# 基于股票评论的投资者情绪对持有股票收益影响的研究

孟 雪, 李艳茹, 胡 锋 *

(曲阜师范大学 统计与数据科学学院, 山东 曲阜 273165)

**摘 要**: 投资者行为易受互联网舆论的影响, 进而造成股票收益的波动. 分析投资者情绪对股票收益的影响方式有利于投资者规避投资风险, 促进我国股票市场稳定发展. 基于东方财富股吧 2020 年 7 月至 2021 年 2 月上证股票的评论数据, 利用加权朴素贝叶斯分类模型构建了投资者情绪因子, 并对情绪因子的构建方式进行了改进. 随后将情绪因子引入中国版 Fama-French 三因子模型, 针对单只股票和持股期为 1 个月的投资组合, 基于线性回归、长短期记忆神经网络模型, 从线性、非线性两个角度研究了投资者情绪对其持有股票收益率的影响. 结果表明, 投资者情绪对股票收益率具有非线性的正向影响. 前一日的投资者情绪会对当日股票收益产生影响, 投资者在研究期望收益率时需予以考虑.

**关键词**: 投资者情绪; 股票收益; 加权朴素贝叶斯分类模型; 中国版 Fama-French 三因子模型; 长短期记忆神经网络模型

## 1 引言

中国互联网络信息中心发布的《中国互联网络发展状况统计报告》显示, 截至 2020 年 12 月, 我国网民规模达 9.89 亿, 互联网普及率达 70.4%. 互联网已经成为舆论发布、传播的主要载体, 每个人既创造和传递舆论, 又深受其影响. 这种现象同样出现在金融投资领域, 投资者会在互联网中了解上市公司的财务状况和新闻动态, 交流投资见解, 吸取他人观点, 并在交流过程中形成对股票市场的预期, 进而做出决策. 事实表明, 投资者情绪会对股票市场产生一定影响. 如 2018 年国际贸易关系紧张, 投资者普遍对 A 股市场信心较低, A 股走势持续下跌. 2020 年受新冠肺炎疫情影响, 全球股市全线杀跌, 离岸人民币单日暴跌超 500 点, 亚太股市全面失守. 但由于我国疫情防控及时, A 股市场随即成为全球股市的避风港. 随着我国经济的发展以及全球局势的改变, 国内金融网络舆情日益活跃、热点频发, A 股市场波动频繁, 投资风险提高. 健全金融监管体系, 守住不发生系统性金融风险的底线, 是习近平新时代中国特色社会主义思想在金融领域的根本要求, 如何利用网络舆情更精确的预测股票收益、降低投资风险、完善现代金融监管体系、健全风险全覆盖监管框架成为投资者、监管者和研究者共同关注的问题. 在此背景下, 本文基于东方财富股吧 2020 年 7 月至 2021 年 2 月投资者发表的股票评论构建了较准确的情绪因子, 并将传统模型、深度学习方法与金融理论相结

**收稿日期**: 2021-11-01
**资助项目**: 国家自然科学基金 (11801307); 山东省优秀青年资助项目 (ZR2016JL002); 山东省研究生教育教学改革研究项目 (SDYJG19178)
* 通信作者

合, 从线性、非线性两个角度研究了投资者情绪对 A 股收益率的影响方式. 为了使研究更具现实意义, 本文分别研究了投资者投资单只股票和投资组合两种投资策略的股票收益率. 研究结果提高了股票收益的预测精度, 为投资者科学理性的做出投资决策提供了理论支持.

## 2 文献综述

对于投资者情绪, De Long 等 [1] 提出了噪声交易者模型, 认为投资者情绪是股票市场中的随机噪音. 随着行为金融学的发展, 研究者发现投资者情绪是可以度量的, 并给出了多种度量方法, 主要有间接度量方法和直接度量方法. 间接度量方法即利用已有的金融指标代替投资者情绪, 或将多个金融指标综合起来代替投资者情绪. 最具代表性的是基于主成分分析法构建的 BW 指数 [2], 该指数在后续的研究中被极广泛的应用 [3−4]. 然而, 部分金融指标的数据不易获取, 成本较高. 直接度量方法可以改善这一点, 即根据互联网中投资者的态度直接获取投资者情绪. 我国研究者最初采用了调查问卷的方式直接构建情绪因子 [5], 但这种方法的时效性较低. 张信东和原东良 [6] 基于微博 “牛市” 和 “熊市” 等词语每天出现的次数构建了情绪因子, 提高了时效性, 但该方法依赖于研究者个人对关键词的选取, 构建的情绪因子不够客观. 随着机器学习技术的发展, 研究者采用了文本挖掘算法筛选关键词并进行文本分类, 从而准确客观地判别了股评的情绪基调. 金秀等 [7] 采用 Bayes 分类算法对新浪股吧评论进行分析, 从质化信息的情绪基调、量化信息的张贴程度和强度信息的关注水平三个维度构建了情绪因子. 北京大学国家发展研究院构建的中国投资者情绪指数利用了支持向量机 (SVM) 模型从上市公司的相关文本中提取了投资者情绪. 根据欧阳资生等 [8] 的总结, Bayes 分类模型与 SVM 模型相比能在保证分类效果的前提下简化计算.

现有研究大多采用传统文本挖掘算法, 且研究对象主要是微博评论或公司新闻文件. 然而, 微博用户复杂多样, 构建的情绪因子噪音较多; 公司文本、新闻数据大多反映了机构投资者和上市公司的投资看法, 此类投资者的非理性程度相对较弱. 石勇等 [9] 研究了不同来源的社交媒体所代表的不同投资者群体态度对股市影响的差异性, 发现代表中小投资者态度的股吧情绪对股市的影响大于代表机构投资者态度的新闻情感数据. 因此, 为更好的反映中小投资者的看法, 本文以东方财富股吧为例, 基于股吧投资者的评论信息, 采用改进的贝叶斯文本分类模型构建情绪因子.

对于股票收益的影响因素, 金融学给出了较为成熟的理论, 其中, Fama 和 French[10] 提出的 Fama-French 三因子模型是重要的收益率解释模型. 该模型基于对股票收益影响程度较大的金融指标构建了三个影响因子, 进而拟合了三因子影响股票收益率的线性模型. 该模型认为投资者是理性的, 未考虑市场中的非理性因素和不确定因素.

在行为金融学理论中, 股票市场中隐藏着大量的非理性因素和不确定因素, 例如网络舆情、媒体报道、国家经济政策、国际环境等, 这些因素均在不同程度上影响着投资者的股票收益情况. 研究表明, 中小投资者在股票市场中面对的信息十分冗杂, 每位投资者接触信息的广度和深度也有差异, 但这些信息最终都会体现在投资者形成的个人投资预期中. 也就是说, 非理性因素和不确定因素会在不同程度上影响投资者情绪, 使得投资者调整投资决策, 进而影响股票收益. 在网络舆情方面, 樊鹏英等 [11] 发现投资者会通过综合分析各类信息形成对金融市场的预期, 并在各网络平台中进行交流, 进而影响他人的投资行为, 因此各平台中以

主观形式存在的舆情信息是影响投资者决策的重要因素. 在媒体报道方面, 饶育蕾和王攀 [12] 研究发现媒体关注度会影响投资者情绪, 进而对股票的短期收益产生正向影响. 在政策方面, 张前程和杨德才 [13] 以货币供应量作为货币政策代理指标, 发现货币政策能够对投资者情绪产生影响. 张博等 [14] 的研究表明, 投资者在接收到货币政策信号后会基于一系列心理过程对信息进行加工处理, 进而改变投资行为, 这一过程表现为投资者情绪, 即货币政策会影响投资者情绪. 此外, 重大突发事件也会造成股票收益的波动, 山立威 [15] 发现突发事件引起的焦虑、恐慌等负面情绪才是影响股票收益率的主要因素.

基于上述原因, 许多研究者针对投资者情绪对股票收益的影响进行了研究. 许天阳 [16]、张艾莲等 [17]、李岩等 [18] 将投资者情绪引入该模型, 得到 Fama-French 四因子模型, 模型的拟合效果得到改善. 然而, 对于 A 股市场, 照搬基于欧美市场构建的 Fama-French 三因子模型会使得结果不理想. Liu 等 [19] 将 Fama-French 三因子模型改进为适用于 A 股市场的中国版三因子模型, 更好的解释了学术界在中国市场上发现的大多数收益率截面异象, 解释力度更强. 此外, Liu 等 [19] 等将换手率作为情绪因子引入中国版三因子模型, 有效弥补了三因子的不足.

上述研究构建了投资者情绪对股票收益的线性预测模型, 随着研究的不断深入, 研究者们发现情绪因子与股票收益之间存在非线性特征. 黄润鹏等 [20] 基于微博发帖构建了情绪因子, 并利用 SVM 模型预测了股票收盘价的上涨或下跌情况. 张信东和原东良 [6] 基于向量自回归、神经网络等方法, 发现情绪因子可以用于预测上证指数的走势. 然而, 这些研究采用的非线性模型较为传统, 所得的结果不够全面. 长短期记忆神经网络模型 (LSTM) 可以有选择的学习长期的、有价值的信息, 杨青和王晨蔚 [21]、黄婷婷等 [22]、耿晶晶等 [23] 利用该模型对股票的金融时间序列数据进行预测, 发现该模型在预测精度和稳定性两方面具有优势, 在金融市场的预测研究中有广阔的应用前景. 目前的研究虽然利用机器学习提高了股票收益的预测精度, 但未能将机器学习技术与金融理论相结合, 导致变量冗余或缺失, 模型效率不高, 结论不够全面. 针对这一点, 潘水阳等 [24] 利用神经网络模型对 Fama-French 三因子进行拟合, 发现神经网络模型的样本外拟合优度优于三因子模型, 脱离了依靠增加解释变量个数提高模型解释力的研究框架, 但未考虑情绪因子.

综上, 现有研究未能实现投资者情绪、金融理论及机器学习技术的结合, 存在考虑不全面、变量冗余或缺失、模型较传统、模型效率低等问题. 本文针对这些问题进行改进, 内容安排如下: 首先提出研究假设, 介绍理论模型, 其次构建股票收益的影响因子, 然后进行投资者情绪影响股票收益的实证分析, 最后进行稳健性检验并得出结论.

## 3 研究假设

### 3.1 投资者情绪对股票收益率影响的合理性

投资者情绪一直是行为金融学研究的热点问题, 常用于解释各种市场异象. 影响投资者情绪的因素主要有以下两方面: 一是投资者的心理因素, 投资者在信息处理和行为决策中是有限理性的, 从而引发的过度自信、锚定效应、损失厌恶等心理偏差. 另一方面, 投资者情绪还受到社会环境、国家政策、国际关系、群体行为等外界因素的影响. 情绪群体的羊群效应是影响市场稳定性的重要因素, 造成羊群效应的原因主要是信息的不确定性. 对于个体投资

者, 由于资金、技术的限制, 个体投资者获取信息的成本更高, 获取收益的能力更弱, 为了趋利避险, 个体投资者通常依赖他人收集、分析信息, 盲目信任所谓的 “内幕消息”, 进而影响投资收益. 此外, 媒体舆论、社会压力或突发事件也会使投资者出现规避风险的从众心理, 致使大众行为趋于统一. 机构投资者则会因为名誉、报酬等原因产生互相模仿的行为, 或因采用类似的经济模型、市场信息和信息处理技术而产生高度同质性 [25].

如今, 我国的互联网普及率高, 但网络舆情不稳定, 信息繁杂, 这种活跃、开放的信息交流进一步加快了投资者情绪的形成, 扩大了投资者情绪的影响规模, 投资者会根据当下的情绪氛围调整投资决策, 进而影响收益情况. 由此, 本文提出假设 H1.

H1: 投资者在互联网中表达的情绪信息可以用于度量投资者情绪, 并对投资者的股票收益产生一定影响.

### 3.2 情绪信息认可度对投资者情绪的影响

股吧作为股票投资者交流的专业性平台, 在股票市场中扮演着重要角色, 现有研究通常利用股吧投资者的发帖内容度量投资者情绪, 但这种度量方法较为片面, 未能考虑投资者参与股吧讨论时情绪的相互影响. 根据 Noelle-Neuman [26] 的 “沉默的螺旋” 理论, 人们会避免由于单独持有某些观点而被孤立. 因此, 当人们看到自己赞同的观点受到广泛支持时会更积极地表达自身观点, 反之则会尽量保持沉默, 最终舆论中占据优势地位的观点会愈演愈烈, 处于劣势的观点会愈发沉默. 这一理论同样适用于股吧, 投资者倾向于浏览与自己观点相似的帖子, 随着热门帖的浏览量增多, 更多的投资者会提高对其观点的信任, 而浏览量低的帖子会因无人点击而逐渐被人忽略. 因此, 股吧中情绪信息的认可度会促进情绪氛围的形成, 进而影响投资者情绪. 本文提出假设 H2.

H2: 考虑情绪信息的认可度, 即浏览量可以提高情绪因子度量的准确性, 且对股票收益的预测能力更强.

### 3.3 投资者情绪对股票收益的非线性影响

在研究股票收益时, 传统金融学模型大多基于市场有效性假说在线性模式下进行构建, 然而, 我国股票市场中存在着大量的非理性因素, 投资者行为也不是相互独立的, 市场有效性假说难以成立. 实证研究表明, 传统金融模型存在的局限性, 曹红辉等 [27]、牛晓健和吴客形[28] 验证了股票市场存在非线性特征, 投资者情绪不稳定会导致股票价格的剧烈变化. 根据文献综述中的介绍, 目前的研究重点主要在于如何将深度学习方法与金融理论结合起来, 寻找深度学习中的经济学原理, 从而构建出更优良的金融计量模型, 提炼出适合金融市场的框架或规律性方法 [29]. 由此, 本文提出假设 H3.

H3: 基于传统金融模型和深度学习方法可构建出预测股票收益率的非线性模型, 投资者情绪对股票收益率具有非线性影响, 非线性模式下的股票收益率预测模型精度更高.

### 3.4 前一时刻的投资者情绪对股票收益的影响

长期实验结果表明, 投资者情绪对股票收益的影响是持续不断的, 某一时刻的投资者情绪会对后期的股票收益产生影响. 从投资者角度, 个体投资者获取信息存在延迟, 接受信息并形成稳定的投资情绪也需要时间, 且个体投资者的有限理性行为会在短期内保持高涨. 从社会环境角度, 某一热点话题造成的社会环境和舆论氛围具有持续性, 在新的热点出现之前, 投资者的羊群效应会持续影响投资者情绪, 进而影响投资决策. 然而, 我国金融网络舆情具有出

现快速、消退迅速且灵活多变的特点 [8], 故舆论氛围的持续性是短期的. 基于此, 本文提出假设 H4.

H4: 前一期的投资者情绪会对当期的股票收益产生显著影响.

### 3.5 投资者情绪对股票投资组合收益的影响

在现有文献中, 除了单只股票的收益率, 研究者也针对股票投资组合的收益率进行了分析. 赵胜民等 [30] 基于 Fama-French 五因子模型对投资组合超额收益率进行回归分析, 研究了股票前景理论价值与预期收益率之间的逻辑关系. 刘澜飚和郭亮 [31] 参考 Fama 和 French 的做法研究了广义失望厌恶导致的下行风险与投资组合预期收益率的关系, 当改变投资组合的构建方式时, 其关系具有稳健性. 在金融学理论中, 投资组合收益率的构建方法多种多样, 本质上是在一定持股期内, 确定各股票收益率的投资权重, 如市值加权投资组合即以各股票的市值比作为投资权重, 最优投资组合的投资权重则是在保证收益的前提下使得投资风险最小得到的. 在现实生活中, 投资者会基于不同目标选择不同的投资组合作为投资决策, 即选择不同的投资权重. 投资者情绪作为影响投资决策的重要因素, 也会在一定程度上影响投资组合收益率. 因此, 本文假定投资组合中各股票的持股期为 1 个月、投资者以投资风险最小为目标购买股票池中的股票, 提出假设 H5.

H5: 在一定持股期内, 投资者情绪会对其持有投资组合的收益率产生显著影响.

## 4 模型介绍

### 4.1 TF-IDF 算法

TF-IDF 算法是常用的特征词加权算法, 可以用于评估词的重要程度 [32]. TF 表示词频 (Term Frequency), 即特征词在各文本中出现的频率, 与特征词的重要程度成正比. IDF 是逆文本频率 (Inverse Document Frequency), 指特征词在文本集中的类别区分能力:

$$IDF_j = \log\left(\frac{N}{N(t_j)+1}\right), \quad j = 1, 2, \cdots n, \tag{1}$$

其中 $N$ 为总文本数, $N(t_j)$ 为文本集中包含特征词 $t_j$ 的文本数, IDF 值与特征词的重要程度成反比. 此时特征词的权重表达式为:

$$\omega_{ij} = \frac{n_{ij}}{\sum\limits_{j=1}^{n} n_{ij}} \times IDF_{ij}, \quad i = 1, 2, \cdots, m,\ j = 1, 2, \cdots n. \tag{2}$$

长期实验表明, 在分类问题中, 特征词在各个类别中并非是均匀分布的, 有些特征词会在某类别的大多数文本中出现, 此时 TF-IDF 算法会使得此类特征词的 IDF 值很小, 虽然其代表性较强, 但依旧会被掩盖. 针对这一缺陷, 隗中杰 [33] 提出用类 $c_k$ 中特征词 $t_j$ 的聚集程度来表达词的 “独特性”, 即对 IDF 的计算方法进行改进:

$$p_{jk} = \frac{N(t_j, c_k)}{N(t_j, c_k) + N(\bar{t}_j, c_k)}, \quad p_j = \frac{N(t_j, c_k) + N(t_j, \bar{c}_k)}{N} = \frac{N(t_j)}{N},$$

$$IDF'_{jk} = \log\left(\frac{p_{jk}}{p_j} \times N + 0.01\right), \quad j = 1, 2, \cdots n, \quad k = 1, 2, \cdots, K, \tag{3}$$

其中, $N(t_j, c_k)$ 表示类 $c_k$ 中包含特征词 $t_j$ 的文本数, $N(\bar{t}_j, c_k)$ 表示类 $c_k$ 中不包含特征词 $t_j$ 的文本数, $N(t_j, \bar{c}_k)$ 表示其他类中包含特征词 $t_j$ 的文本数. $p_{jk}$ 与 $p_j$ 的比值即为该特征词的

聚集程度, 聚集程度越大, 说明特征词在该类中更具有代表性. 如此, 就将类别信息加入到了 TF-IDF 算法中, 改进的权重表达式为:

$$\omega'_{ijk} = \frac{n_{ij}}{\sum_{j=1}^{n} n_{ij}} \times IDF'_{jk}. \tag{4}$$

进一步, 可将上述权重进行归一化处理, 防止 TF-IDF 算法出现对长文本的偏向.

### 4.2 基于 TF-IDF 加权的朴素贝叶斯文本分类模型

朴素贝叶斯文本分类模型的基本思想是: 在特征词之间相互独立的前提假设下, 依据极大后验假设将未分类的文本划分到最可能属于的类别中去. 对某文本 $x_i$, 后验概率最大时类别的取值即为该文本的分类结果, 用 $c_i$ 表示:

$$c_i = \arg \max_{k=1,2,\cdots,K} P(c_k) \prod_{j=1}^{n} P(t_j | c_k). \tag{5}$$

然而, 当各类之间文本数分布不均衡时, 文本数较少的类别特征可能会被文本数多的类别特征湮灭, 使得后验概率偏向文本数较多的类别, 且各类别中每个特征词的重要程度也有所不同. 针对这些不足, 本文采用加权朴素贝叶斯文本分类模型, 即利用 TF-IDF 算法对各个特征词的条件概率进行加权, 从而反映各个特征词之间重要程度的差异, 并在一定程度上弱化条件独立假设 [34]. 根据上节介绍, TF-IDF 算法在分类问题中可进一步优化, 故本文利用改进的 TF-IDF 算法对各类特征词的条件概率进行加权, 将权重引入公式 (5) 中:

$$\omega'(t_j, c_k) = \prod_{i=1}^{m} \omega'_{ijk},$$

$$P(x_i | c_k) = \prod_{j=1}^{n} \left[ P(t_j | c_k)\, \omega'(t_j, c_k) \right].$$

相应的最大后验概率对应的类别为:

$$c_i = \arg \max_{k=1,2,\cdots,K} P(c_k) \prod_{j=1}^{n} \left[ P(t_j | c_k)\, \omega'(t_j, c_k) \right].$$

### 4.3 LSTM 模型

在本研究中, 股票收益往往会受到历史数据的影响, 此时传统的神经网络模型, 如 BP 神经网络模型无法记忆数据之前存储的信息, 而循环型神经网络 (Recurrent Neural Network, RNN) 允许数据在网络中向前或向后流动, 解决了数据的前后关联问题. RNN 模型的最大特点是模型当前的输出不仅依赖于模型的输入, 还与前面时刻的输出有关, 隐藏层的各个神经元不是无连接的, 如图 1.

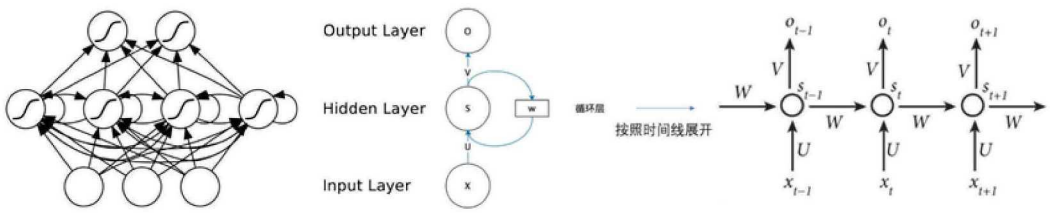


图 1　循环神经网络

图 1 中, $X$ 代表输入层, $S$ 为隐藏层, $O$ 为输出层. 隐藏层会在各时刻之间, 通过 $W$ 建立时间上的联系, 其中 $W$ 是每个时间点之间的权重矩阵. 可以看到, $t$ 时刻隐藏层的输出 $S_t$ 除了与输入层 $X_t$ 有关, 还受到上一时刻隐藏层的输出 $S_{t-1}$ 的影响.

在实际应用中, 传统 RNN 模型往往会因为层数过多出现梯度不稳定的问题, 该问题本质上是由于导数的链式法则导致了连乘的形式. 长短期记忆神经网络 (Long Short-Term Memory, LSTM) 模型是一种特殊的 RNN 模型, 可以有选择的学习长期的、有价值的信息, 它使用累加的形式计算状态, 避免了梯度消失. 此外, RNN 模型的隐藏层只是一些可以储存过去信息的简单记忆细胞, 而 LSTM 的记忆细胞多了三道门, 分别称为输入门, 输出门和遗忘门. 输入门决定了某时刻是否有信息输入记忆细胞, 某时刻是否有信息从记忆细胞中输出取决于输出门, 遗忘门的作用是控制某时刻记忆细胞里的信息是否被遗忘, 如果信息无意义, 模型会选择遗忘该信息. 三道门相当于给隐藏层加入了三个示性函数, 从而对输入模型的信息进行了筛选, 模型效率更高, 拟合效果更好. 图 2 反映了 LSTM 模型的隐藏层变化.

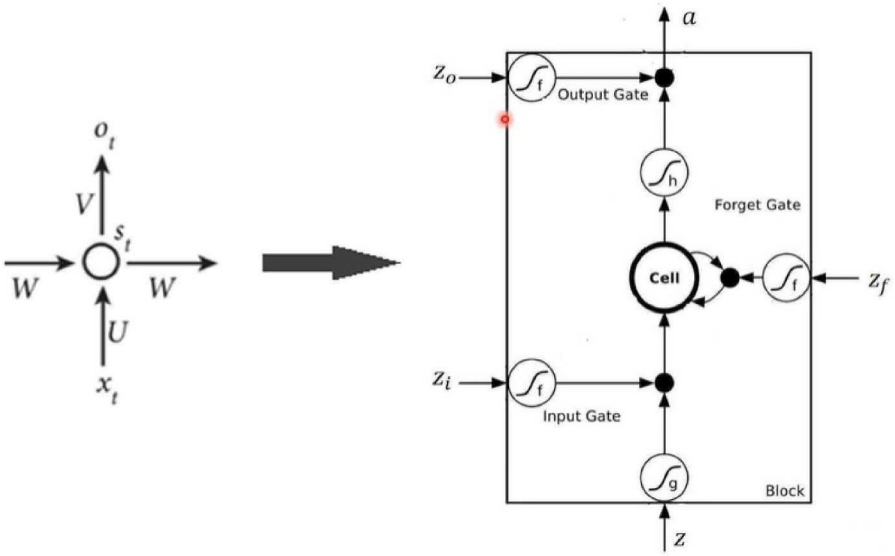


图 2 LSTM 隐藏层

图 2 中, LSTM 模型的输出值为 $a$, 与 RNN 模型的 $O$ 存在差异, $Z$, $Z_i$, $Z_f$, $Z_o$ 均有输入层 $x_t$ 的参与, 计算公式为:

$$Z = tanh\left(W\left[x_t, h_{t-1}\right]\right),$$
$$Z_i = \sigma\left(W_i\left[x_t, h_{t-1}\right]\right),$$
$$Z_f = \sigma\left(W_f\left[x_t, h_{t-1}\right]\right),$$
$$Z_o = \sigma\left(W_0\left[x_t, h_{t-1}\right]\right),$$

其中, $Z$ 是通过时刻 $t$ 的输入 $x_t$ 和上一时刻储存在记忆细胞里的隐藏层信息向量 $h_{t-1}$ 拼接后与权重参数向量 $W$ 点积, 再经过激活函数 tanh 得到的数值. 其他三个公式为门控装置, 使用的激活函数为 sigmoid 函数, 得到一个 0-1 之间的数值, 用来作为输入门的控制信号, 1 表示该门完全打开, 0 表示该门完全关闭.

## 5 研究设计

### 5.1 数据来源及预处理

研究针对 A 股市场, 从上证 50 的所有成分股中选取部分股票构成股票池. 根据 Liu 等[19] 的研究, 为去除 “壳污染”, 将所有股票按 2020 年 6 月末的总市值由高到低排序, 删去总市值后 30% 的股票. 再剔除数据缺失严重、上市时间短和现已退市的股票, 最终保留了 16 只股票进行研究, 见表 1.

**表 1 股票池股票**

| 股票代码 | 股票名称 | 股票代码 | 股票名称 |
|---|---|---|---|
| 600519 | 贵州茅台 | 601166 | 兴业银行 |
| 601398 | 工商银行 | 600030 | 中信证券 |
| 601318 | 中国平安 | 601088 | 中国神华 |
| 601628 | 中国人寿 | 601601 | 中国太保 |
| 601288 | 农业银行 | 600000 | 浦发银行 |
| 600036 | 招商银行 | 600887 | 伊利股份 |
| 601857 | 中国石油 | 600309 | 万华化学 |
| 600028 | 中国石化 | 600016 | 民生银行 |

本文从东方财富股吧中爬取了表 1 中每只股票 2020 年 7 月至 2021 年 2 月的评论数据, 包括每日的评论内容、评论时间和评论浏览量, 再从 wind 数据库中获取各股票对应日期的财务数据, 包括收益率、市盈率和总市值, 并将所有数据分为训练集和测试集.

对文本数据进行预处理. 首先, 将文本数据按其包含的情感倾向分为积极情绪股评和消极情绪股评两类. 然后进行分词, 为提高分词效果, 将搜狗词库 “股票基金词库大全”、“A 股简称” 以及北京大学国家发展研究院构建的 “中国金融情绪词典 (GB)” 加入到原有的中文词典中作为分词词典. 最后删除文本数据中的停用词. 处理后的股评变为多个词组成的词向量. 分别筛选出两类股评中的高频词, 再利用卡方检验进行特征选择, 最终保留 24 个积极类特征词和 24 个消极类特征词.

### 5.2 变量选取与构建

1) 投资者情绪因子

利用改进的加权朴素贝叶斯文本分类模型, 分别对股票池中 16 只股票每日的评论进行情感分类, 并统计分类结果中积极类和消极类股评的个数. 基于统计结果, 本文根据 Mao 等[35] 的做法构建了情绪因子:

$$MS_t = \ln\left(\frac{1+P_t}{1+N_t}\right), \tag{6}$$

其中 $P_t$ 为 $t$ 日积极类股评的个数, $N_t$ 为 $t$ 日消极类股评的个数. 为体现每条股评的影响力差异, 使浏览量高的帖子在情感上更具有代表性, 本文基于浏览量对不同影响力的股评赋予了不同的权重 $w_i$:

$$w_i = \begin{cases} 4, & \text{if } u_i > q_3, \\ 3, & \text{if } q_2 < u_i \le q_3, \\ 2, & \text{if } q_1 < u_i \le q_2, \\ 1, & \text{if } u_i \le q_1, \end{cases}$$

其中, $u_i$ 表示股评的浏览量, $q_1$, $q_2$, $q_3$ 为当天所有股评浏览量的 1/4 分位数、中位数和 3/4 分位数. 将权重引入公式 (6), 可得情绪因子为

$$WMS_t = \ln\left(\frac{1+BP_t}{1+BN_t}\right), \tag{7}$$

$$BP_t = \sum_i^N w_{it} I_{\{positive\ \ comment\}}, \quad BN_t = \sum_i^N w_{it} I_{\{negative\ \ comment\}}.$$

2) 其他影响因子

除投资者情绪因子外, 股票市场中还存在着许多影响股票收益率的金融指标, 例如账面市值、股票市值、市盈率等. Fama 和 French [10] 选取了多个可能影响股票收益的金融指标进行研究, 并通过 Fama-MacBet 多变量回归模型比较了这些指标的拟合效果. 拟合效果越好, 说明该指标对股票收益的解释能力越强. Liu 等 [19] 按照相同的思路构建了中国版 Fama-French 三因子模型, 即利用 Fama-MacBeth 多因子回归模型比较了多个指标对股票收益率解释能力大小. 研究结果显示, 市盈率 (Earnings-to-Price, EP)、市场价值 (Assets-to-Market, AM) 对股票收益率的解释程度较高. 本文基于中国版 Fama-French 三因子模型的思想选取了市场价值、市盈率及市场组合收益率构建了影响股票收益的三个影响因子——规模因子、价值因子和市场资产组合因子.

规模因子 (Small Minus Big, SMB) 反映的是上市公司的市场价值大小引发的股票收益率差异, 可用 AM 表示. 当研究对象为多只股票的组合收益率时, SMB 的计算方法如下: 将需要研究的股票按 AM 的大小平均分为两组 (Small、Big), 再按 EP 由大到小分为三组, 即前 30%(Value), 中间 40%(Neutral), 后 30%(Growth). 最终形成六个组, 如表 2 所示. 至此, 为了度量大规模投资组合 (B/V、B/M、B/G 组) 和小规模投资组合 (S/V、S/M、S/G 组) 市值的差异, 按如下公式进行计算,

$$SMB = \frac{1}{3}(S/V + S/M + S/G) - \frac{1}{3}(B/V + B/M + B/G), \tag{8}$$

其中, S/V 的取值为市值后 50% 且市盈率前 30% 股票的市值加权组合收益率, 其余类似.

**表 2 因子构建分组标准**

| | | EP | | |
|---|---|---|---|---|
| | | Value (top 30%) | Middle (middle 40%) | Growth (bottom 30%) |
| AM | small (bottom 50%) | S/V | S/M | S/G |
| | big (top 50%) | B/V | B/M | B/G |

价值因子 (Value Minus Growth, VMG) 衡量的是上市公司由于市盈率的不同导致的收益率差异, 可用 EP 进行计算. 当研究对象为多只股票的组合收益率时, VMG 的计算方法与 SMG 类似. 首先按照表 2 的方法将投资组合中的股票进行分组, 然后计算高市盈率投资组合和低市盈率投资组合的市值加权组合收益率, 公式如下所示:

$$VMG = \frac{1}{2}(S/V + B/V) - \frac{1}{2}(S/G + B/G). \tag{9}$$

市场资产组合因子 (Market, MKT) 反映了股票市场整体的收益情况. 本文利用股票池中所有股票的市值加权组合收益率 $R_{mkt}$ 与市场无风险收益率 $R_f$ 的差表示, 即超额收益率, 其中, 市场无风险收益率为中国人民银行公布的人民币三个月整存整取利率除以 365. MKT

的计算公式如下所示:

$$MKT = R_{mkt} - R_f. \tag{10}$$

## 6 实证分析

### 6.1 股评文本分类的实证分析

针对训练集文本数据, 由每日积极类和消极类股评的文本数及两类股评含特征词的文本数可得先验概率和条件概率, 再利用朴素贝叶斯文本分类模型对测试集的股票评论进行自动分类, 得到混淆矩阵, 如表 3 所示. 由表 3 可知, 该模型的分类正确率为 72.08%, 召回率为 72.12%, 查准率为 71.77%, $F$ 检验统计量为 0.72, 分类效果良好.

**表 3　测试集股评的朴素贝叶斯文本情感分类情况**

| | 实际属于积极类 | 实际属于消极类 | 合计 |
|---|---|---|---|
| 预测属于积极类 | 3443 | 1760 | 5203 |
| 预测属于消极类 | 1316 | 4500 | 5816 |
| 合计 | 4759 | 6260 | 11019 |

下面利用加权朴素贝叶斯文本分类模型和改进的加权朴素贝叶斯文本分类模型进行分类, 其中, TF-IDF 权重和改进的 TF-IDF 权重可由公式 (1) (4) 得到, 部分结果见表 4. 可以看出, 改进后的权重使得各个特征词的重要性差异更明显.

**表 4　改进前后的 TF-IDF 权重**

| 特征词 | 传统 TF-IDF 权重 $\omega_j$ | | | | 改进的 TF-IDF 权重 $\omega_j'$ | | | |
|---|---|---|---|---|---|---|---|---|
| | 1.4 | 1.5 | 12.3 | 12.31 | 1.4 | 1.5 | 12.3 | 12.31 |
| 上涨 | 0.037 | 0.032 | 0.031 | 0.053 | 0.105 | 0.145 | 0.182 | 0.123 |
| 加油 | 0.04 | 0.028 | 0.036 | 0.029 | 0.126 | 0.113 | 0.177 | 0.071 |
| 利好 | 0.029 | 0.013 | 0.015 | 0.132 | 0.085 | 0.085 | 0.152 | 0.429 |
| 突破 | 0.034 | 0.037 | 0.026 | 0.005 | 0.125 | 0.137 | 0.102 | 0.02 |
| 牛市 | 0.022 | 0.006 | 0.038 | 0.014 | 0.049 | 0.037 | 0.204 | 0.059 |
| 垃圾 | 0.011 | 0.023 | 0.071 | 0.111 | 0.074 | 0.057 | 0.167 | 0.128 |
| 跑路 | 0.013 | 0.112 | 0.044 | 0.012 | 0.076 | 0.027 | 0.122 | 0.077 |
| 韭菜 | 0.023 | 0.056 | 0.058 | 0.023 | 0.188 | 0.181 | 0.158 | 0.147 |
| 亏损 | 0.054 | 0.066 | 0.062 | 0.043 | 0.451 | 0.161 | 0.143 | 0.21 |
| 倒闭 | 0.066 | 0.014 | 0.107 | 0.022 | 0.343 | 0.032 | 0.248 | 0.157 |

为进一步检验加权朴素贝叶斯模型的优越性, 本文将 SVM 模型应用于同样的训练集, 并对同样的测试集进行了文本分类. 下面对各个模型的分类效果进行对比, 见表 5. 结果显示, 改进的加权朴素贝叶斯分类模型的分类性能最优, 对加权方式的改进是有效的.

**表 5　文本分类模型分类性能对比**

| | Recall | Precession | Accuracy | $F$ |
|---|---|---|---|---|
| 朴素贝叶斯分类模型 | 72.12% | 71.77% | 72.08% | 0.7194 |
| 加权朴素贝叶斯分类模型 ($\omega_j$) | 76.80% | 73.86% | 75.87% | 0.753 |
| 改进的加权朴素贝叶斯分类模型 ($\omega_j'$) | 80.21% | 79.86% | 79.87% | 0.8003 |
| SVM 文本分类模型 | 80.06% | 79.70% | 78.95% | 0.7988 |

### 6.2 投资者情绪对股票收益率的线性影响实证分析

1) 模型设计

本文根据 Liu 等 [19] 的做法构建了中国版 Fama-French 三因子模型, 即

$$R_t - R_f = a + b_1 MKT_t + b_2 SMB_t + b_3 VMG_t + \varepsilon_t, \tag{11}$$

其中, $R_t$ 表示 $t$ 日收盘时股票的收益率, $R_f = 0.003014$, $MKT_t$、$SMB_t$、$VMG_t$ 分别表示 $t$ 日开盘时市场资产组合因子、规模因子、价值因子的值, $a$、$b$ 为待估参数, $\varepsilon_t$ 为误差项.

为了研究投资者情绪对股票收益率的影响, 本文在上述三因子模型的基础上加入投资者情绪因子 $MS_t$, 得到中国版 Fama-French 四因子模型:

$$R_t - R_f = a + b_1 MKT_t + b_2 SMB_t + b_3 VMG_t + MS_t + \varepsilon_t. \tag{12}$$

进一步, 为了考察基于浏览量加权的投资者情绪因子对股票收益率的预测能力是否优于未加权的情绪因子, 将 (12) 式中的 $MS_t$ 替换为 $WMS_t$, 即

$$R_t - R_f = a + b_1 MKT_t + b_2 SMB_t + b_3 VMG_t + WMS_t + \varepsilon_t. \tag{13}$$

事实上, 投资者多数情况下会选择同时投资多只股票, 故本文分别研究了购买单只股票、购买股票投资组合两种投资策略下的收益率, 其中股票投资组合收益率采用最优投资组合收益率. 本文在马柯维茨均值方差模型的假设下, 计算了持股期为 1 个月时股票池中所有股票每日的最优投资组合收益率 $R\_best_t$, 并将公式 (12)、(13) 中的因变量进行替换, 得到

$$R\_best_t - R_f = a + b_1 MKT_t + b_2 SMB_t + b_3 VMG_t + MS_t + \varepsilon_t, \tag{14}$$

$$R\_best_t - R_f = a + b_1 MKT_t + b_2 SMB_t + b_3 VMG_t + MS_t + \varepsilon_t. \tag{15}$$

2) 回归结果

首先计算模型中所需的变量的值. 根据公式 (6)、(7) 的计算方法, 可得各股票加权前后的投资者情绪因子值, 对于多只股票的综合情绪因子, 将各股票的情绪因子按照股评数占比相加得到. 再计算持股期为一个月时每日的最优投资组合收益率以及相应的股票超额收益率, 最终得到单只股票及最优投资组合的超额收益率. 在进行拟合之前, 为了避免出现伪回归, 先对各变量进行平稳性检验. 绘制时序图, 初步判定各个因子的时间序列数据是平稳的, 如图 3. 再进行 ADF 检验, 发现除情绪因子外, 其余因子的 $P$ 值均小于 0.01, 通过 ADF 检验. 对于情绪因子, 根据 VAR 模型滞后阶数的确定方法, FPE、AIC、SC 及 HQ 准则均显示滞后一阶的模型是平稳的, 故认为情绪因子时间序列数据的最优滞后阶数为 1 阶. 由此, 在下面的研究中考虑前一期的情绪因子.

下面以浦发银行股票收益率为例, 利用多元线性回归模型对公式 (11)、(12)、(13) 进行线性拟合, 结果如表 6.

由表 6 可知, 三个线性回归模型的调整 $R^2$ 值和 $F$ 检验都显著, 即回归方程是显著的. 加入情绪因子后的回归模型拟合效果更佳, 调整 $R^2$ 的值由 0.686 提高到 0.8 以上, 说明投资者情绪因子对股票收益率有一定解释能力. 当情绪因子取 MS 时, 四个因子对股票超额收益率的影响都是正向的, 市场因子和情绪因子的影响较大, 但情绪因子 $t$ 检验表现不佳. 当情绪因子取 WMS 时, 四个因子同样对股票超额收益率有正向影响, 但 WMS 的 $t$ 检验依旧不理想. 对比前两个回归模型的调整 $R^2$ 值可以看出, 情绪因子取 WMS 时, 模型解释能力稍强, 但增强效果并不明显, 不能很好的说明 WMS 的优越性. 由于情绪因子的 $t$ 检验表现差, 故进行回

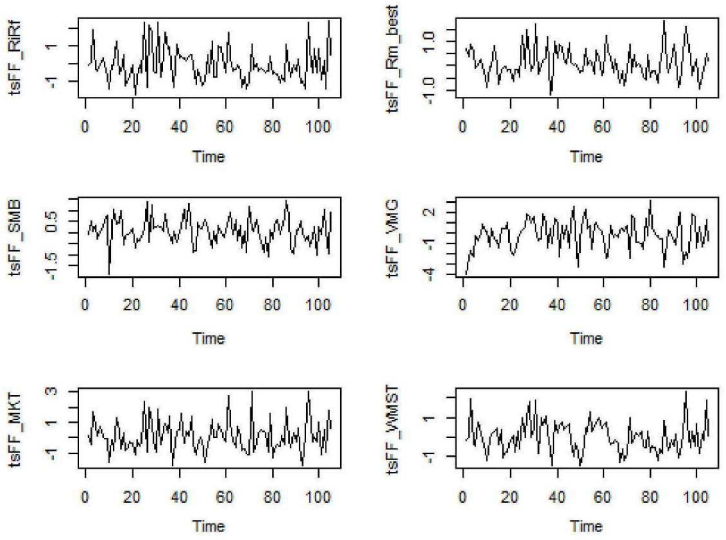


图 3 各变量时序图

**表 6 浦发银行超额收益率与四因子的线性拟合结果**

| 因子 | 系数 | 标准差 | $t$ 值 | $P$ 值 | $F$ 检验 –$P$ 值 | 调整 $R^2$ |
|---|---|---|---|---|---|---|
| Panel A | 情绪因子 MS、三因子与股票收益率回归 | | | | | |
| 常数 | –0.217 | 0.068 | –3.184 | 0.001 ** | 0.000*** | 0.803 |
| SMB | 0.181 | 0.072 | 2.521 | 0.013 * | | |
| VMG | 0.252 | 0.033 | 7.563 | 0.000*** | | |
| MKT | 0.831 | 0.061 | 13.59 | <0.000*** | | |
| MS | 0.399 | 0.231 | 1.726 | 0.087 . | | |
| Panel B | 情绪因子 WMS、三因子与股票收益率回归 | | | | | |
| 常数 | –0.166 | 0.057 | –2.886 | 0.004** | 0.000*** | 0. 848 |
| SMB | 0.206 | 0.075 | 2.735 | 0.007** | | |
| VMG | 0.252 | 0.035 | 7.152 | 0.000*** | | |
| MKT | 0.851 | 0.061 | 14.057 | <0.000*** | | |
| WMS | 0.216 | 0.155 | 1.397 | 0.106 . | | |
| Panel C | 三因子与股票收益率回归 | | | | | |
| 常数 | –0.114 | 0.044 | –2.583 | 0.011* | 0.000*** | 0.686 |
| SMB | 0.198 | 0.075 | 2.626 | 0.009** | | |
| VMG | 0.263 | 0.035 | 7.593 | 0.000*** | | |
| MKT | 0.904 | 0.047 | 19.293 | <0.000*** | | |

归诊断, 诊断结果显示模型不存在异方差性、自相关性和多重共线性. 然而, 股票超额收益率与情绪因子的皮尔逊相关系数仅有 0.37, 认为两者之间线性关系弱. 当因变量采用最优投资

组合超额收益率时, 也出现了同样的问题, 如表 7 所示.

**表 7 最优投资组合超额收益率与四因子的线性拟合结果**

| 因子 | 系数 | 标准差 | $t$ 值 | $P$ 值 | $F$ 检验 –$P$ 值 | 调整 $R^2$ |
|---|---|---|---|---|---|---|
| Panel D 情绪因子 MS、三因子与股票最优投资组合收益率回归 | | | | | | |
| 常数 | –0.008 | 0.056 | –0.143 | 0.887 | 0.000*** | 0.75 |
| SMB | 0.122 | 0.06 | 2.029 | 0.045 * | | |
| VMG | 0.226 | 0.027 | 2.812 | 0.041 * | | |
| MKT | 0.427 | 0.051 | 13.59 | <0.000 *** | | |
| MS | 0.307 | 0.193 | 1.683 | 0.091 . | | |
| Panel E 情绪因子 WMS、三因子与股票最优投资组合收益率回归 | | | | | | |
| 常数 | 0.069 | 0.035 | 1.971 | 0.051 . | 0.000*** | 0. 883 |
| SMB | 0.107 | 0.064 | 4.735 | 0.000 * | | |
| VMG | 0.252 | 0.035 | 0.478 | 0.633 | | |
| MKT | 0.851 | 0.061 | 6.851 | <0.000 *** | | |
| WMS | 0.216 | 0.155 | 3.645 | 0.002 * | | |

事实上, 最优投资组合收益率的拟合模型效果较差, 从表 7 的回归结果中可以看出, 情绪因子对最优投资组合股票收益的解释程度增加, 其他因子的解释能力减弱, 最终各个因子对股票收益的解释能力均处于较低水平. 下面考虑采用非线性模型进行拟合.

### 6.3 投资者情绪对股票收益率的非线性影响实证分析

1) 模型设计

本文采用了机器学习技术研究投资者情绪对股票收益率的非线性影响. 与线性模型相比, 非线性模型更关注拟合结果, 而非模型的渐进影响机制. 基于中国版四因子模型, 可构建出相应的非线性拟合模型

$$R_t - R_f = F\left(MKT_t,\ SMB_t,\ VMG_t,\ MS_t\left(WMS_t\right)|A,\ B,\ \varepsilon\right), \tag{16}$$

其中 $R_t$ 表示 $t$ 日收盘时股票的收益率, $A$, $B$ 为待估参数, $F$ 为非线性映射, $\varepsilon$ 为误差项.

为检验非线性模型能否提高股票收益率的预测精度, 以及对情绪因子的改进是否有效, 基于公式 (16), 本文设计以下四个非线性拟合模型:

MI1: 情绪因子取 MS, 输出变量为某只股票超额收益率;

MI2: 情绪因子取 WMS, 输出变量为某只股票超额收益率;

MP1: 情绪因子取 MS, 输出变量为最优投资组合超额收益率;

MP2: 情绪因子取 WMS, 输出变量为最优投资组合超额收益率.

下面基于 LSTM 模型构建深度学习网络, 在 python 语言环境中构建三层 LSTM 层和一层全连接层, 方向为单向, 采用 Adam 优化器进行优化训练, 再采用 Keras 中的回调函数 ReduceLR- OnPlateau 在训练过程中优化学习率, 最大 epoch 为 100, 最大训练步数为 $2\times10^4$.

2) 拟合结果

对于模型 MI1 和 MP1, 即当情绪因子取 MS 时, 拟合结果见图 4. 图 4 左图为单只股票

超额收益率的拟合情况, 右图为最优投资组合超额收益率的拟合情况. 可以看出, 模型拟合效果较好, 尽管股票超额收益率的预测值相比实际值整体偏大, 但走势基本一致. 对于模型 MI2 和 MP2, 即当情绪因子取 WMS 时, 拟合情况如图 5 所示. 相比图 4, 拟合效果更佳, 预测值在走势与实际值基本一致的前提下更接近实际值.

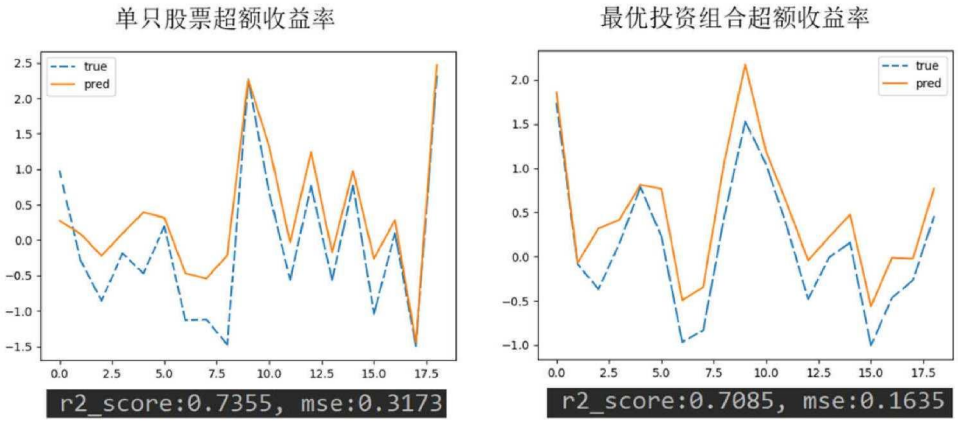


图 4　单只股票/最优投资组合模型超额收益率预测值与实际值 (MI1/MP1)

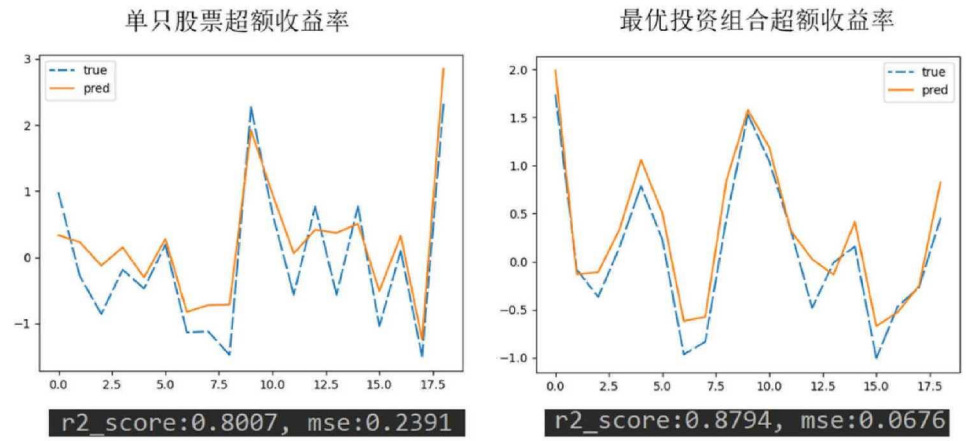


图 5　单只股票/最优投资组合模型超额收益率预测值与实际值 (MI2/MP2)

进一步考虑滞后一阶的情绪因子对收益率的影响, 将 $t-1$ 日的情绪因子纳入模型. 由于股票市场的敏感性和多变性, 前一日的投资者情绪对当日股票的收益率的影响强度通常低于当日的投资者情绪, 故将投资者情绪因子重新定义为:

$$WMS'_t = \frac{2}{3}WMS_t + \frac{1}{3}WMS_{t-1}.$$

由此, 在 MI2、MP2 的基础上, 可以得到以下两个模型:

MI3: 情绪因子取 $WMS'$, 输出变量为某只股票超额收益率;

MP3: 情绪因子取 $WMS'$, 输出变量为最优投资组合超额收益率.

利用 LSTM 模型对 MI3、MI3 进行拟合, 模型预测结果见图 6, 在两种投资策略下, 加入前一日的投资者情绪能使得模型对股票收益率的预测更准确, 前一日的情绪因子会对当日的

股票收益率产生显著影响.

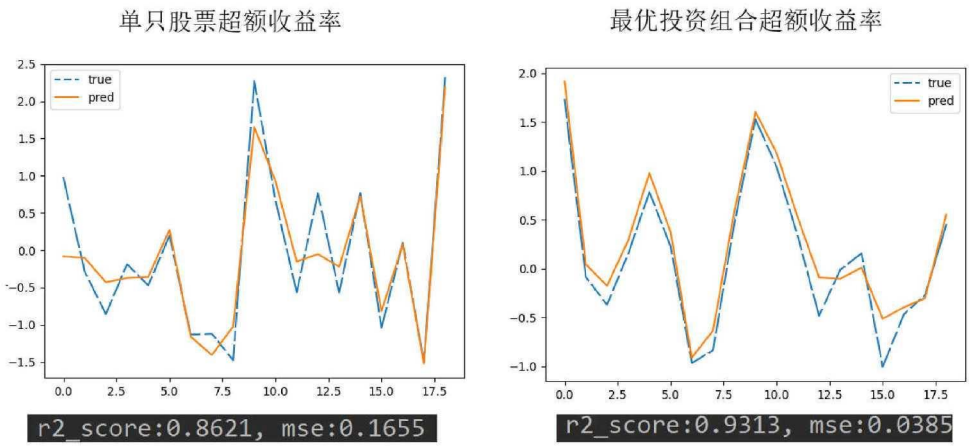


图 6 单只股票/最优投资组合模型超额收益率预测值与实际值 (MI3/MP3)

相比传统的神经网络模型, LSTM 模型具有更好的预测效果. 这是因为传统神经网络模型, 如 BP 神经网络、卷积神经网络等, 较依赖于研究者对模型的设计, 具有一定的主观性. 下面以 BP 神经网络模型为例, 对 MI2、MP2 进行拟合, 预测效果见图 7. 可以看出, 尽管预测值与实际值大小近似, 但误差相对较大, 且两者走势不一, 预期收益率有时高于真实的收益率, 有时低于真实的收益率. 这不利于投资者观察预期收益率的走势, 模型的应用性较差.

进一步, 将情绪因子剔除, 仅对其余三个因子进行非线性拟合, 拟合效果如图 8. 显然, 模型拟合效果相对较差, 情绪因子的确会对股票的超额收益率产生显著的影响.

为更好的比较上述模型的预测效果, 下面计算模型的样本外均方误差 MSE 及决定系数 $R^2$. 预测值与实际值相差越小, MSE 的值越小, $R^2$ 的值越大, 模型拟合效果越好. 由表 8 可知不同收益率预测模型的均方误差和决定系数的表现情况.

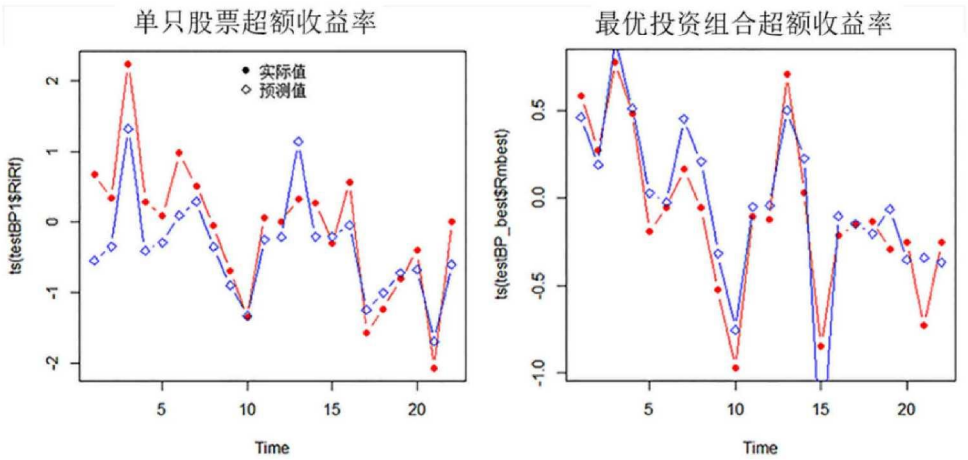


图 7 单只股票/最优投资组合超额收益率预测值与实际值 (MI2/MP2, BP 神经网络)

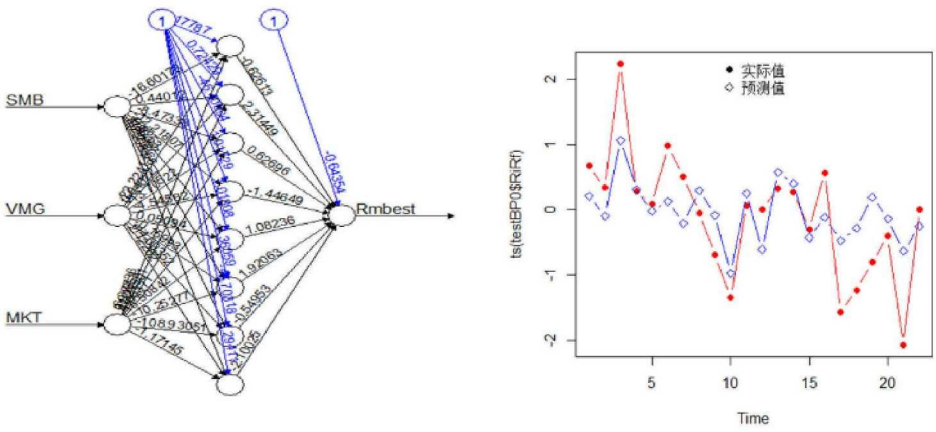


图 8　无情绪因子时模型超额收益率预测值与实际值 (BP 神经网络)

**表 8　预测模型评价指标对比**

| 模型 | | MI1 | MI2 | MI3 | MP1 | MP2 | MP3 |
|---|---|---|---|---|---|---|---|
| 多元线性回归模型 | MSE | 0.3817 | 0.332 | – | 0.5231 | 0.3208 | – |
| | $R^2$ | 0.6336 | 0.6954 | – | 0.6183 | 0.6322 | – |
| BP 神经网络 | MSE | 0.3872 | 0.3416 | – | 0.3012 | 0.2593 | – |
| | $R^2$ | 0.6114 | 0.6285 | – | 0.6517 | 0.6788 | – |
| LSTM 模型 | MSE | 0.3173 | 0.2391 | 0.1655 | 0.1635 | 0.0676 | 0.0385 |
| | $R^2$ | 0.7355 | 0,8007 | 0.8621 | 0,7085 | 0.8794 | 0.9313 |

实验结果证明, 投资者情绪对股票收益具有显著的正向影响, 假设 H1 成立. 对于情绪因子的取值, 采用加权情绪因子的模型预测效果普遍优于未加权的情绪因子. 以 LSTM 模型为例, 模型 MI2 的均方误差为 0.2391, 低于模型 MI1 的 0.3173, 决定系数为 0.8007, 高于模型 MI1 的 0.7355; 模型 MP2 的均方误差 (0.0676) 和决定系数 (0.8794) 也优于模型 MP1 的均方误差 (0,1635) 和决定系数 (0,7085). 因此, 在研究股票超额收益率时, 应采用基于浏览量加权的情绪因子度量投资者情绪, 假设 H2 成立.

对于假设 H3, 横向对比表 8 的结果可知, 在两种不同的投资策略下, 非线性模型普遍优于线性模型, 具有更好的预测效果, 且 LSTM 模型表现最优. 因此, 本文认为投资者情绪对股票收益率具有非线性影响, 基于 LSTM 构建的非线性模型具有更高的预测精度, 假设 H3 成立.

加入前一日的情绪因子时, 模型的均方误差和决定系数表现更优, 且在两种投资策略中, 采用 LSTM 模型拟合的 MI3、MP3 预测准确率分别是最高的. 因此, 在研究股票超额收益率时, 需考虑前一时刻的情绪因子, 假设 H4 成立.

观察表 8 的 MP1、MP2、MP3 列可知, 当投资者的持股期为一个月时, 投资者情绪对其持有投资组合的收益率具有显著的正向影响, 且这种影响为非线性的, 前一时刻的投资者情绪也会影响当期的投资组合超额收益, 假设 H5 成立.

## 7 稳健性检验

为保证本文结论的稳健性, 对研究过程进行以下调整: 第一, 爬取同一时间段的上证指数、深证综指的股票收益率、市盈率、总市值及东方财富股吧上证指数吧、深证股吧的评论信息代替本文数据. 第二, 本文按 8:2 的比例分离了训练集、测试集, 将该比例更换为 7:3 后按照相同的步骤重新进行数据分析. 调整后的研究结论与前文基本一致, 说明本文对模型的改进及研究结果具有一定的稳健性.

## 8 结论与建议

本文从东方财富股吧中提取了投资者的情绪信息, 在现有研究的基础上建立了更准确的情绪因子, 结合上证股票的金融指标数据, 研究了投资者情绪对股票收益率的影响, 并构建了股票收益率的预测模型. 具体结论如下: 1) 在短期内, 投资者情绪对股票收益率具有正向影响. 2) 投资者情绪的变化对股票收益率的波动均具有显著的非线性影响, 非线性模型在预测股票收益率时具有更高的预测精度. 投资者在预测未来收益时, 使用线性模型有时会使得投资者情绪失去意义, 忽略市场中的非理性因素, 投资者对收益率的预期出现偏差. 3) 将前一日的情绪因子纳入非线性预测模型时, 模型的预测效果更佳. 情绪作为一种信息, 并非是立即失效的, 而是被储存在股票市场中, 持续对股票的价格和收益率产生影响. 投资者在关注股票收益率时, 需考虑投资者情绪产生的滞后影响.

在情绪因子构建的研究中, 本文得到了以下结论: 1) 考虑股票评论在社交平台中的影响力差异能够更好地度量投资者情绪. 在构建投资者情绪因子时, 依据每个股评的浏览量对股评进行加权, 这样构造出的情绪因子相比传统情绪因子对股票收益的预测效果更佳. 2) 加权朴素贝叶斯文本分类模型可以有效的改善传统模型的缺陷, 进一步改进该模型的加权方式可以继续优化分类效果. 分类效果的优化是提高情绪因子准确性的基础, 随着机器学习技术的发展, 股票评论的分类效率将持续改善.

随着经济的发展以及社会的不断进步, 国家和人民对金融风险控制的要求不断提高. 在这个背景下, 中小投资者应不断学习金融知识, 树立理性识别各类信息的意识, 采用科学方法调整投资决策. 机构投资者应充分利用自身的资金和技术优势, 不断更新构造投资组合的技术手段, 提高服务效率. 上市公司应完善信息披露制度, 提高公司信用, 发布切实可靠的信息, 做到让投资者安心. 市场监管部门需随时关注网络舆情, 在股市波动较大时做到及时预警和正确引导, 提高投资者信心, 加强投资者教育, 顺应媒体传播的新形势、新变化, 积极营造良好的投资环境, 激发股市活力.

### 参考文献

[1] De Long J B, Shleifer A, Summers L H, et al. Noise trader risk in financial markets[J]. Journal of Political Economy, 1990, 98(4): 703-738.

[2] Baker M, Wurgler J. Investor sentiment and the cross-section of stock returns[J]. Journal of Finance, 2006, 61(4): 1645-1680.

[3] 皇甫玉婷, 刘澄, 王未卿. 投资者情绪对资本市场稳定性影响的实证分析 [J]. 统计与决策, 2019, 35(14): 174-177.

[4] 刘婧, 温雅丽, 许金玲. 公司治理、投资者情绪与盈余反应 [J]. 统计与决策, 2020, 36(7): 154-157.
[5] 王美今, 孙建军. 中国股市收益、收益波动与投资者情绪 [J]. 经济研究, 2004, 10: 75-83.
[6] 张信东, 原东良. 基于微博的投资者情绪对股票市场影响研究 [J]. 情报杂志, 2017, 36(8): 81-87.
[7] 金秀, 姜尚伟, 苑莹. 基于股吧信息的投资者情绪与极端收益的可预测性研究 [J]. 管理评论, 2018, 30(7): 18-27.
[8] 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.
[9] 石勇, 唐静, 郭琨. 社交媒体投资者关注、投资者情绪对中国股票市场的影响 [J]. 中央财经大学学报, 2017(7): 45-53.
[10] Fama E F, French K R. Common risk factors in the returns on stocks and bonds[J]. Journal of Financial Economics, 1993, 33(1): 3-56.
[11] 樊鹏英, 杨音, 张正平, 陈敏. 个股投资者情绪与股票收益率的关系——基于股评信息视角的研究 [J]. 数学的实践与认识, 2021, 51(16): 305-320.
[12] 饶育蕾, 王攀. 媒体关注度对新股表现的影响——来自中国股票市场的证据 [J]. 财务与金融, 2010(3): 1-7.
[13] 张前程, 杨德才. 货币政策、投资者情绪与企业投资行为 [J]. 中央财经大学学报, 2015(12): 57-68.
[14] 张博, 邑文秀, 杨熙安. 投资者情绪生成机理的研究 [J]. 中国管理科学, 2021, 29(1): 185-195.
[15] 山立威. 心理还是实质: 汶川地震对中国资本市场的影响 [J]. 经济研究, 2011, 46(4): 121-134+146.
[16] 许天阳. 网络社交媒体中投资者情绪对股票市场的影响研究 [J]. 上海管理科学, 2018, 40(3): 67-74.
[17] 张艾莲, 郭升刚. 投资者情绪对异质股票市场的非对称影响 [J]. 统计与信息论坛, 2020, 35(4): 113-118.
[18] 李岩, 金德环. 投资者情绪与股票收益关系的实证检验 [J]. 统计与决策, 2018, 34(15): 166-169.
[19] Liu J N, Stambaugh R F, Yuan Y. Size and value in china[J]. Journal of Financial Economics, 2019, 134(1): 48-69.
[20] 黄润鹏, 左文明, 毕凌燕. 基于微博情绪信息的股票市场预测 [J]. 管理工程学报, 2015, 29(1): 47-52.
[21] 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.
[22] 黄婷婷, 余磊. SDAE-LSTM 模型在金融时间序列预测中的应用 [J]. 计算机工程与应用, 2019, 55(1): 142-148.
[23] 耿晶晶, 刘玉敏, 李洋等. 基于 CNN-LSTM 的股票指数预测模型 [J]. 统计与决策, 2021, 37(5): 134-138.
[24] 潘水阳, 刘俊玮, 王一鸣. 基于神经网络的股票收益率预测研究 [J]. 浙江大学学报 (理学版), 2019, 46(5): 550-555.
[25] Froot K A, Scharfstein D S, Stein J C. Herd on the street: informational inefficiencies in a market with short-term speculation[J]. Journal of Finance, 1992, 47(4): 1461-1484.
[26] Noelle-Neumann E. The Spiral of silence: a theory of public opinion[J]. Journal of Communication, 1974, 24(2): 43-51.
[27] 曹红辉, 杨欣, 申慧. 股票市场非线性随机游走检验 [J]. 中央财经大学学报, 2003(4): 24-28.
[28] 牛晓健, 吴客形. 多层次资本市场条件下交易者互动对资产价格的影响——基于中国股票和债券市场的研究 [J]. 复旦学报 (社会科学版), 2020, 62(2): 180-191.
[29] 苏治, 卢曼, 李德轩. 深度学习的金融实证应用: 动态、贡献与展望 [J]. 金融研究, 2017, 443(5): 111-126.
[30] 赵胜民, 刘笑天. 引入投资者偏好的多因子模型——基于前景理论视角的分析 [J]. 中国经济问题, 2019(2): 106-121.
[31] 刘澜飚, 郭亮. 中国 A 股市场下行风险研究——基于广义失望厌恶行为视角 [J/OL]. 中国管理科学. https://doi.org/10.16381/j.cnki.issn1003-207x.2021.0820.
[32] Jones K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972, 28(1): 11-21.
[33] 隗中杰. 文本分类中 TF-IDF 权重计算方法改进 [J]. 软件导刊, 2018, 17(12): 39-42.

[34] 饶丽丽, 刘雄辉, 张东站. 基于特征相关的改进加权朴素贝叶斯分类算法 [J]. 厦门大学学报 (自然科学版), 2012, 51(4): 682-685.

[35] Mao H N, Counts S, Bollen J. Quantifying the effects of online bullishness on international financial markets[J]. European Central Bank-Statistics Paper series, 2015, 9: 1-23.

# A Research on the Influence of Investor Sentiment on A-share Returns Based on Stock Comments

MENG Xue, LI Yan-ru, HU Feng

(School of Statistics and Data Science, Qufu Normal University, Qufu 273165, China)

**Abstract:** The behavior of investors is susceptible to the influence of internet public opinion, then it will cause the fluctuation of stock returns. Analyzing the influence of investor sentiment on stock returns is helpful for investors to avoid investment risks and promote the stable development of China's stock market. Based on the comment data of Shanghai stock market from July 2020 to February 2021 of Oriental Fortune stock bar, this paper uses weighted naive Bayes classification model to construct investor sentiment factor, and improves the construction method of sentiment factor. Then the sentiment factor is introduced into the Chinese Fama-French three-factor model. Aiming to the single stock and one-month stock portfolio, the influence of investor sentiment on stock return rate is studied from linear and nonlinear perspectives based on the linear regression and long short-term memory model. The results show that investor sentiment has a nonlinear positive effect on stock return rate. In addition, investor sentiment of the previous moment will have an impact on the stock returns of the present moment, and investors need to take it into consideration when studying the expected return rate.

**Keywords:** investor sentiment; stock returns; weighted naive bayesian classification model; Chinese Fama-French three-factor model; long short-term memory model