型构建的详细步骤：

（1）数据收集：

本研究通过自动化工具如八爪鱼(Octoparse)从股吧“上证指数吧”获取社交媒体平台搜集投资者情绪相关的文本数据和并从锐思数据库等数据库抓取宏观经济指标和股市交易数据。同时本研究所选取的样本时间段是自 2015 年 12 月至 2023 年 8 月，用于检验预测效果的时间段为 2023 年 9 月至 2024 年 4 月，除去节假日（此期间股票不交易）。

（2）数据预处理：

本研究训练了一个自定义的关于中文文本的词嵌入模型，该模型的工作包含：划分数据集（训练集：测试集=8:2）、中文分词、去除中文常用停用词和标点符号（调用 python jieba 库）以及文本向量化（将按上述工作处理好的文本数据转换为 tokenized Document 后调用 MATLAB 内置 train Word Embedding 函数）。

（3）情绪分析：

在本研究中，情绪分析是利用机器学习技术从网络上投资者留言文本数据中识别和量化情绪倾向的关键步骤。为了实现这一目标，本研究采用了两种先进的机器学习模型：卷积神经网络（CNN）和随机森林（RF），并分别为这些模型准备了合适的训练数据集。

首先，使用 10 万多条微博评论训练的卷积神经网络（CNN）模型，通过结合人工筛选和 word2vec 算法，构建了约 12 万条文本的中文金融情感词典（姜富伟等，2020）。CNN 模型擅长捕捉文本中的局部模式和上下文信息，包括反语评论，其准确率约为 80%。

其次，随机森林（RF）模型利用大连理工大学的中文情感词汇本体库，通过人工筛选和 word2vec 算法扩充，构建了中文金融情感词典（姜富伟等，2020），形成了约 2 万条文本字词的数据集。RF 模型通过构建多个决策树来识别文本中的字词或短语特征，具有自动特征选择能力和高效的并行处理能力，准确率高于 80%。

综上所述，CNN 模型侧重于捕捉文本中的局部模式和上下文信息，而 RF 模型则专注于识别和利用文本中的字词或短语作为分类特征。这两种方法的结合，使得研究能够从多个维度分析和理解投资者情绪，为构建综合情绪指数和预测股票收益率提供了坚实的基础。最终，这些模型的输出被用来构建网络情感因子（NSF），该因子通过量化社交媒体上的投资者情绪，为后续的股票收益率预测模型提供了重要的情绪指标。

（4）情绪指数构建：

①数据收集

本研究收集宏观经济指标，包括但不限于市盈率（PE）、市场换手率（TOR）、消费者信心指数（CCI）以及新增开户者数增长率（NEW）。这些指标被广泛认为是反映市场情绪和投资者信心的有效代理变量。

从网络上投资者留言中提取情绪数据，利用之前训练好的 CNN 和 RF 模型对网络文本进行情感分析，得到网络情感因子（NSF）。这一步骤涉及大量的文本数据预处理工作，包括分词、去除停用词和标点符号，以及将文本转换为向量形式。

②主成分分析（PCA）

本研究借鉴 Baker 和 Wurgler（2006）的方法，采用主成分分析（PCA）技术，将宏观经济指标与基于网络文本数据的投资者情绪因子相结合，以构建更准确可靠的情绪指数。PCA 通过识别数据模式，

将相关变量转化为一组线性不相关的主成分，以捕捉投资者情绪的动态变化。研究中选取了 4 个宏观经济指标及其滞后一期值，连同两组网络情感因子，共计 10 个代理变量进行 PCA 分析，以构建综合情绪指数，代表数据集的主要变异性。

通过选择前几个主成分，构建了综合情绪指数，这些指数能够代表原始数据集中的主要变异性。

③情绪指数的构建

通过 PCA，本研究提取了前五个主成分（sent_RF 前 5 个主成分方差累计贡献率达 84.393%，sent_CNN 前 5 个主成分方差累计贡献率达 85.292%），并利用这些主成分对投资者情绪指标进行加权平均，从而构建了两个时间序列指标：sent_RF 和 sent_CNN。这两个指标分别代表了基于随机森林和卷积神经网络分析得到的特征数据集。

通过计算得到的主成分得分，结合宏观经济指标和网络情感因子，构建了综合情绪指数。这一指数不仅包含了传统的市场情绪代理变量，还融入了社交媒体舆论分析的结果，从而能够更全面地捕捉市场情绪的细微变化。

（5）模型建立与训练：

①数据集划分

在机器学习实践中，为了准确评估模型的泛化能力，本研究按照时间将数据划分为 2 个阶段:第一阶段（2015 年 12 月-2023 年 8 月）和第二阶段（2023 年 9 月-2024 年 3 月）。本研究将第一阶段数据作为训练集用于模型构建，训练集是用于训练模型的数据，通过它模型能学习到数据的特征和规律。得到用第一阶段数据训练好的模型后，本研究将第二阶段数据作为测试集，用训练好的模型预测收益率，以对模型的预测效果进行评价。

②预测模型

预测模型一

模型一的数据选取基于 K 线实体占比和股票收益率之间存在高度显著的相关性（徐钟荣，张军成，2019），选取 4 个股票历史交易数据（开盘价，收盘价，最高价，最低价）构造 K 线实体占比，对上证指数收益率进行预测。

$$\text{K线实体占比} = \frac{\text{收盘价} - \text{开盘价}}{\text{最高价} - \text{最低价}} \quad \text{公式（1）}$$

预测模型二

模型二的指标选取借鉴了姚远、姚贝贝和钟琪（2019）和文凤华等人（2014）关于能够反映股民情绪的传统经济指标选取的研究，选取了 4 个传统的能够反映股民情绪的经济指标（市盈率，市场换手率，消费者信心指数，新增开户者增长率数据）。对其进行主成分分析以后构建传统投资者情感指数和 K 线实体占比对上证指数收益率进行预测。

预测模型三

在获取网络上投资者留言文本数据后，基于随机森林（RF）和卷积神经网络（CNN）分别进行网络情感因子（NSF）的初步构建。同时本研究参照姚远、姚贝贝和钟琪（2019）和文凤华等人（2014）研究中关于能够反映股民情绪的经济指标的选取，选取了 4 个传统的能够反映股民情绪的经济指标（市盈率，市场换手率，消费者信心指数，新增开户者增长率数据）和网络情感因子（NSF_RF 和 NSF_CNN）分别进行主成分分析，再共同构建综合情绪指数，最后将 2 个综合情绪指数结合 K 线实体占比数据，也就是构建二维综合投资者情绪指数对上证指数收益率进行预测。

③模型训练方法选取

在本研究中，模型建立的方法选取聚焦于利用长短期记忆网络（LSTM）和双向长短期记忆网络（BiLSTM）来预测股市收益率，考虑到这些模型在处理时间序列数据中的长距离依赖问题时的优势。同时，辅以反向传播（BP）算法优化网络参数，并考虑了卷积神经网络（CNN）在捕捉局部时间模式上的潜力，以期构建一个综合考虑多维度信息、能够有效预测市场情绪变化及其对股票收益率影响的机器学习框架。

1）模型评估：

使用均方误差(MSE)和相关性分析(R^2)等统计指标来评估模型的预测准确性和泛化能力。

2）结果分析与结论：

对比不同情绪指标对模型预测能力的影响，分析综合情绪指数在短期预测方面的优势，并提出未来研究的方向。

整个研究方法体现了跨学科的研究视角，将金融学、行为学、数据科学和机器学习等领域的技术相结合，以期为股市预测提供更为科学和精确的方法论。通过逐步集成更多的情绪指标和特征，本研究的

模型能够更全面地捕捉市场情绪的细微变化，从而为投资者提供更为敏感和准确的市场反馈。

4 模型数据分析与结论

4.1 预测模型性能对比

（1）模型训练效果表现

该部分基于上述四种模型分别对我们建立的时间（2015 年 12 月-2023 年 8 月）序列指标和相应的收益率进行回归分析，并多次训练和改进模型，调整相对较好的模型参数，用于下一部分（2023 年 9 月-2024 年 3 月）的预测。

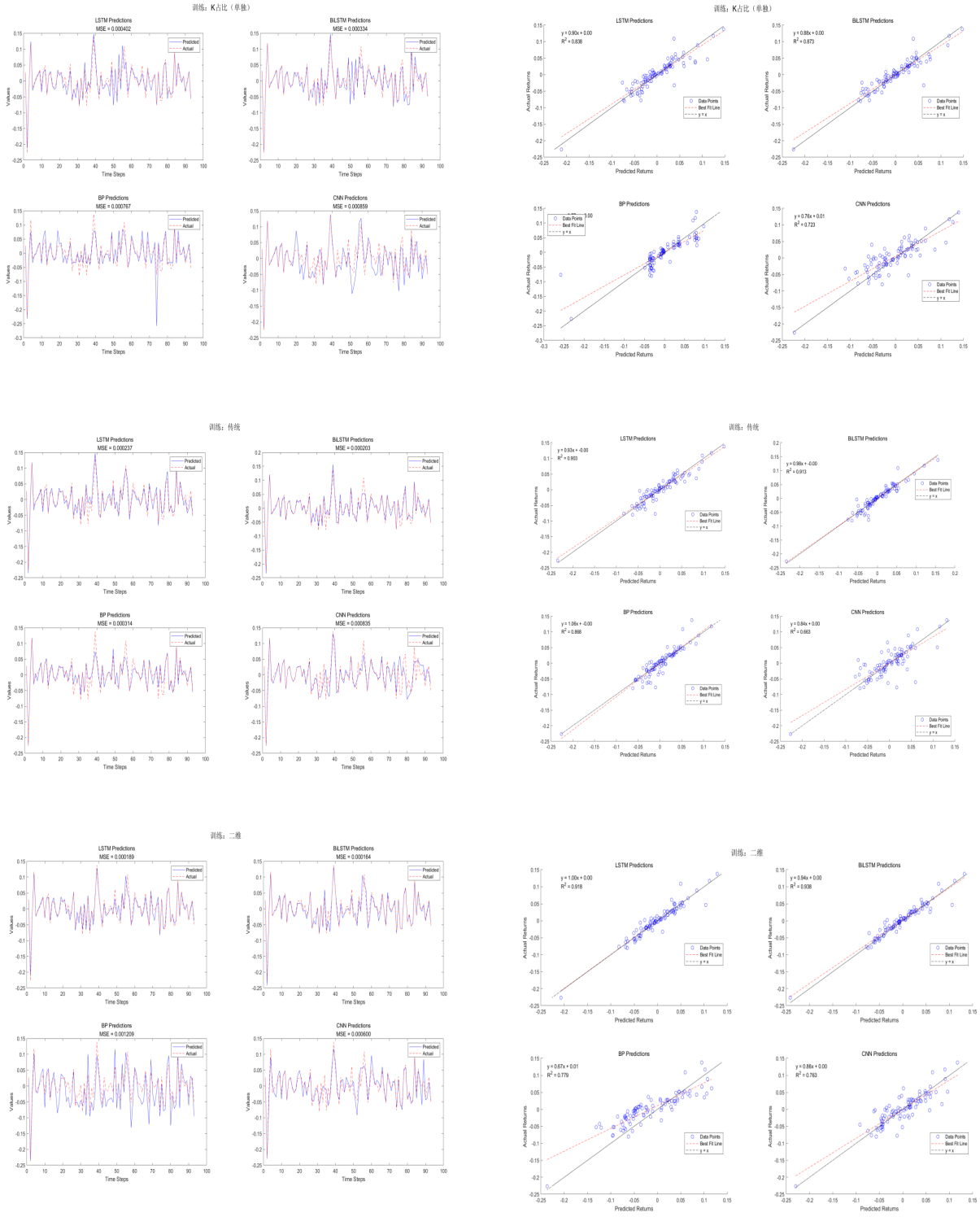

图 1　训练集上三种预测模型综合回归对比图和散点图

（2）模型对比分析结论

①LSTM 和 BiLSTM 在金融市场预测中的应用

在本项研究中，长短期记忆网络（Long Short-Term Memory, LSTM）和双向长短期记忆网络（Bidirectional Long Short-Term Memory, BiLSTM）因其在处理序列数据时能够同时考虑时间上的前后信息，因此在预测任务中表现出了较为卓越的性能。这一现象可能归因于金融市场数据所固有的强烈时间序列特性，LSTM 和 BiLSTM 模型通过其独特的记忆机制和信息传递方式，能够更加有效地捕捉并利用这些特性，从而在预测金融市场的未来走势时具有更高的准确性。

②相关系数在模型预测性能评估中的作用

本研究使用相关系数衡量模型预测准确性，发现 LSTM 和 BiLSTM 模型在预测相关性上表现最佳。因此，研究将深入分析这两种模型的预测结果，同时利用其他模型的结果来评估泛化能力，确保研究结果的全面和可靠。

③情绪指标对模型预测能力的影响

在对比预测模型一与预测模型二、预测模型三的性能时，本研究得出了重要的结论：无论是传统的情绪指标还是本研究提出的新情绪指标，均能对原有模型的预测能力产生显著的增强效果。这表明情绪指标作为金融市场分析的重要组成部分，无论是传统的还是创新的，都能为模型提供有价值的信息，从而提高预测的准确性。

④新旧情绪指标的比较与权衡

通过对比预测模型二与与预测模型三的性能，本研究可以得出结论：新情绪指标在提升模型性能方面比传统指标更有效，但泛化能力较弱。因此，在实际应用中，研究者需要平衡新指标的增强效果和其泛化能力的局限，以在金融市场分析和预测中达到最佳效果。

综上所述，本研究通过对比不同模型的性能，深入探讨了情绪指标在金融市场预测中的应用，并提出了新情绪指标对原有模型具有显著增强效果的观点。同时，本研究也指出了新情绪指标在泛化能力上的局限性，为未来的研究和实践提供了宝贵的经验和启示。

4.2 基于情绪指数的预测模型结果对比

（1）模型预测效果表现

为了更好评估模型，现在特地选取 2023 年 9 月至 2024 年 3 月相关指标对应的月数据作为验证集，通过上述分析过程构建综合情绪指标，用训练好的对应模型分别直接对收益率预测，得到结果与真实值的比较如下所示：

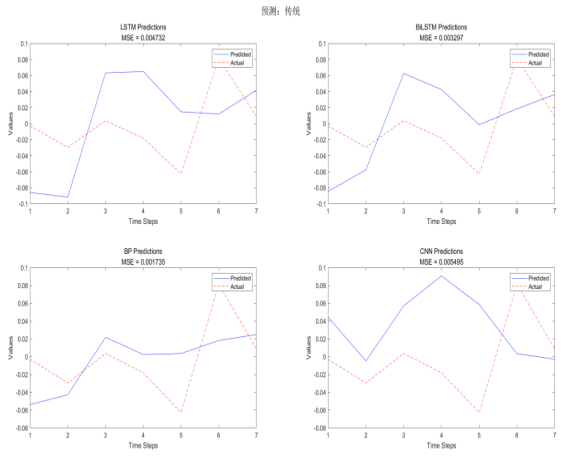

图 2　传统投资者情绪指标预测收益率效果图

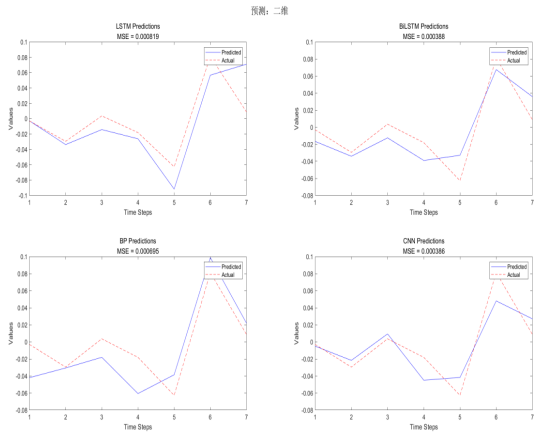

图 3　综合投资者情绪指标预测收益率图

（2）预测结果对比分析结论

①综合投资者情绪指标短期预测方面具有明显优势

引入社交媒体舆论分析的这一创新之处，使得综合投资者情绪指标在短期预测方面具有明显优势。这种综合投资者情绪指标的核心在于其对社交媒体舆论的分析。通过监测和评估投资者在社交平台上的讨论、评论和情绪表达，该指标能够捕捉到市场情绪的细微变化，从而为投资者提供更为即时和动态的市场信息。此外，它还结合了传统经济指标，以确

保模型在分析时能够综合考虑市场的宏观和微观因素。它能够快速响应市场新闻、政策变动、公司事件等因素的影响，为投资者提供更为敏感和准确的市场情绪反馈。这种快速响应能力，尤其在市场波动剧烈或出现重大事件时，显得尤为重要。

②综合投资者情绪指标的实时性和动态性

综合投资者情绪指标的实时性和动态性，使其能够不断适应市场的变化，及时更新和调整预测模型。这种灵活性和适应性，为投资者在快速变化的市场中做出及时反应提供了有力支持。

综上所述，本研究创建的综合投资者情绪指标，通过结合社交媒体舆论分析和传统经济指标，为市场预测提供了一种新的视角和工具。它不仅提高了短期预测的准确性，而且增强了模型对市场即时情绪变化的敏感度，有助于投资者更好地理解市场动态，制定更为科学的投资决策。随着社交媒体在金融领域影响力的不断增强，这种综合指标的应用潜力和价值将愈发显著。

参考文献

[1] 王晴. 网络情感因子对股票收益率的影响研究[J]. 金融，2014，13(4), 835-840.

[2] 王改银. 中国股市收益率可预测性的实证检验[J]. 金融，2023，13(6), 1214-1225.

[3] 叶皓珏, & 金辉. 投资者情绪对股票收益影响的实证分析[J]. 商业全球化，2020，8(3), 63-73.

[4] 徐钟荣, & 张军成. K 线实体占比对股票短期收益率影响的实证分析. 金融理论与教学，2019,(2), 41-44.

[5] 文凤华, 肖金利, 黄创霞, 陈晓红, 杨晓光. 投资者情绪特征对股票价格行为的影响研究[J]. 管理科学学报，2014, 17(3): 60-70.

[6] 姚远, 姚贝贝, 钟琪. 投资者情绪对股票收益率的影响研究——基于上证 A 股数据的实证分析[J]. 经济研究导刊，2019(5): 88-91.

[7] 雒庆举, 吕鹏博. 基于投资者情绪的 IPO 首日收益研究[J]. 经济与管理研究，2010.

[8] 张丹, 廖士光. 中国证券市场投资者情绪研究[J]. 经济研究，2004.

[9] 王乾. 基于 LSTM 的多特征股票趋势预测研究[D]. 武汉理工大学，2019.

[10] 林亮. 高频投资者情绪指数的构建及其预测能力分析[D].南开大学,2023.

[11] 李小晗. 情绪周期与股票收益 — 基于中国股票市场月相效应的检验[J]. 中国会计评论，2009，7(4), 384-418.

[12] 姜富伟，孟令超，唐国豪. 媒体文本情绪与股票回报预测.经济学(季刊),2021(4)，1323-1344.

[13] 昌晓明. 投资者情绪对股票收益率及波动率预测的研究——基于 VIX 和时变风险厌恶[D]. 长沙理工大学，2022.

[14] Antweiler, W., & Frank, M. Z. Is all that talk just noise? The information content of internet stock message boards[J]. The Journal of Finance, 2004, 59(3), 1259-1294.

[15] Ndlovu, B., Faisa, F., Resatoglu, N. G., & Tursoy, T. The impact of macroeconomic variables on stock returns: A case of the Johannesburg Stock Exchange[J].Revista De Statistica, 2018,(2),87-104.

[16] Baker, M., & Wurgler, J. Investor Sentiment and the Cross-Section of Stock Returns[J]. The Journal of Finance, 2006，61(4), 1645-1680.

[17] Fuwei Jiang, Joshua Lee, Xiumin Martin, and Guofu Zhou.Manager Sentiment and Stock Returns[J].Journal of Financial Economics,2009, 132(1), 126-149.