第3卷 第4期 计 量 经 济 学 报 Vol. 3, No. 4
2023年10月 China Journal of Econometrics Oct., 2023

doi: 10.12012/CJoE2023-0061

另类数据在中国股票市场投资中有用吗？——基于财经短视频、图像、文本数据的探究

何 勇 [1], 李琪琪 [2], 焦 丽 [2], 黄文萱 [1]

(1. 山东大学金融研究院, 济南 250100; 2. 山东大学数学学院, 济南 250100)

摘 要 当前, 另类数据的应用为金融投资领域的学者和从业者提供了新的视角. 本文构建了基于因子增广回归与深度神经网络的预测模型, 实现了从财经类短视频和财经新闻等另类数据学习交易信号, 并构建了中国股票市场投资交易策略. 首先, 将抓取的财经新闻匹配相应股票代码, 并分解为文本数据和图像数据. 其次, 将文本数据输入文本数据学习框架, 求解因子回归方程计算新闻文本得分情况; 图像数据输入基于迁移学习搭建的图像识别深度神经网络模型, 计算图像情绪指数和图像得分. 对于抓取的财经类短视频, 包含两步, 第一步剥离音频数据转换为文本数据, 利用训练好的文本数据学习框架计算短视频文本得分; 第二步提取视频的关键帧, 利用训练好的图像模型计算视频图像得分; 文本得分与图像得分求和得到短视频数据得分. 最后将财经类短视频得分、新闻报道的文本得分和图像得分求和得到股票投资信号, 并将之作为构建投资组合的依据, 制定合适的投资策略. 研究结果表明, 财经类短视频和财经新闻数据中包含了与股价相关的信息, 能够有效预测市场变化, 并为投资者带来超额收益. 实证研究证实, 另类数据在中国市场中具有重要影响力. 通过对另类数据进行综合分析, 本文为投资者提供了一个全面且有效的交易信号提取方法, 有助于优化投资策略并实现更高的实际收益.

关键词 中国股票市场; 财经新闻分析; 财经短视频; 量化投资

收稿日期: 2023-06-26
基金项目: 国家自然科学基金 (12171282, 11801316); 全国统计科学研究重点项目 (2021LZ09)
Supported by National Natural Science Foundation of China (12171282, 11801316); National Statistical Scientific Research Key Project (2021LZ09)
作者简介: 通信作者: 何勇, 博士, 教授, 博士生导师, 山东大学齐鲁青年学者, 研究方向: 金融统计、计量经济学, E-mail: heyong@sdu.edu.cn; 李琪琪, 硕士, 研究方向: 金融统计, E-mail: 202122926@mail.sdu.edu.cn; 焦丽, 硕士, 研究方向: 金融统计, E-mail: 202211956@mail.sdu.edu.cn; 黄文萱, 硕士, 研究方向: 大数据分析和机器学习, E-mail: 202231998@mail.sdu.edu.cn.

Is Alternative Data Useful in China Share Market Investment? — Empirical Study Based on Financial Short Video, Image and Text Data

HE Yong[1], LI Qiqi[2], JIAO Li[2], HUANG Wenxuan[1]

(1. Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan 250100, China;
2. School of Mathematics, Shandong University, Jinan 250100, China)

Abstract Currently, the application of alternative data provides a new perspective for scholars and practitioners in the field of financial investment. This paper builds an analysis platform based on the FarmPredict (factor-augmented regularized model for prediction) framework and deep neural network model, realizing the task of learning trading signals from alternative data such as financial short videos and financial news thereby constructing trading strategies for the China share market. Firstly, match the captured financial news with their corresponding stock code and decompose it into text data and image data. Secondly, the text data is input into the FarmPredict learning framework. We construct and screen the text bag of words by which the phrases are decomposed into common factors and specific factors, and then calculate the score of the news text by the factor regression; We then input the image data into the image recognition deep neural network Google Inception v3 model framework built by the transfer learning technique, thereby outputting the probability that the image represents positive/negative emotions and the image sentiment index and image score. For the captured financial short video, it contains two steps. The first step is to strip the audio data and convert it to audio text data, and use the trained FarmPredict framework to calculate the text score of the short videos; the second step is to extract the key frames of the video, and use the trained image model to calculate the video image score; the text score is summed up with the image score to get the short video data score. Finally, the financial short video score, the text score and the image score of the news report are summed to obtain the stock investment signal, which is used as the basis for constructing the China share stock portfolio and formulating an appropriate investment strategy. Finally, the financial short video score, the text score and the image score of the news report are summed to obtain the stock investment signal, which is used as the basis for constructing the China share stock portfolio and formulating an appropriate investment strategy. The research results show that financial videos and financial news data contain information related to stock prices, which can effectively predict market changes and bring excess returns to investors. The empirical study confirms the importance of alternative data in the Chinese market. By comprehensively analyzing alternative data, this paper provides investors with a comprehensive and effective trading signal extraction method, which can help optimize investment strategies and achieve higher real returns.

Keywords China share market; financial news analysis; financial short video; quantitative investment

1 引言

随着信息技术的革新以及互联网普及率的提升, 另类数据大量产生. 另类数据, 即非传统数据, 泛指区别于传统金融数据的有价值的信息和数据 (廖理 (2021)). 这类数据具有种类繁多、数据量大、实时性高、颗粒度细等特点, 是金融领域的重要研究方向之一. 如今, 个人投资者纷纷选择在互联网上浏览财经新闻、上市公司年报等信息并在网络公众平台发表自己对中国股市的看法. 这些非结构性数据均属于另类数据范畴, 在预测股市走向、舆情分析和投资者意见摩擦等方面有一定使用价值. 越来越多的非结构性数据, 被运用到金融学研究中 (周颖刚等 (2022)). 现有的投资策略构建大多仅考虑了股票收盘价等传统金融数据, 而较少考虑具有非结构化特征的另类数据. 因此将另类数据作为数据来源引入投资策略的构建中, 探究其在金融市场投资的绩效表现具有重要的意义.

目前另类数据的应用处于初步发展阶段. Schumaker and Chen (2009) 发现上市公司季度报告或突发的新闻报道会极大地影响证券价格, 并将其用于股价预测, 实验结果表明文本数据可有效应用于金融市场. Froot et al. (2017) 从多个来源的另类数据中挖掘出了与零售企业销售相关的实用指标, 这些指标在解释企业季度销售、收入和盈利方面非常有利. 林建浩等 (2022) 基于新闻文本情绪构造机器学习投资策略并取得了较为可观的累计收益. 然而, 现有的另类数据研究大多围绕文本数据展开讨论, 较少考虑图片、音频、视频等其他形式的数据. 依托于自媒体的迅速发展, 国内金融市场产生了海量的视频、音频等非结构化数据, 而机器学习算法具备处理非结构化数据的天然优势, 能够提取深层次的潜在特征. 因此, 本文将自适应性强的机器学习算法用于金融市场中的另类数据分析, 旨在探究能否构建出具有超额收益的投资策略, 研究另类数据在投资中的价值.

本文研究的另类数据包含三种数据类型: 文本数据、图像数据和视频数据, 其中文本数据和图像数据由财经新闻分解得到, 视频数据为财经类短视频. 对于文本数据, 本文使用 Fan et al. (2021) 提出的 FarmPredict (factor-augmented regularized model for prediction) 框架学习交易信号. FarmPredict 框架是一种基于因子模型和稀疏正则化的文本数据学习框架, 可以将文本数据分解为公共因子和特异因子, 通过求解因子回归方程, 计算文本数据得分. 对于图像数据, 本文通过迁移学习引入预训练的 Google Inception v3 模型, 修改网络输出层为二分类层, 输出图片的积极情绪概率和消极情绪概率. 随后, 应用最小二乘法求解以图片情绪指数为自变量, 股票次日收益率为因变量组成的回归方程, 计算图像数据得分. 对于视频数据, 第一步剥离音频数据转换为文本数据, 利用训练好的 FarmPredict 框架计算视频数据文本得分; 第二步提取视频的关键帧, 利用训练好的图像模型计算视频图像得分; 文本得分与图像得分求和得到短视频数据得分. 至此, 本文完成了从文本、图像和视频多种数据中挖掘交易信号的任务. 接下来, 根据文本、图像和视频的得分计算股票的投资信号并构建投资策略, 通过实际数据检验策略在中国股票市场的表现.

本文以 2021 年 1 月至 2023 年 3 月国内权威财经网站的财经新闻和短视频平台头部财经博主的视频作为数据源. 经过数据清洗后的新闻总量为 78,668 篇, 视频总数为 29,830 个, 累计时长超过 5,960 分钟. 图片数据选用了 2018 年 1 月至 2023 年 3 月间国内权威财经网站的新闻配图, 共计 1,254 张. 需要强调的是, 在实证分析中, 文本、图像、视频数据的时间

区间是一致的, 均为 2022 年 1 月至 2023 年 3 月. 为了保证模型的有效性, 本文设置滚动窗口定期更新参数. 回测结果显示, 测试期内累计对数收益率随时间稳步增长, 日波动率平缓. 从判断量化投资策略优劣的指标数据来看, 本文构建的投资策略的年化收益率优于同期中证 500 指数收益, 表明另类数据具有明确的投资价值; 夏普比率大于 1 (在合理区间内), 说明投资策略有良好的稳健性; 最大回撤小于 8%, 证明投资策略是低风险的. 此外, 本文发现将图片数据和视频数据纳入信息源后, 投资策略的年化收益率与夏普比率分别提升了 4.52% 和 13.48%, 最大回撤下降了 0.45%. 这意味着模型具有更准确的交易信号预测能力, 由此制定的交易策略具有更好的稳健性和更低的风险性.

本文的贡献主要包括三个方面. 在数据方面, 数据来源更具兼容性. 模型数据输入以另类数据为主, 传统经济数据为辅, 提高了模型的普适性. 应用尚未被投资市场广泛使用的另类数据挖掘市场信息, 为个人投资者制定交易策略提供了新思路. 在模型方面, 本文提出的 FarmPredict 模型与深度神经网络模型相结合的组合具有一定的创新性, 该组合模型可以通过处理另类数据为投资者和市场分析师提供更加有效的风险管理工具. 这种组合模型应用于另类数据的处理, 更好地了解其在不同场景下的有效性和适用性, 从而为后续研究提供有益的启示. 在应用价值方面, 通过实证分析验证了财经新闻数据 (包括文本、图像和视频信息) 在股票市场中的预测价值. 基于我们的研究结果, 投资者和市场分析师可以利用财经新闻数据构建更加有效的交易策略. 监管机构也可以通过关注财经新闻中的信息, 更好地监测市场动态和潜在风险.

后文安排如下: 第二部分是文献综述部分, 整理了另类数据与文本分析、媒体情绪以及机器学习相关的前沿研究成果; 第三部分介绍基于 FarmPredict 文本数据处理框架和深度神经网络模型的另类数据投资信号提取方法; 第四部分为实验设计与实证分析, 阐述本文的数据提取、模型训练及构建交易策略的方法, 报告实验结果, 分析交易策略的综合表现; 第五部分是模型鲁棒性测试, 分析了模型对训练参数的敏感性; 最后是本文的研究结论与未来展望.

2 文献综述

本文主要与三类文献紧密相关, 分别是另类数据与实证金融研究、媒体情绪与资产定价研究、机器学习与投资策略研究. 本文关注以上三类文献的交叉融合, 利用机器学习算法提取媒体情绪并将其应用在投资组合策略中, 对不同领域的研究做出边际贡献.

2.1 另类数据与实证金融研究

另类数据作为投资研究中使用的新型数据, 在实证金融方面发挥着越来越重要的作用. 另类数据的信息价值吸引了资本市场中的投资者, 特别是希望得到超额回报的买方参与者, 他们希望通过另类数据中的有效信息来提高自身的信息解读能力和信息处理效率, 从而获取超额收益 (廖理等 (2021)). 通过研究另类数据的信息含量, 学者们认为另类数据包含了基本面信息, 并且可以预测股票的未来收益 (Zhu (2019)).

另类数据与传统金融数据不同, 即它不通过常规渠道获得, 但是在实际应用中与传统结构化数据类似, 同样在监测金融市场情绪、舆情分析和投资者意见摩擦等方面有一定使用价值. Hirshleifer and Teoh (2003) 认为投资者获取信息和解读信息能力有差异, 也会影响对企

业盈余预测的结果. 因此, 投资者希望通过不同的信息渠道寻求新的信息, 并积极发展新技术来提高获取信息的效率 (Jame et al. (2016)). 而 Chordia et al. (2014) 认为在当今市场透明度较高的情况下, 从公开数据中获取超额收益将变得越来越困难, 因此金融市场的投资者更加注重另类数据的价值. 虽然我国学术界对另类数据的研究起步较晚, 但发展迅速 (王正位等 (2022)), 依托于我国发达的移动互联网和网络平台, 投资者可以分析网络平台上存积的大量用户数据, 挖掘其中有用的价值信息, 拓展金融市场的信息渠道. 然而, 目前关于另类数据的学术研究大多针对国外健全的资本市场 (Jagtiani and Lemieux (2018)), 而用户多、信息量大的国内资本市场鲜有研究. 因此, 将另类数据应用于中国股票市场的研究具有重要学术价值和实践意义.

在实证金融研究中, Bollen et al. (2011) 发现从社交媒体信息中得出的集体情绪状态与道琼斯工业平均指数在时间上存在相关性. 以此为起点, 另类数据开始在学术界和对冲基金行业中受到广泛关注. 在学术界, Chen et al. (2014) 通过社交媒体的传播, 研究了投资者意见, 并预测了未来的股票收益率; Huang (2018) 利用亚马逊网站上顾客评论的数据集, 验证了投资者可以获取产品质量和价值信息, 同时证明了公司产生现金流的能力在很大程度上取决于它为客户创造的价值; Tang (2018) 收集了个人 Twitter 观点, 发现通过大量的聚合信息可以显著预测公司季度营业收入和超额收益, 并且在控制了来自传统媒体的信息和观点后 (Bartov (2018)), 结论依旧成立. 在对冲基金行业, 一些非常成熟的量化基金在很长一段时间内都在使用另类数据, 这些数据在业界被使用的时间远远早于 “另类数据” 一词的流行.

在基于图像或视频数据的实证金融研究文献中, Jiang et al. (2020) 利用机器学习分析图像数据-股价价格图表, 得出最能预测回报的价格模式; Gomez-Cram and Grotteria (2022) 利用视频数据处理技术, 为新闻发布会视频中发音的词加上时间戳, 并将这些词与高频金融数据对齐, 实证证明了视频数据对资产定价与货币经济的影响; Hu and Ma (2020) 量化了视觉, 声音和语言维度的说服力, 并表明了量化得到的说服力通过引导投资者形成不准确信念而影响结果.

2.2 媒体情绪与资产定价研究

在行为经济学中, 情绪是影响个人行为和决策的重要因素之一 (姜树广等 (2013)). Bollen et al. (2011) 发现这种规律也适用于金融市场, 即公众情绪与经济指标相关, 甚至可以预测经济指标. 新闻是引导公众情绪的重要媒介, 财经新闻报道作为金融市场重要的信息来源, 也是常见的另类数据, 广泛应用于预测股价变动、衡量投资者情绪和控制投资风险等领域.

前沿研究已经意识到媒体情绪会对金融资产价格产生较大影响. Mullainathan and Shleifer (2005) 发现媒体在传播报道内容的基础上, 也会向公众释放其情绪和观点, 从而影响大众对事件的看法与态度. 从这一角度来说, 媒体并不是一个客观中立的信息传播者, 而是市场情绪和大众舆论的引导者. 媒体可以通过新闻内容影响公司的决策与投资行为 (Dyck et al. (2007)). 很多学者认为资本市场对新闻报道的反应是非常迅速有效的, Huberman and Regev (2001) 发现即使是早已被 Nature 杂志和多家知名媒体报道过的 “治癌新药” 文章, 当其再次刊登在纽约时报时, 依旧引起了相应上市公司的股票大涨. Jeffrey and Green (2002) 通过研究美国 CNBC 电视台, 发现个股意见的正面报道会在 1 分钟之内完全反应在价格上.

而 Rinallo and Basuroy (2009) 注意到新闻媒体存在着特有情绪和报道偏差, 并对新闻报道的客观性与有效性提出质疑. 对财经新闻而言, Tetlock et al. (2008) 则认为媒体情绪在公共空间中的扩散会广泛影响市场投资者对未来股票收益和公司业绩的判断与预测, 进一步影响金融市场中的资产定价问题. Jacob et al. (2019) 认为媒体情绪通过引导舆情, 影响市场注意力与投资者的交易行为, 同样会导致资本市场中的价格波动.

在实际应用中, 由于使用新工具获取信息的方式得到普及, 许多工作被商品化, 社交媒体是投资者们分析金融市场的有力工具, 投资者在选择金融产品时, 不再完全依赖于专家建议, 而是转向听取同行意见 (Chen and Xie (2008)). 而社交媒体中的媒体情绪也是影响市场预测的重要因素, 例如张宗新等 (2021) 发现通过引入媒体文本情绪, 可以提高对股票收益的样本外预测能力. 游家兴等 (2012) 研究认为媒体情绪易在公共环境中广泛迅速传播, 且在传播过程中出现过度重复和强调的现象, 使得媒体情绪在投资者之间交叉传递. Jiao et al. (2020) 发现社交媒体对股票波动率和换手率有很好的预测能力, 这一发现有效验证了回声室效应的存在, 即投资者会把重复的信息当作真实信息. 从横截面角度来说, Fang and Peress (2009) 则认为更高的投资回报率发生在那些很少被媒体关注或报道的公司中. 还有学者将目光转向了悲观媒体情绪, Chan (2003) 研究发现负面新闻可对股票收益造成长达 12 个月的负面影响. Tetlock (2007) 就媒体悲观情绪对收益的影响进行了逻辑分析, 发现如果媒体展示纯粹的悲观情绪, 那么资产会出现短期低收益和长期收益反转现象. 研究说明, 媒体关注会影响市场中的交易量与资产定价. 因此, 针对媒体情绪的研究在金融市场中发挥着越来越重要的作用.

2.3 机器学习与投资策略研究

在计算机技术飞速发展的时代背景下, 基于大量数据的机器学习算法应运而生, 它能够训练、分析和处理看似无关的另类数据, 得到有用的信息或结论. 例如, 机器学习在股票预测与投资策略制定中就展现出了优于传统计量模型的盈利效果 (Gu et al. (2020)).

高效的投资策略离不开合理的股票投入, 动态、复杂、非线性是股票市场的鲜明特征, 如何选出优质股票是一项具有挑战性的工作 (林耀虎等 (2022)). 关于股票未来收益的预测, 现有的文献研究主要集中在两个方面: 一是以线性自回归模型为代表的计量经济学模型, 这些模型为保证结果的准确性通常需要做一些条件严格的假设. 二是以支持向量机、神经网络等为代表的机器学习模型, 这些模型不需要严格假设, 且处理非线性问题的效率更高. 已有学者将机器学习算法应用至股票收益预测中, 如 Krauss (2021) 在预测标准普尔 500 指数的涨跌情况时使用了决策树、深度神经网络等算法, 提高了预测效率; Lin et al. (2006) 将 Elman 神经网络应用在股票价格的预测问题上, 并与向量自回归模型做对比, 发现使用机器学习算法的动态投资组合模型表现更佳; Freitas et al. (2009) 基于神经网络算法改进传统的自回归移动模型, 使模型达到更好的预测水平. 借助人工智能的相关算法, 金融从业者可以利用大量的财经新闻去对股市波动做出高效分析, 通过选定的财经指标对股票指数的涨跌进行预测 (林建浩等 (2022)), 揭示财经新闻与股票市场之间存在的某种规律, 帮助投资者更好的决策. Zhang et al. (2016) 用向量表示文本数据, 利用卷积神经网络对输入数据进行语义学习, 预测股价变动, 改进个股的预测效果. 预测信息在投资问题中至关重要, 姜富伟等 (2021) 改进了传统 CAPM 模型在处理高维数据时容易出现过拟合和维数陷阱的不足, 利用前沿的机器

学习算法对高维数据进行降维和建模. Campbell and Thompson (2008) 认为利用预测的信息, 主动型投资者可获得比买入持有型投资者更高的夏普比率. Gu et al. (2020) 在使用深度学习模型预测股票市场中发现, 相比传统模型, 引入机器学习算法的模型盈利能力更强, 制定的投资策略更有效. 在规避风险能力上, Chinco et al. (2019) 意识到风险溢价中高维预测因子的函数形式具有不确定性, 传统的模型很难预测具有动态特性的风险溢价问题, 而机器学习能够很好地近似复杂的非线性关系, 避免过拟合偏差. 研究表明, 机器学习在解决投资策略问题时具有更好的处理方案和更高的效率, 可以提高模型盈利能力, 降低投资风险 (Zhang et al. (2016), Gu et al. (2020)).

近年来, 随着机器学习在金融市场的大量应用, 直接用机器学习选择股票构建投资组合, 成为了投资者们日渐关注的问题. Mayew et al. (2020) 认为如何使用机器学习算法改进传统的计量经济学模型是未来研究的重要方向. 得益于机器学习算法的数据驱动特征, 从数据中挖掘未知信息, 可以拓宽经济学研究范围 (肖争艳等 (2022)). 因此, 机器学习能够利用金融数据中包含的有关资产价格变动的信息, 对未来资产的价格做出准确预测, 进一步制定可盈利的资产配置方案.

3 理论模型

本文理论模型的流程图展示在图 1. 本文的核心是基于财经新闻文本数据、图像数据和短视频数据构建交易策略, 并检验该策略在中国股票市场的综合表现. 本节将阐述如何对文本数据、图像数据和短视频数据计算得分, 得分即数据中隐藏的交易信号.

对于文本数据, 本文使用的基础模型是 Fan et al. (2021) 提出的 FarmPredict 框架. FarmPredict 框架分为三个部分实现: 第一部分是利用无监督方法从词频矩阵中学习潜在因子, 通过 SVD 分解 (singular value decomposition) 将文章转换为由多个公共因子和特异因子组成向量, 并通过调整特征值阈值法确定公共因子的数量. 第二部分, 根据因子与学习目标的相关性筛选特异因子, 该步骤可以降低因子维度, 提高运算效率. 第三部分, 结合公共因子和筛选出的特异因子预测目标变量. 完成上述流程后, 模型可以自主计算新闻文本数据得分.

对于图像数据, 第一步是通过迁移学习的方式微调预训练的 Google Inception v3 模型, 将原始框架的输出层改为两个节点: 图像代表积极情绪的概率和图像代表消极情绪的概率. 第二步根据图像积极情绪概率构建图像情绪指数. 第三步, 应用最小二乘法求解回归方程, 其中图像情绪指数为自变量, 股票次日收益率为因变量. 完成上述流程后, 模型可以自主计算新闻图像数据得分.

对于短视频数据, 分两部分实现交易信号的提取, 第一部分是从视频中剥离音频, 利用讯飞 AI 语音转写接口将音频转写成文本, 将得到的文本数据输入 FarmPredict 框架计算视频文本得分. 第二部分是从视频中提取关键帧, 利用图像模型计算关键帧图像得分, 将关键帧图像的平均得分作为视频数据的图像得分. 文本得分及图像得分之和为视频数据的得分. 完成上述流程后, 模型可以自主计算财经类短视频得分.

接下来, 根据新闻标题及内容与股票代码的匹配结果, 获得股票与投资信号对应关系. 投资信号为文本数据、图像数据和短视频数据得分之和. 于是, 可以依据每支股票每日的投资信号, 构建合适的投资策略. 3.1∼3.3 节详细说明了上述算法的实现步骤.

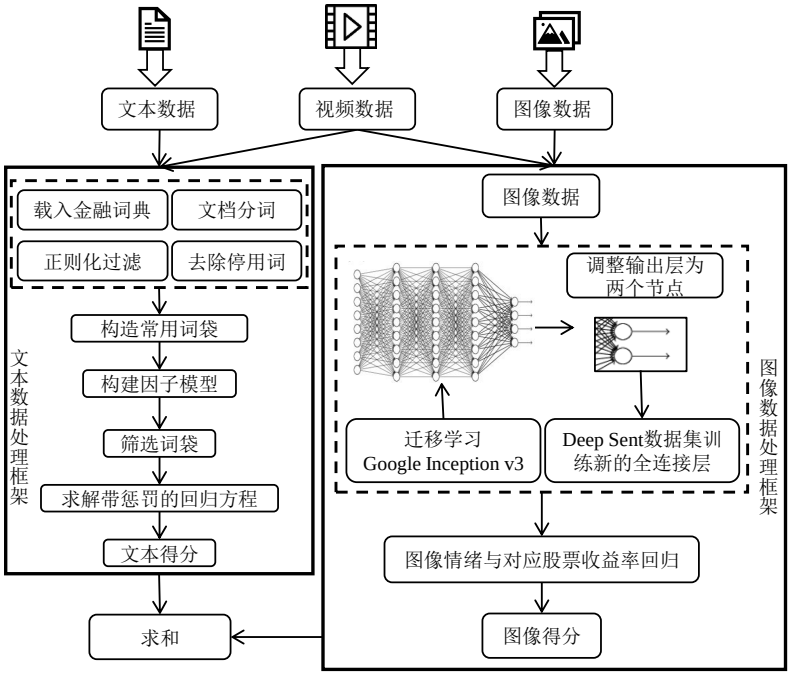


图 1　混合数据的交易信号提取流程图

3.1 基于 FarmPredict 框架的文本交易信号提取模型

3.1.1 建立因子模型并学习潜在因子

设 $\boldsymbol{D}$ 为从收集数据中提取到的全部词汇的集合. 在数据收集中, $\boldsymbol{D}$ 含有超过 224 万个不同的词汇或短语, 但是其中大部分词汇只在单一文章中出现, 机器很难学习到它们的语义. 因此, 需要设计算法清洗掉这部分不常见的词汇. 设 k_j 为含有词汇 j 的文章数, κ 为阈值, 当 $k_j \geq \kappa$ 时 j 词属于常用词袋集合 $\boldsymbol{D}^{\text{freq}}$, 即:

$$\boldsymbol{D}^{\text{freq}} = \{j\text{-th word in } \boldsymbol{D} : k_j \geq \kappa\}, \tag{1}$$

式中阈值 κ 为超参数, 通过调节 κ 的取值以保证集合 $\boldsymbol{D}^{\text{freq}}$ 中词汇的全面性. 具体来说, 阈值 κ 是一个经验参数, 用于筛选出对股票市场具有较大影响力的文本信息, 阈值 κ 的确定方法参考了 Fan et al. (2021) 将阈值 κ 设置为 $\boldsymbol{D}^{\text{freq}}$ 词集的单词数约 10,000 左右, 以在 $\boldsymbol{D}^{\text{freq}}$ 的全面性和不常用词引入的噪声之间取得平衡.

设 $\boldsymbol{X_i}$ 为词频向量, 其中 $\boldsymbol{X_{i,j}}$ 是第 i 篇文章中单词 j 的词频, 单词 j 属于常用词袋 $\boldsymbol{D}^{\text{freq}}$. 词频向量 $\boldsymbol{X_i}$ 可以表示为多个公共因子和特异因子组成的向量, 表达式如下:

$$\boldsymbol{X_i} = \boldsymbol{B f_i} + \boldsymbol{u_i}, \quad i = 1, 2, \cdots, n, \tag{2}$$

其中 $\boldsymbol{B}$ 表示因子载荷矩阵, $\boldsymbol{f_i} \in R^k$ 表示 k 维公共因子向量, $\boldsymbol{u_i} \in R^{|\boldsymbol{D}^{\text{freq}}|}$ 是与 $\boldsymbol{f_i}$ 无关的特异因子向量. 因子模型也可以改写成矩阵形式:

$$\boldsymbol{X} = \boldsymbol{BF} + \boldsymbol{U}, \tag{3}$$

式中 $\boldsymbol{X}$ 和 $\boldsymbol{U}$ 的维数都是 $n \times |\boldsymbol{D}^{\mathbf{freq}}|$, $\boldsymbol{F}$ 的维数是 $n \times k$, 由原来的 $|\boldsymbol{D}^{\mathbf{freq}}|$ 减少到 k. 式中 $\boldsymbol{X}$ 是可以获得的, $\boldsymbol{F}$、$\boldsymbol{B}$ 和 $\boldsymbol{U}$ 可以通过 SVD 分解得到.

3.1.2 构建情绪词汇集合并筛选特异因子

在已知因子载荷 $\boldsymbol{B}$ 的情况下, 通过因子与学习目标 $\boldsymbol{Y}$ 间的相关性进一步筛选特异因子, 构建情绪词汇集合 $\boldsymbol{S}$. 设 $\widehat{\boldsymbol{Y}}_{\boldsymbol{u}}$ 为 $\boldsymbol{Y}$ 与公共因子 $\boldsymbol{F}$ 线性拟合得到的残差向量. 设定阈值 α, 当 $\widehat{\boldsymbol{Y}}_{\boldsymbol{u}}$ 与单词 j 的相关性超过阈值 α 时, 认为词汇 j 在集合 $\boldsymbol{S}$ 中. 本文阈值 α 将被调整为选择大约 1000 个词汇. 算法实现过程如下:

$$\boldsymbol{S} = \{j : |\mathrm{corr}(\boldsymbol{U}_{\boldsymbol{j}}, \widehat{\boldsymbol{Y}}_{\boldsymbol{u}})| > \alpha\} \cap \{j : k_j \geq k\}. \tag{4}$$

3.1.3 训练预测模型

由 3.1.1 节知词频向量 $\boldsymbol{X}_{\boldsymbol{i}}$ 可以写成公共因子 $\boldsymbol{f}_{\boldsymbol{i}}$ 和特异因子 $\boldsymbol{u}_{\boldsymbol{i}}$ 的组合, 因此可以用因子代替 $\boldsymbol{X}_{\boldsymbol{i}}$ 进行预测. 设 t 日为新闻发布日, 以股票收益率 $Y_{i,t}$ 为因变量, $\boldsymbol{f}_{\boldsymbol{i}}$ 和 $\boldsymbol{u}_{\boldsymbol{i}}$ 作为自变量, 建立模型如下:

$$Y_{i,t} = a + \boldsymbol{b}^{\top}\boldsymbol{f}_{\boldsymbol{i}} + \boldsymbol{\beta}^{\top}\boldsymbol{u}_{\boldsymbol{i}} + \varepsilon_i, \tag{5}$$

式中的 ε_i 是模型引入的噪声项, $Y_{i,t}$ 是学习目标. $\mathrm{cp}_{i,t}$ 为新闻 i 对应股票在 t 日的收盘价, $Y_{i,t}$ 计算公式如下:

$$Y_{i,t} = \ln(\mathrm{cp}_{i,t}) - \ln(\mathrm{cp}_{i,t-1}). \tag{6}$$

使用带惩罚的最小二乘法估计的回归参数, 表达式如下:

$$(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\beta}}) = \underset{(\boldsymbol{a},\boldsymbol{b},\boldsymbol{\beta})}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_{i,t} - \boldsymbol{a} - \boldsymbol{b}^{\top}\boldsymbol{f}_{\boldsymbol{i}} - \boldsymbol{\beta}^{\top}\boldsymbol{u}_{\boldsymbol{i},\hat{\boldsymbol{S}}})^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{7}$$

其中, $\boldsymbol{u}_{\boldsymbol{i},\hat{\boldsymbol{s}}}$ 是情绪词汇集合 $\boldsymbol{S}$ 中 $\boldsymbol{u}_{\boldsymbol{i}}$ 的分量形式, 惩罚参数 λ 将通过交叉验证选择.

至此, 完成了对 FarmPredict 框架的训练. 对于给定的新文章中的向量 $\boldsymbol{X}_{\mathbf{new}}$, 利用第一步给定的因子载荷矩阵 $\widehat{\boldsymbol{B}}$, 得到公共因子 $\boldsymbol{f}_{\mathbf{new}}$ 以及特异因子 $\boldsymbol{u}_{\mathbf{new}}$, 公式如下:

$$\boldsymbol{f}_{\mathbf{new}} = (\hat{\boldsymbol{B}}^{\top}\hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^{\top}\boldsymbol{X}_{\mathbf{new}},\ \boldsymbol{u}_{\mathbf{new}} = \boldsymbol{X}_{\mathbf{new}} - \hat{\boldsymbol{B}}\hat{\boldsymbol{f}}_{\mathbf{new}}. \tag{8}$$

新文章预测股票收益率过程可以由 (9) 式刻画. 针对同一天有多个新闻对应同一支股票的情况, 采用等权平均的方法, 使用多个文本数据得分的等权平均作为该股票在该天的文本得分.

$$\hat{\boldsymbol{Y}}_{\mathbf{new}} = \hat{a} + \hat{\boldsymbol{b}}^{\top}f_{\mathbf{new}} + \hat{\boldsymbol{\beta}}^{\top}\boldsymbol{u}_{\mathbf{new},\hat{\boldsymbol{S}}}, \tag{9}$$

式中 $\hat{\boldsymbol{Y}}_{\mathbf{new}}$ 就是模型从文本中学习到的交易信号.

3.2 基于 Google Inception v3 模型的图片交易信号提取

本节将描述如何建立可以从图片中捕获交易信号的模型. 受 Obaid et al. (2022) 启发, 本文将 Google Inception v3 模型应用于解决图片交易信号的提取问题, 并应用于中国股票市场投资.

Google Inception v3 模型是一个卷积神经网络模型. 在 ImageNet academic 比赛中表现卓越, 并被广泛用于实践和研究. 预训练的 Google Inception v3 模型是在 1,331,167 张带标签的图片上训练的, 每个推理的计算成本为 50 亿乘法的网络, 使用接近 2500 万个参数. 由此可见, 训练全新的图像分类模型提取交易信号计算成本高昂.

为了节约计算成本, 本文使用 Google Inception v3 模型已存储的知识来解决图片交易信号的提取问题. 通过给 Google Inception v3 模型提供一个以情绪为标签的图片数据集, 将模型的输出层调整为两个节点, 分别代表图片的积极情绪概率和图片的消极情绪概率. 为了研究图片情绪与股票收益率之间的关系, 本文根据图片积极情绪概率构建了图片情绪指数 M_i. 具体而言, 先将财经新闻配图作为模型输入. 假设新闻 i 中含有 N_i 张配图, 则将其全部作为框架的输入数据, 计算每张图像的积极情绪概率 $M_{i,j}$, 其中 $1 \leq j \leq N_i$. 假定同一篇新闻中每张图片的曝光量都是一样的, 于是新闻 i 的情绪指数 M_i 可以由下式计算:

$$M_i = \frac{\sum_{j=1}^{N_i} M_{i,j}}{N_i}. \tag{10}$$

接下来, 将新闻匹配的学习目标 $Y_{i,t}$ 作为因变量, 新闻的图像情绪指数作为自变量, 构造回归方程, 最小二乘法求解表达式如下:

$$(\hat{c}, \hat{d}) = \arg\min_{(c,d)} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_{i,t} - c - dM_i)^2 \right\}. \tag{11}$$

使用新文章配图预测股票收益率的过程可以表示如下:

$$\hat{Y}_{\text{new}} = \hat{c} + \hat{d} M_{\text{new}}, \tag{12}$$

式中 M_{new} 表示新文章的图像情绪指数, $\widehat{Y}_{\text{new}}$ 表示模型从图像中学习到的交易信号.

3.3 视频交易信号提取

财经类短视频属于语言类视频, 包含音频信息和图像信息. 本节的主要工作就是利用财经类短视频的特点, 充分挖掘视频的音频数据和图像数据提取交易信号.

视频交易信号学习流程展示在图 2, 共包含两部分. 第一部分计算视频音频得分, 首先从视频数据中剥离音频数据, 使用讯飞 AI 语音转写接口将音频转换为文本数据, 根据文本数据及视频发布时间为其匹配对应的股票收益率, 保留匹配结果唯一的音频文本; 其次根据常用词袋 $\boldsymbol{D}^{\text{freq}}$ 将音频文本转换为词频矩阵 $\boldsymbol{X}$; 最后将词频矩阵输入 FarmPredict 框架, 计算视频数据对应的文本得分. 第二部分计算视频图像得分, 首先利用视频采集数据接口, 对预处理后的视频进行关键帧提取; 然后将所得关键帧图像输入 Google Inception v3 模型, 针对有多个图像对应同一支视频的情况, 我们采用了等权平均的方法, 得到对应视频的图像情绪得分. 文本得分及图像得分之和为视频数据的交易信号.

4 实验设计与实证分析

4.1 数据说明

本文借助财经数据接口包 Tushare 提取了 “东方财富网” 的新闻快讯, 将其作为本文使用的文本数据; 对于图像数据和短视频数据, 本文基于 Scrapy 自建爬虫实现了对数据的抓取.

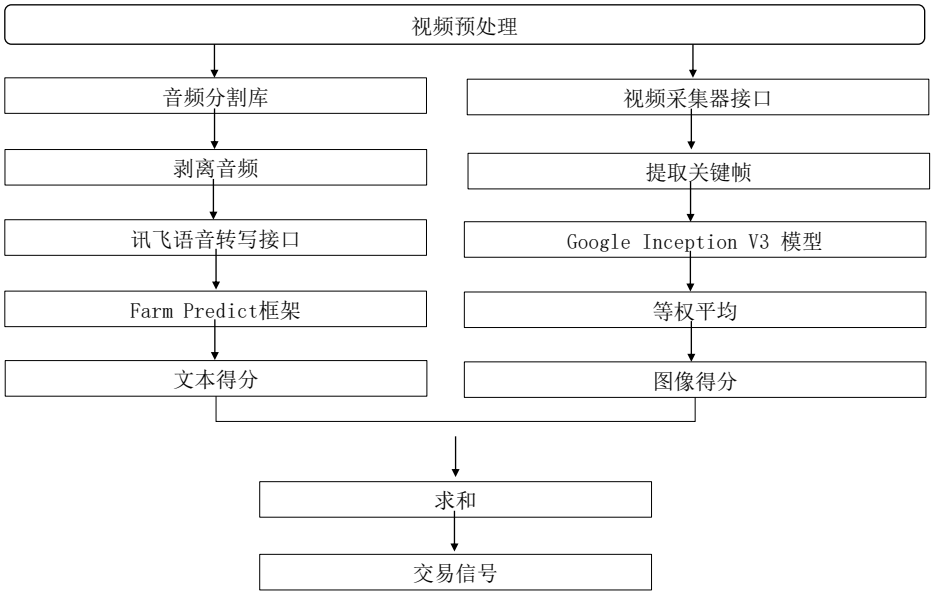


图 2 视频交易信号提取

4.1.1 文本数据

TuShare 是一个基于 Python 的开源财经数据接口包, 开发人员对股票等金融数据进行了收集、清洗加工和存储等工作, 能够为金融分析人员提供快速、整洁和多样的便于分析的数据. 本文中主要使用 Tushare 的 “新闻快讯” 接口, 通过设置接口参数返回相应的财经新闻, 返回值包含新闻内容、新闻标题和发布时间等字段. 接口提供的金融信息来自于国内权威的财经网站, 这些财经网站各有侧重点, 数据质量存在差异. 为了选取合适的金融网站, 本文选取了以下五个指标来综合评价网站, 分别如下:

1) Alexa 排名, 指三个月内网站的用户链接数以及页面浏览量的几何平均数, 每三个月更新一次, 是对全球网站的综合排名.

2) 百度权值, 代表网站的受欢迎程度, 主要通过网站的关键词排名来估算能给网站带来的流量, 划分等级为 0∼10.

3) PR 值, 全称为 PageRank, 是用来体现网页等级的通用标准, 划分等级为 0∼10.

4) 反向链接, 指从其他网站引入到该网站的链接数量.

5) 有效数据含量, 指清洗后的数据占原始数据的比例. 在 TuShare 接口下有四个主流新闻网站, 本文选择 2021 年 3 月 1 日至 2021 年 3 月 31 日的数据, 进行清洗, 计算有效数据含量.

表 1 展示了四个网站的相应指标. 可以看到, “东方财富网” 在 Aleax 排名和 PR 值上表现很好, 反向链接的数量虽然稍逊于 “新浪财经网”, 但比 “云财经” 要好得多. 关于有效数据比例, “云财经” 为 32%, 是四个平台中最好的一项, 但是它的 Aleax 排名, 反向链接数量较差. 综合考虑, 本文将选取 “东方财富网” 为金融信息获取网站.

接口提取的原始数据如表 2 所示. 虽然接口开发人员已经对数据进行了初步清洗, 但接

口返回值仍存在新闻内容为空、新闻字段不完整等问题, 所以再次对数据进行处理是很有必要的. 预处理过程分为三步: 第一步是根据新闻标题及内容匹配相应的股票代码, 本文研究的股票代码范围不包含金融行业股票及股票名称含有 ST (Special Treatment)、*ST、SST、S*ST 及 PT (Particular Transfer) 等关键字段的股票; 第二步删除匹配结果不唯一的新闻, 这是因为同时描述多支股票的新闻会对语义学习造成障碍; 第三步, 根据新闻发布时间和相应的股票代码为数据匹配学习目标. 处理后的数据示例展示在表 3. 最终, 文本数据源为 “东方财富网” 的财经新闻, 时间范围是 2021 年 1 月 1 日 00 : 00 到 2023 年 3 月 19 日 24 : 00, 收集的新闻总量为 367,851 篇, 清洗后为 78,668 篇.

表 1　权威财经网站综合排名

	Alexa 排名	百度权值	PR 值	反向链接	有效数据含量
东方财富	269	5	7	11,822	27%
新浪财经	37	8	7	51,880	16%
云财经	110,618	5	0	1,630	32%
同花顺	2,055	4	0	78,854	26%

表 2　未经预处理的金融新闻文本格式示例

发布日期	新闻内容	新闻标题
2022/1/3 20:29:02	【恒瑞医药 2 款创新药获批以创新实力迎新年 “开门红”】岁末年初, 恒瑞医药研发创新喜获丰收: 国家药品监督管理局于同一天批准公司 2 款创新药上市. 至此, 恒瑞医药已上市创新药达到 10 款. (经济观察网)	恒瑞医药 2 款创新药获批以创新实力迎新年 “开门红”
2022/1/3 20:23:27	【】	腾讯游戏: 相关产品已在华为游戏中心恢复上架
2022/1/3 20:14:16	【欧佩克下调一季度全球石油供应过剩预测正值考虑下一次增产】就在欧佩克 + 讨论是否再次增产的前一天, 欧佩克下调对本季度全球石油市场过剩供应的预测. 欧佩克 + 代表称预计周二会议将推进温和增产, 最新预测或许会鼓励他们做这个决定.	欧佩克下调一季度全球石油供应过剩预测正值考虑下一次增产

4.1.2 图像数据

在图像数据获取中, 同文本分析类似, 研究所采用的图片金融信息主要来自国内权威的财经网站: 新浪财经、腾讯财经、网易财经、财联社, 表 4 展示了不同平台上提供的图片数据的时间区间, 由于金融财经类的图片数据量有限, 本文中并未区分图片是来自于哪个平台网站, 使用 Python 语言开发的爬虫框架 Scrapy 抓取数据.

图片数据预处理过程与文本数据类似, 共分为三步: 第一步, 根据图片所在财经新闻的标题及内容匹配相应的股票代码; 第二步删除匹配结果不唯一的新闻配图; 第三步, 根据新闻发

布时间和相应的股票代码为数据匹配学习目标. 处理后的数据示例展示在表 3. 最终, 用于交易信号提取的图片数据源为 2018 年 1 月至 2023 年 3 月间财经新闻配图, 共计 1,254 张.

表 3 预处理后的金融新闻文本格式示例

发布日期	新闻内容及标题	收益率
2022/10/3	隆平高科: 水稻、玉米及棉花新品种通过国家审定【隆平高科: 水稻、玉米及棉花新品种通过国家审定】隆平高科 (000998) 1 月 3 日晚间公告, 根据《中华人民共和国农业农村部公告第 500 号》, 第四届国家农作物品种审定委员会第八次会议审定通过了 677 个稻品种、919 个玉米品种、39 个棉花品种及 86 个大豆品种, 其中含公司及下属公司自主培育或与他方共同培育的 88 个水稻新品种、50 个玉米新品种和 1 个棉花新品种. (证券时报)	3.1814
2022/10/3	四川九洲: 子公司拟挂牌转让捷能科技 5% 股权【四川九洲: 子公司拟挂牌转让捷能科技 5% 股权】四川九洲公告, 公司控股子公司九州科技拟通过产权交易所以公开挂牌的方式转让所持有的捷能科技 5% 股权. 该次交易以评估报告为依据, 挂牌底价为 17.198 万元, 挂牌交易完成后, 公司将不再持有捷能科技的股权. (财联社)	−0.9385
2022/10/3	双乐股份: 入选江苏省专精特新 “小巨人” 企业名单【双乐股份: 入选江苏省专精特新 “小巨人” 企业名单】双乐股份 (301036)1 月 3 日晚间公告, 公司入选江苏省专精特新 “小巨人” 企业名单. (证券时报)	3.1403

表 4 主流财经网站特点

网站	板块	是否含有配图	时间范围	发表时间精确度
新浪财经	经济新闻	有	2004.01 – 至今	年 – 月 – 日 – 时 – 分 – 秒
腾讯财经	金融市场	有	2022.09 – 至今	年 – 月 – 日 – 时 – 分 – 秒
网易财经	个股新闻	有	2022.07 – 至今	年 – 月 – 日 – 时 – 分 – 秒
财联社	财经新闻	有	近两周	年 – 月 – 日 – 时 – 分 – 秒

4.1.3 视频数据

本文选取抖音及快手作为短视频数据的来源, 主要基于以下考虑: 首先, 作为国内头部的短视频平台, 庞大的用户群体和均衡的用户分布保证了视频数据的影响力与影响范围. 据统计, 抖音平台日活跃用户超 7 亿, 人均单日使用时长超过 2 小时, 财经兴趣人群数量达 1.05 亿, 财经类作品发布量达 937 万, 视频播放量达 434 亿. 快手平台日活跃用户超 3 亿, 平均每日的使用时间为 111.5 分钟, 其中财经创作者共有 1.5 万, 财经类作品量达 166 万, 累积获得 294 亿曝光量. 庞大的受众群体保证了视频数据源的基础流量和影响力.

此外, 抖音平台以一二线用户群为主, 快手短视频平台以三四线城市的下沉用户为主, 两个平台的用户分布基本涵盖了国内各个城市, 确保了视频数据影响力的基础范围. 视频账号的选择依据了《快手 2020 财经科技创作者生态报告》和《2020 抖音财经内容生态报告》中的高质量优质头部财经博主账号榜单. 选取了这两个榜单交集的账号创作内容作为视频数据源, 从而确保所选账号具有较高的权威性和影响力. 选取的优质头部财经博主在两平台的粉丝量如表 5 所示.

视频数据预处理过程共分为四步: 第一步, 从视频中剥离音频, 将音频转换为文本; 第二步, 依据视频标题及音频文本数据匹配相应的股票代码; 第三步删除匹配结果不唯一的视频数据; 第四步, 根据视频发布时间和相应的股票代码为数据匹配学习目标. 视频数据通常以科普经济常识和介绍经济新闻为主, 对于那些无法直接对应到个股的视频, 通过预处理将其排除在研究范围之外, 以确保实证分析集中在与个股相关的视频数据上. 最终, 用于交易信号提取的视频数据源时间区间为 2022 年 1 月至 2023 年 3 月, 清洗后的视频数据测试集包含 29,830 个, 累计时长超过 5,960 分钟.

表 5　优质财经博主及粉丝量

数据来源	直男财经	韩秀云讲经济	珍大户	叶檀财经	暴躁财经
快手平台粉丝量/万人	205.6	456.8	165.5	207.5	263.4
抖音平台粉丝量/万人	1552.2	1148.5	783.2	552.0	303.0
粉丝量合计/万人	1757.8	1605.3	948.7	759.5	566.4

4.2 实验设计

4.2.1 文本交易信号提取

在将清洗后的文本数据输入 FarmPredict 框架前, 需要对文本进行分词处理, 将文本转换成词频矩阵. 文本分词是将整篇财经新闻变成由中文词语组成的集合. 本文选择 jieba 中文分词系统作为分词工具, 其优点是代码开源, 便于研究过程进行个性化扩展.

分词处理过程为: 载入金融词典、读入文本数据、文本过滤、去除停用词. 经处理, 总词袋包含 2,448,717 词, 词频 50 以上的常用词袋有 5,460 词, 由常用词袋构造的训练集维数超 18 万. 通过训练集估计得到因子载荷和带惩罚的回归方程参数, 然后可以获得对应的文本数据交易信号.

4.2.2 图片交易信号提取

在将图片数据输入模型进行计算前, 需要训练好微调后的 Google Inception v3 模型参数, 具体而言, 训练过程有下面三个步骤.

第一步, 选择数据集. 微调后的模型使用 DeepSent 数据集进行训练, DeepSent 是 You et al. (2015) 收集的数据集, 由 1,269 张带情感标签的图像组成. 本文保留 10% 的数据集作为验证集, 剩余图片作为训练集.

第二步, 载入预训练参数. 从 PyTorch 导入经过预训练的 Google Inception v3 模型, 预训练参数在 ImageNet 数据集学习得到.

第三步, 超参设置. 超参数是指那些在训练前或者训练中人为的进行调整的参数. 训练过程使用 6GB GPU 进行计算, 通过反复对比实验验证, 确定批次大小为 96. 本文使用 ResNet-18 (He (2016)) 和 ResNeXt-50(32x4d) 设计了一组对比实验, 选择准确率作为评估指标, 由此确定网络结构为 ResNet-18. 根据源代码库的分类实验, 可以确定当迭代次数为 200 左右时, 模型的训练误差和测试误差能够达到一个应用可以接受的范围. 因此, 迭代次数固定为 200.

最终, 模型训练参数设定如下: 批次大小为 96, 迭代次数为 200, 特征提取方法为

ResNeXt-50(32x4d), 学习率为 0.001, 动量为 0.9, 在配置为 3090GPU+i9 12900k CPU 服务器计算. 图 3 展示了迭代次数与模型在不同数据集上的准确率的关系, 经 200 次迭代后, 训练集的准确率稳定在 97%, 验证集的准确率稳定在 80%. 本文准确率与文献中其他照片情感分类模型所获得的值相似, 例如 Campos (2017) 比较了在 DeepSent 上训练的 CNN (convolutional neural networks) 模型, 最终验证集的准确率保持在 78.3% 和 83.0% 之间. 另外, 为便于理解模型的分类能力, 图 4 展示了模型的部分预测结果.

将清洗后的图片输入训练好的模型, 输出结果为图片情绪指数. 对于含有多个配图的新闻, 新闻图片得分为所有配图积极指数的平均值. 接着, 利用最小二乘法求解回归方程, 其中自变量是新闻图片得分, 因变量是股票收益率.

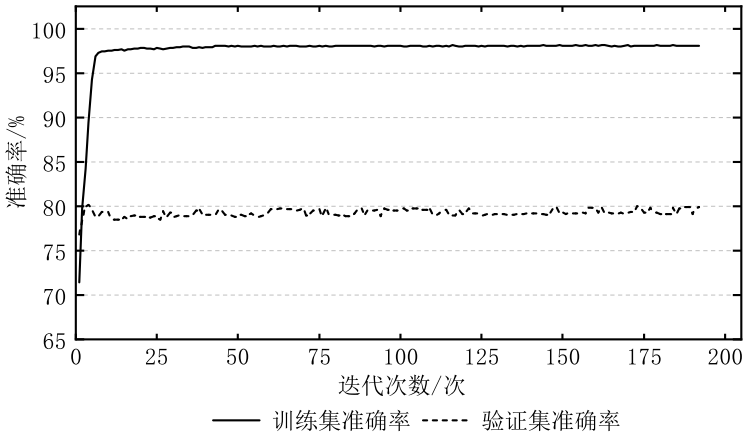
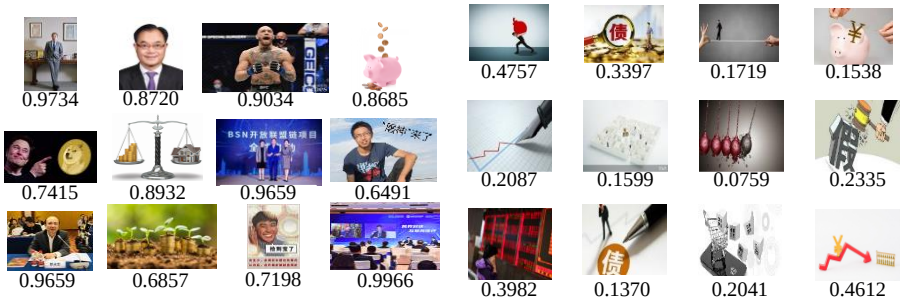


图 3　不同特征提取网络的准确率



(a) 积极图像示例　　(b) 消极图像示例

图 4　深度神经网络输出结果示例

4.2.3 视频交易信号提取

视频交易信号提取分为两部分, 分别完成对视频音频数据、视频图像数据的处理. 具体而言, 在处理视频音频数据部分, 完成视频音频到文本的转换后, 按照文本的预处理步骤处理、加工音频文本数据, 将预处理后的数据输入 FarmPredict 框架获得对应的视频文本得分. 在处理视频数据的图片信息时, 使用关键帧提取技术, 获取视频关键帧图像, 将图像输入微调

后的 Google Inception v3 模型得到视频图像得分. 对于多个关键帧对应同一个视频的情况, 将多个关键帧的图像得分等权平均的结果作为视频图像得分. 文本得分与图像得分求和得到短视频数据的交易信号.

4.2.4 回测窗口设置

模型在月度更新的滚动窗口上进行训练和测试, 图 5 展示了滚动窗口的更新频率与步长. 对于每个窗口, 12 个月的数据用于训练模型, 随后的 1 个月的数据用于测试模型. 测试中每篇文章的预测分数都被记录下来. 在一个窗口的训练和测试完成后, 我们将整个窗口向前滚动 1 个月, 重新进行训练和测试, 如此反复. 第一个窗口使用 2021 年 1 月至 2021 年 12 月的数据进行训练, 2022 年 1 月的数据进行测试. 最后一个窗口使用 2023 年 3 月的数据作为测试期. 总共检查了 15 个窗口, 最终记录了从 2022 年 1 月到 2023 年 3 月每个交易日的预测结果.

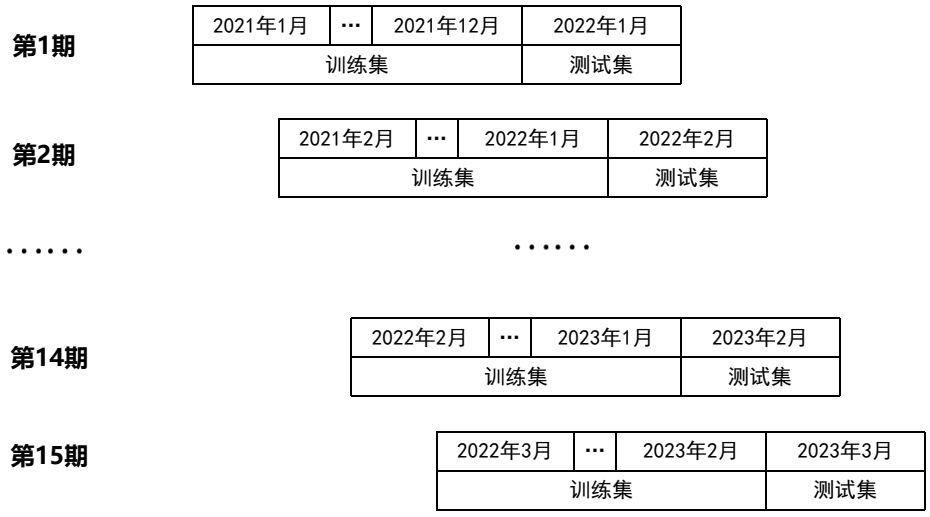


图 5　滚动窗口测试

4.3 实证结果分析

在得到文本数据、图像数据和视频数据的得分后, 根据其与股票匹配结果, 三者求和得到对应股票的投资信号. 下面, 通过预测的投资信号建立投资组合并验证模型的有效性. 投资组合是在每个滚动窗口建立和测试的, 具体如下.

步骤一: 设 t 日为交易日, t 日收市前收集中国股票市场相关的文本数据、图像数据及视频数据, 并根据数据类型计算得分, 三者求和得到对应股票的投资信号.

步骤二: 将股票按其得分从高到低依次排序, 等额投资交易信号为正且排名前 100 的股票.

步骤三: 在 $t+1$ 交易日收市前平仓. 并重复步骤一、步骤二.

本文选用年化收益率、夏普比率和最大回撤作为评估投资策略好坏的指标, 原因如下:

年化收益率: 指投资一年时的预期收益率, 是衡量策略投资收益的基本指标.

夏普比率: 指承担一单位风险带来的收益. 对于有效组合而言, 风险越大, 收益越大. 夏普比率越大.

最大回撤: 指可能发生的最大亏损幅度, 其值等于策略收益曲线上, 当前最高点到后期最低点的回撤幅度的最大值. 它描述了投资者可能面临的最大亏损, 数值越大说明风险越大.

本文计算了四种数据类型的交易信号, 并执行了对应的交易策略, 结果总结在表 6. 混合数据类型指导下的交易策略年化收益率为 23.21%, 夏普比率为 1.41, 最大回撤为 7.31%, 均优于单一数据源模型. 这表明本文所提出的模型具有更准确的交易信号预测能力, 由此制定的交易策略具有更好的稳健性和更低的风险性.

为了直观地理解视频与图像数据对文本数据的辅助效果, 在图 6 绘制了 “文本” 与 “文本 + 图像 + 视频” 两种策略的累积收益对比图. 经过与基准模型 (单一数据源模型) 对比, 混合数据类型指导下的交易策略有着更优秀的表现. 原因是相较于只能利用单一数据类型的框架, 优化后的模型具有文本、图片和视频三者协同识别和处理的能力, 这有效地优化了原始模型对数据的洞察力.

图 7 显示了使用文本、图像和视频三种数据类型的交易策略的累计对数收益率, 与之相比的是中证 500 指数. 虽然在 2 月份受 “俄乌战争” 影响导致股票市场大跌, 模型的交易策

表 6 投资策略判断指标

数据类型	夏普比率	年化收益率	日均基点	最大回撤
视频 + 图片 + 文本	1.41	23.21%	9 bps	7.31%
文本	1.22	18.69%	7 bps	7.76%
图片	−19.28	0.13%	<1	-
视频	−3.35	0.48%	<1	0.82%

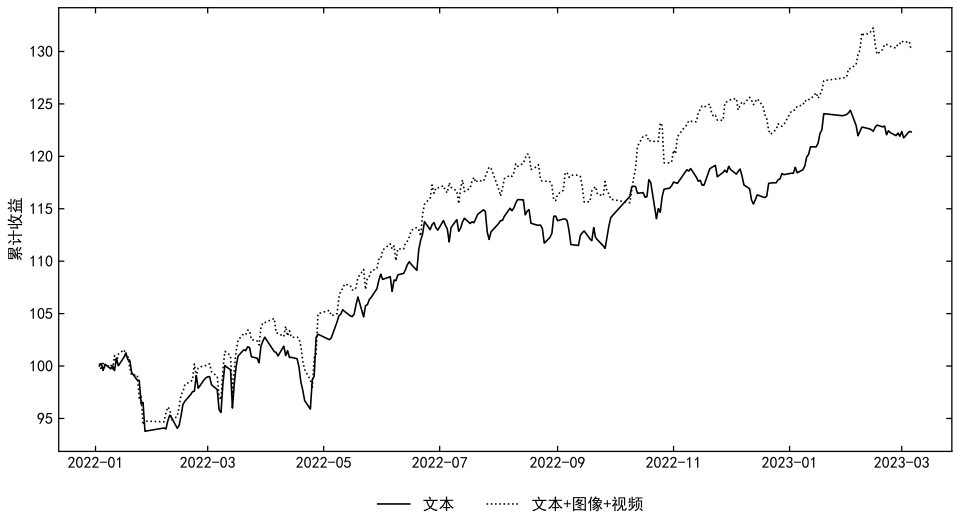


图 6 视频与图像数据的辅助效果

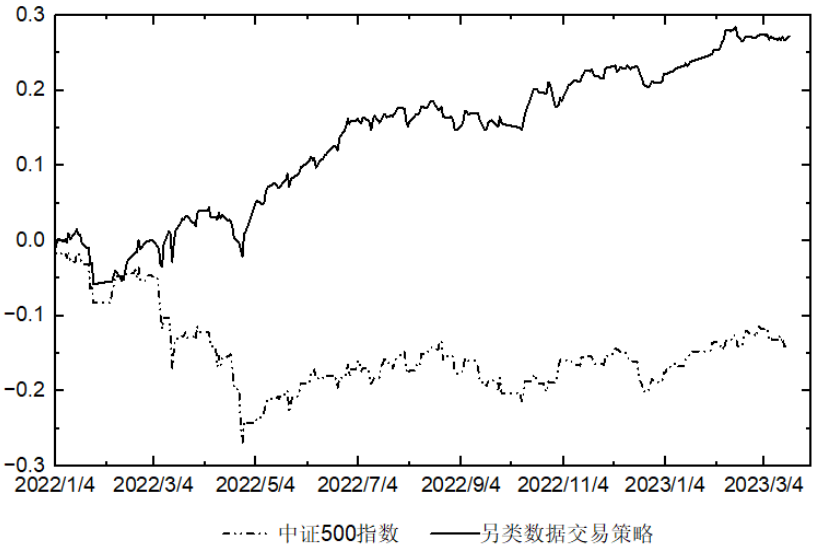


图 7　回测期内累计收益率

略出现负累计收益, 但是股票市场回暖后, 从另类数据中提取到的交易信号可以有效提高交易策略的收益率.

为了研究本文构建的投资组合是否受市场、 行业等各类因子的影响, 本文使用 Fama-French 五因子对各类数据构建的投资组合的超额收益进行回归, 对 alpha 做假设检验, 验证是否存在 “真正” 的超额收益. 具体公式如下:

$$\mathrm{Rirf}_{p,t} = \alpha_p + \beta_p^{\mathrm{RP}}\mathrm{RP}_t + \beta_p^{\mathrm{SMB}}\mathrm{SMB}_t + \beta_p^{\mathrm{HML}}\mathrm{HML}_t + \beta_p^{\mathrm{RMW}}\mathrm{RMW}_t + \beta_p^{\mathrm{CMA}}\mathrm{CMA}_t + \varepsilon_p. \tag{13}$$

其中 RP_t 是 t 时刻的市场风险溢价因子, SMB_t 是 t 时刻的市值因子, HML_t 是 t 时刻的账面市值比因子, RMW_t 是 t 时刻的盈利能力因子, CMA_t 是 t 时刻的投资模式因子. $\mathrm{Rirf}_{p,t}$ 是数据类型 p 在 t 时刻的日收益率.

通过计算得到混合数据的 α_p 的 p 值为 0.007, 在 5% 的置信水平下是显著的. 这说明在混合数据类型指导下构建的投资组合策略确实存在超额收益. 对于单一数据类型, 考虑到文本数据贡献了超 80% 的投资收益, 本文仅验证文本数据是否存在超额收益. 通过计算, 文本数据的 α_p 的 p 值为 0.043, 在 5% 的置信水平下也是显著的, 这表明基准模型的投资组合策略也是存在超额收益的.

5　模型鲁棒性测试

5.1　控制特异因子维数的影响

情绪词组决定了模型提取到的异质性特征, 情绪词组的维数是模型关键参数. 本节将情绪词组维数从 2000 更改至 500, 重新测试模型的投资策略, 研究控制特异因子维数对模型性能的影响, 测试结果归纳于表 7. 实验结果表明, 特异因子维数确实存在对模型预测性能的影

表 7 FarmPredict 框架测试

指标	夏普比率	年化收益率	日均基点	最大回撤
$\|S\|$=2000	1.326	7.560%	3 bps	1.119%
$\|S\|$=1000	1.220	18.690%	7 bps	7.760%
$\|S\|$=500	0.840	11.811%	5 bps	7.538%

注: FarmPredict 框架测试的数据源为纯文本数据类型.

响力, 为特异因子选择合适的维数是影响模型性能的关键因素之一. 这一结果也证实了特异因子确实含有有效的市场信息, 并且因子的信息含量与因子数量相关.

5.2 控制数据源的影响

DeepSent 训练集中的照片来自社交媒体, 可能与样本的专业新闻图像类型存在一定差异, 差异过大时会导致模型分类准确率降低. 因此, 本节使用人工标注的专业新闻照片数据集测试模型, 研究数据源对图片情绪识别的影响, 最终的结果归纳于表 8. 实验结果表明, 数据源对模型情绪识别准确率的影响较弱, 模型对图像数据源不敏感.

表 8 Google Inception v3 Model 框架测试

		真实值	
		积极情绪	消极情绪
预测值	积极情绪	48	8
	消极情绪	2	42

5.3 投资组合设定的影响

为了确定交易策略对模型的影响, 本节分别采用排名前 5, 10, 20, 50, 100 及 120 支股票制定策略并执行回测. 通过对比不同交易策略的优劣, 结果展示在图 8. 实验结果表明, 选择排名靠前的 100 支股票可以获得更大的超额收益.

本文研究的股票代码初始范围为 A 股市场所有股票, 为了防止存在退市风险的股票和金融行业的股票带来的额外影响, 剔除了金融行业股票及股票名称含有 ST、*ST、SST、S*ST 及 PT 等关键字段的股票.

图 9 展示了回测期间每日股票池中股票数量的变化. 剔除极大值和极小值后, 我们发现每日股票数量 (交易信号为正) 的最大值为 91, 这一数字是基于我们所选取的股票样本和投资策略计算得出的. 鉴于中国股票市场存在禁止卖空的规定, 在投资组合中不考虑做空情况下, 我们选取交易信号为正的股票构成交易策略的选择池. 它反映了在特定日子内, 根据本文的投资策略, 具有正交易信号的股票数量. 这并不意味着股票样本量不足, 而是表明在某些日子里, 根据策略, 只有部分股票具有正向交易信号. 因此, 在交易策略设定为 100 支股票时, 意味着我们完全接受了模型的预测结果, 所以选择排名靠前的 100 支股票可以获得更大的超额收益.

5.4 控制交易成本的影响

本文构建交易策略以日度为频率进行换仓, 持有期较短, 高换仓频率会带来高交易成本. 在以往文献中, 费率的设定主要参考所研究地区的费率水平, 一般设定固定的交易费率. 本文

设置交易费率为 0.15%. 表 9 表明, 扣除交易成本后策略年化收益率为 23.210%, 较扣除前降低 17.042%, 夏普比率由 2.532 降低至 1.410. 尽管交易费用消耗了一部分的策略收益, 但达 23% 的年化收益率远远高出同期中证 500 指数收益率, 本文策略仍然具有较高的投资价值.

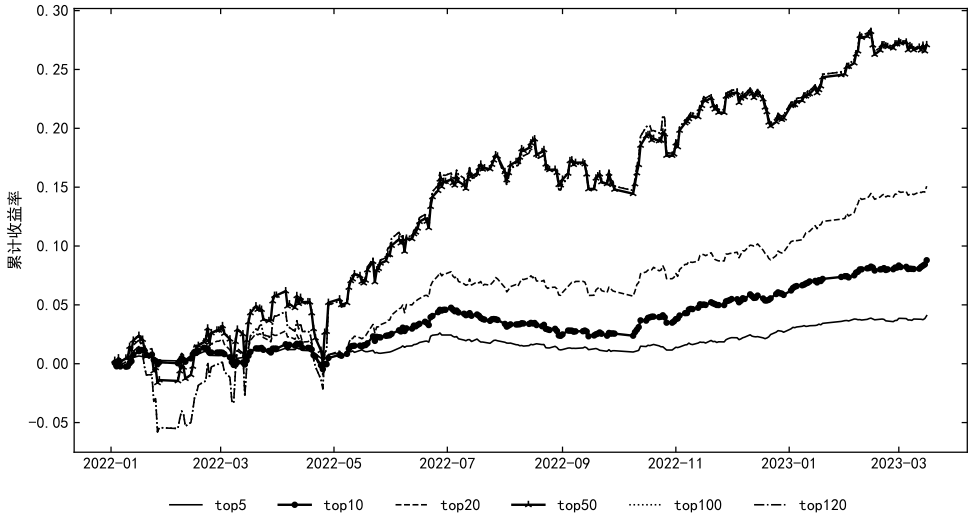


图 8　不同的投资策略结果

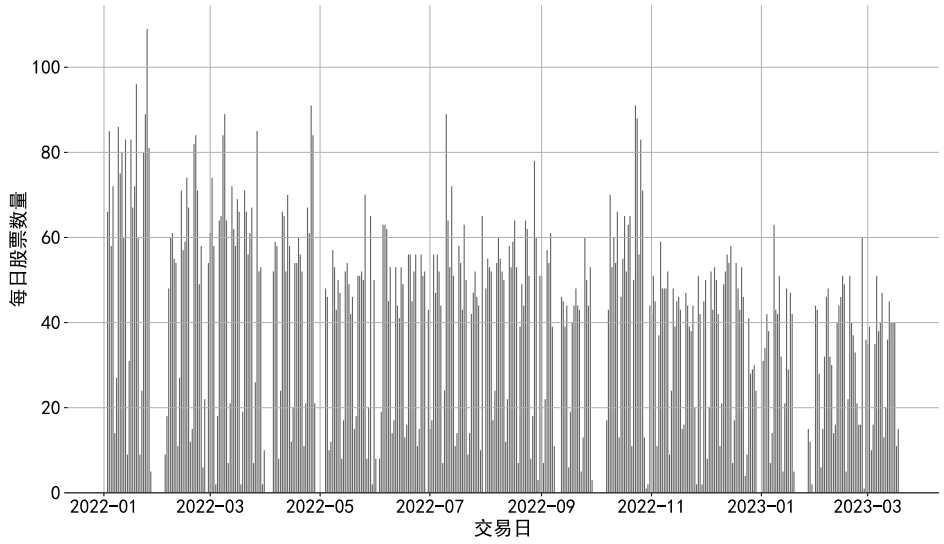


图 9　回测期间每日股票池中的股票数量

表 9　控制交易成本的影响

	夏普比率	年化收益率	日均基点	最大回撤
扣除交易成本前	2.532	40.252%	16 bps	6.235%
扣除交易成本后	1.410	23.210%	9 bps	7.310%

注: FarmPredict 框架测试的数据源为纯文本数据类型.

6 研究结论

本文实现了从财经新闻和财经类短视频中学习投资信号并构建投资组合, 通过实际市场数据考察了构建的投资组合在中国股票市场的表现, 综合实证结果得到以下几点结论: 第一, 本文发现了除财经新闻文本外, 财经新闻配图及财经类短视频也具有潜在的市场信息, 基于此构建的策略可以为投资者带来超额收益. 尽管日度交易策略的交易费用高昂, 但本文策略在扣除交易成本后依然可以获得远超中证 500 指数的超额收益率. 第二, 本文提出了从财经新闻图像和财经类短视频中挖掘市场信息的模型, 提供了使用另类数据研究投资市场的新思路. 第三, 本文验证了另类数据在中国市场具有重要影响力, 补充了另类数据驱动资产价格的实证结果. 关于研究结果的实际应用, 我们认为本研究的主要贡献在于揭示了财经新闻数据 (包括文本、图像和视频信息) 在股票市场中的预测价值, 基于我们的研究结果, 投资者和市场分析师可以利用财经新闻数据构建更加有效的交易策略和风险管理工具. 此外, 监管机构也可以通过关注财经新闻中的信息, 更好地监测市场动态和潜在风险.

本文已经实现了采用关键帧提取技术对视频图像进行关键帧提取, 并使用深度神经网络对视频图像进行了情绪挖掘, 提高了研究的准确性和可靠性, 提出的交易策略虽实现了较为可观的累计收益, 但仍存在一定的改进和探索空间, 具体而言: 其一, 改善另类数据的处理方式, 直接从音频数据提取情绪信息或将进一步提升模型性能 (Gorodnichenko et al. (2023)). 其二, 考虑更多的情感维度: 本研究主要关注了财经新闻中的情感信息, 未来研究可以尝试从其他维度 (如主题、观点等) 对另类数据进行分析, 以更全面地挖掘另类数据的价值. 最后, 本研究主要关注股票市场, 未涉及其他金融市场 (如债券、期货等). 未来研究可以尝试拓展研究范围, 探讨财经新闻数据在其他金融市场的预测价值.

参 考 文 献

姜富伟, 孟令超, 唐国豪, (2021). 媒体文本情绪与股票回报预测 [J]. 经济学 (季刊), 21(4): 1323–1344.
Jiang F W, Meng L C, Tang G H, (2021). Media Textual Sentiment and Chinese Stock Return Predictability[J]. China Economic Quarterly, 132(1): 126–149.

姜树广, 韦倩, (2013). 信念与心理博弈: 理论、实证与应用 [J]. 经济研究, 48(6): 141–154.
Jiang S G, Wei Q, (2013). Belief and Psychological Games: Theory, Evidence and Applications[J]. Economic Research Journal, 48(6): 141–154.

李龙飞, (2021). 空间计量经济学中的空间自回归模型 [J]. 计量经济学报, 1(1): 36–65.
Lee L F, (2021). The Spatial Autoregression Model in Spatial Econometrics[J]. China Journal of Econometrics, 1(1): 36–65.

廖理, (2021). 另类数据: 经济增长的新亮点 [J]. 学术前沿, (6): 22–27.
Liao L, (2021). Alternative Data: A New Area of Economic Growth[J]. Frontiers, (6): 22–27.

廖理, 崔向博, 孙琼, (2021). 另类数据的信息含量研究 —— 来自电商销售的证据 [J]. 管理世界, 37(9): 90–103.
Liao L, Cui X B, Sun Q, (2021). The Information Content of Alternative Data: Evidence from E-commerce Sales[J]. Journal of Management World, 37(9): 90–103.

林建浩, 张一帆, 陈良源, 邓益萌, (2022). 基于新闻情绪的机器学习交易策略 [J]. 计量经济学报, 2(4): 881–908.

Lin J H, Zhang Y F, Chen L Y, Deng Y M, (2022). News Sentiment and Machine Learning Investment Strategy[J]. China Journal of Econometrics, 2(4): 881–908.

林耀虎, 刘善存, 杨海军, (2022). 一种基于机器学习和蜡烛图的股市投资策略研究 [J]. 计量经济学报, 2(1): 126–140.

Lin Y H, Liu S C, Yang H J, (2022). A Novel Stock Investment Strategy Using Fusion of Machine Learning Techniques and Candlestick Charting[J]. China Journal of Econometrics, 2(1): 126–140.

谭松涛, 崔小勇, 孙艳梅, (2014). 媒体报道、机构交易与股价的波动性 [J]. 金融研究, 25(3): 180–193.

Tan S T, Cui X Y, Sun Y M, (2014). Does Institutional Investors' Trading Behavior Exacerbate Stock Market Volatility?[J]. Journal of Financial Research, 25(3): 180–193.

王正位, 崔向博, 廖理, (2022). 线上销售、市场反应与未来股票收益 [J]. 经济学报, 9(2): 146–165.

Wang Z W, Cui X B, Liao L, (2022). Online Sales, Market Reaction and Future Stock Returns[J]. China Journal of Economics, 9(2): 146–165.

肖争艳, 陈衍, 陈小亮, 陈彦斌, (2022). 通货膨胀影响因素识别 —— 基于机器学习方法的再检验 [J]. 统计研究, 39(6): 132–147.

Xiao Z Y, Chen K, Chen X L, Chen Y B, (2022). Identifying the Influencing Factors of Inflation: Reexamination Based on Machine Learning Methods[J]. Statistical Research, 39(6): 132–147.

游家兴, 吴静, (2012). 沉默的螺旋: 媒体情绪与资产误定价 [J]. 经济研究, 47(7): 141–152.

You J X, Wu J, (2012). Spiral of Silence: Media Sentiment and the Asset Mispricing[J]. Economic Research Journal, 47(7): 141–152.

张宗新, 吴钊颖, (2021). 媒体情绪传染与分析师乐观偏差 —— 基于机器学习文本分析方法的经验证据 [J]. 管理世界, 37(1): 170–185.

Zhang Z X, Wu Z Y, (2021). Media's Emotional Contagion and Analyst Optimistic Bias: Evidence Based on the Technique of Machine Learning[J]. Journal of Management World, 37(1): 170–185.

周颖刚, 纪洋, 倪骁然, 谢沛霖, (2022). 金融学的发展趋势和挑战与中国金融学的机遇 [J]. 计量经济学报, 2(3): 465–489.

Zhou Y G, Ji Y, Ni X R, Xie P L, (2022). Development Trend & Challenges of Finance Research and Opportunities of China's Finance[J]. China Journal of Econometrics, 2(3): 465–489.

Bartov E, Faurel L, Mohanram P S, (2018). Can Twitter Help Predict Firm-level Earnings and Stock Returns?[J]. The Accounting Review, 93(3): 25–57.

Bollen J, Mao H, Zeng X J, (2011). Twitter Mood Predicts the Stock Market[J]. Journal of Computational Science, 2(1): 1–8.

Campbell J, Thompson S B, (2008). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?[J]. Review of Financial Studies, 21(4): 1509–1531.

Campos V, Jou B, Giro-i-Nieto X, (2017). From Pixels to Sentiment: Fine-tuning CNNs for Visual Sentiment Prediction[J]. Image and Vision Computing, 65: 15–22.

Chan W, (2003). Stock Price Reaction to News and No-news: Drift and Reversal after Headlines[J]. Journal of Financial Economics, 70(2): 223–260.

Chen H, De P, Hu Y J, Hwang B H, (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media[J]. Review of Financial Studies, 27(5): 1367–1403.

Chen, Xie Y J, (2008). Online Consumer Review: Word-of-mouth as a New Element of Marketing Communication Mix[J]. Management Science, 54(3): 477–491.

Chinco A, Clark-Joseph A D, Ye M, (2019). Sparse Signals in the Cross-section of Returns[J]. The Journal of Finance, 74(1): 449–492.

Chordia T, Subrahmanyam A, Tong Q, (2014). Have Capital Market Anomalies Attenuated in The Recent Era of High Liquidity and Trading Activity?[J]. Journal of Accounting and Economics, 58(1): 41–58.

Christopher K, Xuan A D, Nicolas H, (2017). Deep Neural Networks, Gradient-boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500[J]. European Journal of Operational Research, 259(2): 689–702.

Dyck A, Morse A, Zingales L, (2007). Who Blows the Whistle on Corporate Fraud?[J]. Journal of Finance, 65(6): 2213–2253.

Fan J Q, Xue L R, Zhou Y, (2021). How Much Can Machines Learn Finance From Chinese Text Data?[J]. Social Science Electronic Publishing. Doi: 10.2139/ssrn.3765862.

Fang L, Peress J, (2009). Media Coverage and the Cross-section of Stock Returns[J]. The Journal of Finance, 64(5): 2023–2052.

Freitas F D, Souza A, Almeida A, (2009). Prediction-based Portfolio Optimization Model Using Neural Networks[J]. Neurocomputing, 72(10–12): 2155–2170.

Froot K, Kang N, Ozik G, Sadka R, (2017). What do Measures of Real-time Corporate Sales Say about Earnings Surprises and Post-announcement Returns?[J]. Journal of Financial Economics, 125(1): 143–162.

Gomez-Cram R, Grotteria M, (2022). Real-time Price Discovery via Verbal Communication: Method and Application to Fedspeak[J]. Journal of Financial Economics, 143(3): 993–1025.

Gorodnichenko Y, Pham T, Talavera O, (2023). The Voice of Monetary Policy[J]. American Economic Review, 113(2): 548–84.

Gu S, Kelly B, Xiu D, (2020). Empirical Asset Pricing via Machine Learning[J]. The Review of Financial Studies, 33(5): 2223–2273.

He K, Zhang X, Ren S, Sun J, (2016). Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 770–778.

Hirshleifer D, Teoh S H, (2003). Limited Attention, Information Disclosure, and Financial Reporting[J]. Journal of Accounting and Economics, 36(1–3): 337–386.

Huang J, (2018). The Customer Knows Best: The Investment Value of Consumer Opinions[J]. Journal of Financial Economics, 128(1): 164–182.

Huberman G, Regev T, (2001). Contagious Speculation and a Cure for Cancer: A Nonevent that Made Stock Prices Soar[J]. Journal of Finance, 56(1): 387–396.

Hu A, Ma S, (2021). Persuading Investors: A Video-Based Study[R]. NBER Working Papers, National Bureau of Economic Research.

Jacob B, Ronen F, Shimon K, Matthew R, (2019). Information, Trading, and Volatility: Evidence from Firm-specific News[J]. The Review of Financial Studies, 32(3): 992–1033.

Jagtiani J, Lemieux C M, (2018). The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the Lendingclub Consumer Platform[R]. FRB-Philadelphia: Working Papers (Topic).

Jame R, Johnston R, Markov S, Wolfe M C, (2016). The Value of Crowdsourced Earnings Forecasts[J]. Journal of Accounting Research, 54(4): 1077–1110.

Jeffrey A B, Green T C, (2002). Market Efficiency in Real-time[J]. Journal of Financial Economics, 65(3): 415–437.

Jiao P, Veiga A, Walther A, (2020). Social Media, News Media and the Stock Market[J]. Journal of Economic Behavior & Organization, 176: 63–90.

Jiang J W, Kelly B T, Xiu D, (2020), (Re-)Imag(in)ing Price Trends[J]. SSRN Electronic Journal. Doi:10.2139/ssrn.3756587.

Julapa J, Catharine L, (2019). The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the Lending Club Consumer Platform[J]. Financial Management, 48(4): 1009–1029.

Kelly B, Palhares D, Pruitt S, (2020). Modeling Corporate Bond Returns[J]. Capital Markets: Asset Pricing & Valuation eJournal, ssrn.3720789.

Krauss A, (2021). Assessing the Overall Validity of Randomised Controlled Trials[J]. International Studies in the Philosophy of Science, 34(3): 159–182.

Lily F, Joel P, (2009). Media Coverage and the Cross-section of Stock Returns[J]. Journal of Finance, 64(5): 2023–2052.

Lin C, Huang J, Gen M, Tzeng G, (2006). Recurrent Neural Network for Dynamic Portfolio Selection[J]. Applied Mathematics and Computation, 175(2): 1139–1146.

Mayew W J, Sethuraman M, Venkatachalam M, (2020). Individual Analysts' Stock Recommendations, Earnings Forecasts, and the Informativeness of Conference Call Question snd Answer Sessions[J]. The Accounting Review, 95(6): 311–337.

Mullainathan S, Shleifer A, (2005). The Market for News[J]. American Economic Review, 95(4): 1031–1053.

Obaid K, Pukthuanthong K, (2022). A Picture is Worth a Thousand Words: Measuring Investor Sentiment by Combining Machine Learning and Photos from News[J]. Journal of Financial Economics, 144(1): 273–297.

Rinallo D, Basuroy S, (2009). Does Advertising Spending Influence Media Coverage of the Advertiser?[J]. Journal of Marketing, 73(6): 33–46.

Schumaker R P, Chen H, (2009). Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZF in Text System[J]. ACM Transactions on Information Systems, 27(2): 1–19.

Tang V W, (2018). Wisdom of Crowds: Cross-sectional Variation in the Informativeness of Third-party-generated Product Information on Twitter[J]. Journal of Accounting Research, 56(3): 989–1034.

Tetlock P C, (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market[J]. Journal of Finance, 62(3): 1139–1168.

Tetlock P C, Saar-Tsechansky M, Macskassy S, (2008). More than Words: Quantifying Language to Measure Firms' Fundamentals[J]. The Journal of Finance, 63(3): 1437–1467.

You Q, Luo J, Jin H, Yang J, (2015). Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks[C]// Proceedings of the Twenty-ninth AAAI Conference on Artificial Intel ligence (AAAI): 381–388.

Zhang J N L, Hardle W K, Chen C Y, Bommes E, (2016). Distillation of News Flow Into Analysis of Stock Reactions[J]. Journal of Business & Economic Statistics, 34(4): 547–563.

Zhu C, (2019). Big Data as a Governance Mechanism[J]. The Review of Financial Studies, 32(5): 2021–2061.