

基于新闻文本情绪的区间值股票回报预测研究

张飞鹏¹, 徐一雄¹, 陈曦¹, 周勇²

(1. 西安交通大学经济与金融学院, 西安 710049; 2. 华东师范大学统计学院和统计交叉科学研究院, 统计与数据科学前沿理论及应用教育部重点实验室, 上海 200062)

摘要 投资者情绪与股票市场的价格变动息息相关, 所以正确理解投资者情绪对金融投资者的投资策略选择与监管部门的风险管控具有重要意义. 本文选取国务院新闻文本与金融情感词典, 首先构建一个基于粉丝加权的新闻媒体情绪区间指数, 然后建立自回归条件区间模型, 对我国股市主板市场与不同板块子市场进行预测研究. 实证研究结果表明: 1) 基于粉丝加权的新闻媒体情绪指数能够有效地反映新闻情绪, 且对我国股票市场回报产生显著的负向影响; 2) 不论在主板市场还是不同板块子市场, 粉丝加权的新闻媒体情绪指数对我国股市回报的预测效果显著最优, 为文本情感指标的构建提供了新的思路与实践; 3) 新闻媒体情绪对主板市场股票回报的预测效果随着预测步长的增加而减弱, 但对不同板块子市场股票回报的预测效果具有异质性. 研究表明基于文本挖掘的新闻媒体情绪区间指数对股市区间值回报预测具有重要作用.

关键词 股市预测; 文本挖掘; 情感分析; 区间数据

Forecasting Interval Valued Stock Returns Based on News Media Sentiments

ZHANG Feipeng¹, XU Yixiong¹, CHEN Xi¹, ZHOU Yong²

(1. School of Economics and Finance, Xi'an Jiaotong University, Xi'an 710049, China; 2. MOE Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Academy of Statistics and Interdisciplinary Sciences and School of Statistics, East China Normal University, Shanghai 200062, China)

收稿日期: 2023-04-28

基金项目: 国家自然科学基金 (72171192, 71931004)

Supported by National Natural Science Foundation of China (72171192, 71931004)

作者简介: 张飞鹏, 教授, 博士生导师, 研究方向: 金融计量、经济与金融预测, E-mail: zhangfp@xjtu.edu.cn; 徐一雄, 博士研究生, 研究方向: 金融计量、金融风险, E-mail: xyx123@stu.xjtu.edu.cn; 陈曦, 硕士研究生, 研究方向: 经济与金融预测, E-mail: happychenxi828@163.com; 周勇, 长江学者特聘教授, 国家杰出青年科学基金获得者, 博士生导师, 研究方向: 数量金融、风险管理、管理统计学, E-mail: yzhou@fem.ecnu.edu.cn.

Abstract Since the investor sentiment is closely related to the price movement of Chinese stock market, it is crucial to correctly understand the investor sentiment for both financial investors and regulators for risk management. This paper aims to construct a fan-weighted news sentiment index based on the State Council news texts and financial sentiments dictionary, and then to investigate the predictability of the fan-weighted news sentiment indicator on the main board market and sub-markets of Chinese stock market by an interval-valued autoregressive model. The empirical results show that: 1) Our proposed fan-weighted news sentiment index can not only reflect news sentiment effectively, but also has a significant negative impact on Chinese stock market; 2) Compared with the simple average news sentiment index, our proposed fan-weighted news sentiment index can significantly improve the predictive ability on Chinese stock market under both the main board market and sub-markets, which provides a new idea for constructing sentiment indicators; 3) The predictive ability of news sentiment index gradually decreases over time for the main board market, but it is heterogeneous for the sub-markets. These empirical findings show that the investor sentiment mined from news text is important to predict the stock market performance.

Keywords stock prices forecasting; text mining; emotional analysis; interval-valued data

1 引言

股票市场在金融市场中占据着重要地位, 股票市场回报的可预测性一直是学者们研究的热点问题. 根据 Fama (1970) 的有效市场假说, 如果股票市场是完全有效的, 则在任何时点上的证券交易价格是当下所有可得信息的体现, 且证券价格的变化仅取决于新信息. 在该假说下, 证券价格的变化应该满足随机游走, 即价格的收益序列是相互独立的, 从而股票回报是不可预测的. 注意到, 有效市场假说的前提假设是, 参与市场的投资者是理性人且目标是追求利益最大化. 但是从 1637 年荷兰郁金香事件到 2020 年全球新冠疫情暴发等一系列极端事件都反映出股票市场并不总是满足有效市场假说, 且股票价格会受到市场情绪的影响 (Bollen et al., 2010). 当市场投资者受到诸如过度自信、框定依赖、锚定和调整、损失规避等信念影响时, 投资者在非确定性条件下的市场决策行为不是完全理性的, 可能会出现系统性的认知偏差 (Shefrin, 2002). 当投资者个体情绪受群体情绪感染时, 容易出现盲目跟风的“羊群效应”, 股市高回报时越炒越高, 低回报时过度恐慌. 因此, 证券价格不仅由其内在价值决定, 也会受到投资主体心理因素及行为的影响 (周战强, 2004). 已有文献研究表明, 投资者作为证券市场的主体, 除了认知、文化、阅历等个人因素外, 投资者情绪是影响资产定价的重要主观因素, 与股票收益之间是相互影响、相互制约的关系 (池丽旭等, 2012). 因而, 如何合理地测度投资者情绪, 并对股票回报进行预测成为了学界与业界关注的重要研究方向之一.

在早期研究中, 受收集手段与情绪载体等因素的限制, 投资者情绪多用金融市场指标、经济指数等市场代理变量衡量. Baker and Wurgler (2006)、朱伟骅和张宗新 (2008) 选取封闭式基金的折价率代替投资者情绪发现, 投资者情绪会对股市定价与波动造成影响. 鹿坪和冷军等 (2017) 用消费者信心指数替代投资者情绪也发现类似结论. 余佩琨和钟瑞军 (2009)、

刘维奇和刘新新 (2014) 分别用华鼎多空民意调查结果与投资者月新增开户数代替投资者情绪, 发现机构投资者的情绪可以预测股市, 而个人投资者情绪不具有预测性. 另外, 也有学者通过问卷调查等主观情绪指标构建投资者情绪. 例如, 王美今和孙建军 (2004) 基于中国证券分析师指数构造了中国投资者情绪指数, 发现投资者情绪的变化不仅显著地影响了沪深两市收益, 而且显著地反向修正沪深两市的收益波动. 易志高和茅宁 (2009) 则进一步将封闭式基金折价率等客观指标与央视看盘指数等主观指标结合起来, 构建中国股票市场投资者情绪综合指数 (CICSI), 解释我国股市的一些现象. 卢米雪 (2022) 基于百度搜索指数、股票论坛发帖数量等指标构建 CICSI, 发现极端的投资者情绪会对股市波动产生影响. 上述研究均表明, 投资者情绪与股票市场表现之间存在显著相关性, 但由于代理变量的选取标准的不统一, 已有研究结论存在较大差异, 且不同代理变量对投资者情绪的解释性也值得商榷.

随着计算机和互联网技术的发展, 互联网文本已成为获取市场情绪的重要来源. 相比于传统的金融指标或问卷调查指标, 文本情绪更能直观地反映并替代投资者情绪, 从而基于文本数据的情绪指标逐步成为金融实证研究中的热点方向 (唐国豪等, 2016). 很多研究表明, 文本数据中体现的情感倾向可以有效地预测股票市场回报, 所以文本情绪被广泛应用于资产价格波动预测和行为资产定价等领域的研究中 (Antweiler and Frank, 2004; Li, 2008; Feldman et al., 2010; Price et al., 2012). 但是, 由于股票市场涵盖的信息领域非常广泛, 各类文本信息内涵十分丰富, 因此需要合理地选取文本信息来源, 以及文本分析方法提取文本情绪信息, 进而更好地探究投资者情绪对股市回报的可预测性. 本文在后续文献评述章节从文本分析方法和文本信息来源两方面总结了文本情绪指标的构建及其对股市预测的影响. 综上所述, 考虑到我国股市散户多, 专业性较弱, 易受到市场宏观语调的影响, 而且我国新闻媒体等正式文本的信息密度高, 蕴涵的情绪语调丰富, 传播力度强, 会大范围的引导股市投资者情绪, 进而影响股市回报. 因此, 本文拟采用新闻正式文本为研究对象, 研究新闻媒体情绪对我国股市回报的可预测性.

尽管已有文献对中国股市回报的预测性研究非常丰富 (于志军和杨善林, 2013; 林昱等, 2022; 孙彦林等, 2019), 但大多数文献均采用点值数据进行建模, 难以全面地囊括股票价格的变化信息. 与点值数据不同, 区间数据作为一种新颖的数据形式, 能涵盖价格数值的变化区间, 从而能够包含更全面的数据信息. 因此, 区间数据在股市预测中具有明显的优越性. 例如, 陈炜等 (2022) 基于区间数据对股票市场数据进行预测, 结果表明区间数据在预测精度方面具有优势. 另外, 区间数据模型在原油市场 (Qiao et al., 2019)、汇率市场 (Sun et al., 2018), 宏观经济预测 (周文凯和杨威, 2020) 等领域也被广泛应用, 详见文献评述章节的回顾区间数据建模方法与应用. 值得指出的是, 以往基于投资者情绪对股市回报预测的研究也是基于点值数据模型. 而投资者情绪作为一个相对主观的因素, 本身受到多方面的影响, 例如, 投资主体心理因素与行为 (周战强, 2004), 社会舆论因素 (陆沁晔和陈昊, 2021), 与经济政策不确定性 (刘志峰等, 2023) 等因素. 因此, 由于投资者情绪本身具有不确定性, 使用点值情绪指标虽能反映投资者情绪的单一水平, 但无法同时反映情绪的变化范围与波动情况, 涵盖的情绪信息较低. 特别是, 随着计算机和互联网数据的发展, 文本情绪代替传统的经济指标情绪, 更能直观地表示市场情绪. 然而, 基于点值数据的文本情绪仅能反映情绪变动的水平情况, 而且由于文本内的情绪更加丰富, 点值数据的文本情绪指标所提取的文本情绪信息可能更不充分.

但是, 由于区间数据允许变量在某个区间内取值, 从而将点值文本情绪指标拓展到区间文本情绪指标能够涵盖更多的文本情绪信息. 因此, 本文从区间数据视角出发, 构建区间数据模型, 研究新闻媒体文本情绪对我国股市回报的预测能力. 具体方法为: 首先, 利用网络爬虫技术获取国务院网站新闻媒体报道中金融相关的新闻文本, 基于金融情绪词典构建粉丝加权新闻媒体情绪区间指数; 然后, 构建区间数据条件自回归模型 (ACIX), 研究新闻媒体文本情绪对我国主板股市、不同板块子市场的影响, 进行区间数据预测分析; 最后, 讨论新闻媒体情绪指数对我国股市回报的最优预测步长, 研究新闻媒体情绪对我国股票市场回报的预测性能.

本文的主要贡献在于: 第一, 本文将文本情绪指标从传统的点值数据扩展到区间数据, 以增加数据信息的使用, 提高预测精度. 另外, 本文选取的新闻文本来源是国务院网站新闻板块, 因此能够较好地体现新闻文本的重要性与规范性, 从而本文所构建的新闻媒体情绪区间指数能够反映市场主流情绪, 能对投资者进行正确引导, 进而替代投资者情绪; 第二, 本文将新闻文本传播范围纳入文本情绪指数构建中, 构建粉丝加权的新闻媒体情绪区间指数, 相比于简单平均新闻媒体情绪指数, 本文构建的指数在预测性能方面具有更优异的表现, 这为文本情绪指数的加权构建提供了新思路与新实践; 第三, 本文灵活运用区间数据, 构建新闻媒体情绪区间预测模型, 评估新闻媒体情绪对我国股市回报的预测性能, 且与基准模型进行比较, 是对已有文献研究方法与实践的补充与完善.

本文的结构组织如下: 第2节文献评述, 从文本分析方法和文本来源角度探讨文本情绪指标的构建, 以及它们对股票市场的影响; 第3节介绍数据来源与文本指标, 构建新闻媒体情绪区间指数; 第4节给出区间数据自回归模型估计方法及其统计推断; 第5节提供实证研究结果, 评估新闻文本情绪区间指标对我国股市区间价格的预测性能; 第6节总结全文.

2 文献评述

2.1 金融文本分析方法文献综述

随着计算机技术的发展, 文本大数据的可得性不断提高, 利用文本分析方法提取文本情绪分析金融问题逐渐成为金融实证领域的热点. 相比于与金融市场相关的交易和财务定量数据, 文本情绪指标涵盖面更广, 因而能更好地应用于金融市场研究中 (唐国豪等, 2016). 而文本数据作为非结构化数据, 如何通过对文本内容进行挖掘和数据分析, 提取出有效的文本信息, 需要计算语言学、自然语言处理、信息检索、内容分析和文体学等多方面的综合考虑. 本节从文本分析方法和文本来源角度梳理各类金融文本情绪指标的构建方法及其对各类股票市场回报的可预测性.

从建模分析的角度来看, 现有的文本信息提取方法主要分为基于词袋向量和基于情绪分析两种方法. 基于词袋向量的方法是通过向量空间模型对文本进行建模, 将每个文本信息转换为词语统计量的向量, 如出现的次数等. 词袋方法所形成的向量包含出现的所有重要词句, 因而文字信息非常全面. 但是, 该方法忽略了文档中词语顺序和词语之间的语义关系, 可能会产生歧义, 而且当文档中词语量过大时可能会造成维度灾难 (姚加权等, 2020). Chan and Chong (2017) 讨论不同金融文本信息的提取方法, 结果发现虽然基于词袋方法具有一定预测能力, 但该方法缺乏文本信息中的关键要素——情绪. 然而, 文本情绪反映投资者情绪, 进而影响股票价格的变动, 所以基于词袋的文本分析方法并不适用于投资者情绪对股市回报的

可预测性。因此,结合机器学习的情绪分析法预期比基于词袋方法更具有优势 (Oliveira et al., 2017)。它是将情绪判断转化为文本分类问题,然后运用机器学习方法预测给定文本属于积极或消极类别的概率 (范小云等, 2022)。张宗新和吴钊颖 (2021) 利用上市公司的百度新闻报道作为文本,运用机器学习文本分析方法测算情绪倾向得分,考察媒体情绪对分析师预测行为的影响。Kim and Kim (2014) 采用朴素贝叶斯方法,通过搜集雅虎财经留言板上的用户信息,研究投资者情绪对股票收益、波动性和交易量是否具有预测能力,研究结果并未发现投资者情绪对股市的预测能力。部慧等 (2018) 基于东方财富网股吧帖文与朴素贝叶斯方法构建投资者情绪度量指标,发现该指标对股票市场收益率、交易量和波动性均无预测能力,仅存在当期影响。许雪晨和田侃 (2021) 则将 BERT 模型应用到财经新闻情感分析中,对沪深 300 指数进行预测,有效地提高股指趋势预测的准确率。武静等 (2023) 进一步用深度学习模型对在线新闻与用户评论文本进行情感分析并预测。虽然机器学习方法在预测精度上具有较大优势,但它需要提供机器学习分类的初始训练集。而该训练集往往是随机选取且由人工定义,导致分类种类和方法因人而异,缺乏统一的规范和标准 (唐国豪等, 2016),从而国内股市文本分析预测中未有一致结论。另外一种金融文本情绪测度方法,便是采用成熟的金融词典。基于词典法的文本情绪分析依赖于预先定义的包含情绪得分词汇列表,能够简便且有效地识别情感词汇。尽管 Loughran and McDonald (2011) 从上市公司年报中筛选出的词典适合金融领域研究,从而成为最有影响力的英文金融词典,但是国内还不存在权威的中文金融情绪词典。王靖一和黄益平 (2018) 根据和讯网新闻,构建适用于金融科技领域的情感词典。姚加权等 (2021) 通过词典重组和长短期记忆模型深度学习算法对上市公司年报和社交媒体发帖进行文本分析,构建正式用语情绪词典和非正式用语情绪词典。姜富伟等 (2021) 使用 LM 词典、word2vec 算法和人工筛选等方法构建出一个全新的中文金融文本情感词典,并用该词典提取股票市场的媒体文本情绪,发现媒体文本情绪可以很好地衡量我国股票市场投资者情绪的变化,而且对我国股票回报具有显著的预测能力。

从文本选取的角度来看,文本来源主要分为正式文本和非正式文本两种方式。正式文本主要来源为主流新闻报刊或公司年报,具有流通性广特点,媒体情绪与语调能引导投资者情绪,进而影响股市回报。谭松涛等 (2014) 发现投资者关注媒体报道信息后会对股市产生影响。陆沁晔和陈昊 (2021) 发现媒体报道具有倾向性,能通过影响投资者情绪导致股价波动。于琴和张兵等 (2020) 构建新闻情绪指标,发现媒体报道对股市收益存在媒体强化效应。而且,新闻媒体内容和情绪与上市公司的市场风险和波动性存在明显关联 (顾洪梅和张嫚玲, 2022)。上述研究均表明,新闻媒体报道的正式文本与股市回报之间存在显著关联性。另一种文本来源是非正式文本,主要反映投资者个人的言论及情绪。我国股市的非正式文本主要源自股吧、微博等社交软件。金雪军等 (2012) 基于股吧发帖信息提取投资者意见,构建看涨指数与意见趋同指数,发现看涨指数与股票收益率成正相关,且对第二天股价具有较强预测能力。黄润鹏等 (2015) 发现加入微博情绪信息的预测模型能够获得更高的预测准确率。石善冲等 (2018) 基于微信文本挖掘构建投资者情绪指数,发现它能够稳定预测上证指数收盘价。相比于主流新闻报刊等正式文本,股吧或微博等软件的发帖评论能够直观地反映个体的情绪,对投资者情绪的代替性更强,但是包含大量口语、俚语、网络用语等个性化语言,从而信息密度较低 (钱宇等, 2020)。

2.2 区间数据模型应用文献综述

区间数据, 是指数值变量的变动区间, 广泛存在于经济、金融、社会学等实际应用中. 相比于点值数据, 区间数据包含数据的变动范围, 包含信息丰富, 所以备受学者欢迎. 本节主要对区间数据模型的建模方法与应用研究进行梳理回顾.

区间数据的研究始于 Moore (1996) 提出区间分析, 旨在将一系列数据作为区间来处理, 以解决数学模型中变量不确定取值的问题, 使得计算结果更加准确. 因此, 区间数据所含信息大, 不仅能够同时反映变量的水平与范围信息, 还能直观地给出预测变量的有效区间变化情况. 区间数据的建模研究主要从区间数据点值属性、区间运算下的简单区间回归以及预测等角度入手 (杨威等, 2016). 考虑到区间可以由中点与极差 (或上界与下界) 唯一确定, 一种常用的建模方法便是采用传统的二维变量统计方法对区间数据的中点和极差建立联立模型. 它的基本思想是针对区间数据的中点与极差过程分别建模. Billard and Diday (2000) 利用中点方法 (center method, CM) 来拟合区间数据回归模型, 先对区间中点建立点值回归模型, 然后将所得到的回归系数分别应用于区间的上界和下界得到对应的区间. 接着, Billard and Diday (2003) 提出最小最大方法 (min-max method), 通过区间的上界和下界分别建立回归模型进行区间拟合和预测. Neto and De Carvalho (2008) 提出中点与极差方法 (center and range method, CRM), 首先通过区间中点和区间极差分别建立回归模型, 然后根据区间生成方式给出区间上界和下界, 比 Billard and Diday (2000, 2003) 提出的中点方法更加有效. 而 Arroyo et al. (2011) 认为区间的两个点值过程相互影响, 因此需要在建模过程中加入各自的滞后项, 建立二维向量自回归 (VAR) 模型.

虽然上述建模方法也能给出区间形式的结果, 但他们仍然是在传统点值模型框架下进行研究, 不能从计量经济分析角度给出区间数据模型参数估计的统计性质与相应的经济解释. 正如 Gil et al. (2007) 指出, 利用区间点值属性构建回归模型必须考虑一些重要的限制条件, 譬如区间模型本身所暗含的区间预测下界小于区间上界的假设条件, 导致现有的最小二乘方法无法获得它们的模型参数估计的最优解. 尽管 Neto and De Carvalho (2010) 提出带有限制条件的区间回归模型能保证预测区间结构的一致性, 但是约束条件给参数估计带来额外负担. 为了有效地解决上述研究中区间结构不一致等问题, Blanco-Fernández (2011), Han et al. (2012) 提出基于区间运算法建立的区间线性回归模型. Han et al. (2016) 从区间样本总体出发, 首次提出区间自回归模型 (autoregressive conditional interval model, ACI), 拓展了区间定义和区间运算法则, 给出模型参数估计方法及其统计性质. ACI 模型不仅能够直接对区间型数据建模, 捕捉区间过程的动态特征, 而且最小 D_K -区间距离估计方法能够给出有效的参数估计.

综上所述, 区间数据模型在经济、金融等实际中有广泛的应用. He et al. (2009)、Hu et al. (2010) 分别利用股票市场和抵押贷款等数据, 实证研究表明基于区间运算得到区间预测要比传统基于点值运算和置信区间方法得到的区间预测精准. Yang and Han (2015) 基于区间数据模型和点值自回归模型分别对金融资产价格极差进行预测, 发现区间数据模型对极差波动存在预测优势. 杨威等 (2016) 从区间时间序列模型出发, 利用美国股票市场和中国股票市场数据研究发现区间时序模型比传统点值 AR 模型、VAR 模型更具有稳健的预测优势. 上述研究均表明区间数据模型相比于点值数据模型, 存在明显的预测优势, 并且能直观地表

示出预测区间. 因此, 本文利用区间数据模型对我国股票市场回报进行建模分析预测.

3 新闻媒体情绪区间指数构建

3.1 样本选取与语料来源

由于新闻媒体报道内容非常正式, 文本信息密度高, 包含市场、媒体、政策等多种情绪信息, 且覆盖内容广, 受众人群多, 所以新闻媒体情绪对市场具有较大的潜在影响. 为了凸显新闻媒体所发布信息的重要性, 本文选取中华人民共和国中央人民政府网站 (www.gov.cn) 上新闻板块进行文本爬取. 考虑到 2014 年以前该网站的转载新闻未包含新闻的具体来源, 本文爬取文本数据的时间跨度为 2014/02/03-2022/04/30, 总计 47583 条新闻数据. 新闻文本来源不仅包括官方性质的报刊 (人民日报、新华社等), 具有较大影响力的新闻媒体 (新京报等), 还包括各类地方报刊 (重庆日报等). 值得指出的是, 国务院网站转载的新闻信息通常具有较高的经济政策含义, 对市场投资者的导向性强, 因而能更好地反映投资者情绪.

考虑到新闻文本内容驳杂, 为了减少无关新闻文本信息的干扰, 本文以“金融”“股市”“股票”“经济”等经济金融词汇为索引, 爬取的文本内容必须包含上述词汇, 然后再通过人工筛除不合理新闻, 挑选出与经济金融领域相关的新闻报道, 以凸显出市场情绪对股市回报的可预测性. 图 1 给出本文爬取的金融相关文本条数变化情况. 由图可知, 每月新闻数据相对均衡, 新闻报道转发量在 400 条左右, 只有少数月份出现大幅波动. 例如, 2015 年 6 月是股灾发生, 2016 年 10 月是“一人三户”股市政策推行.

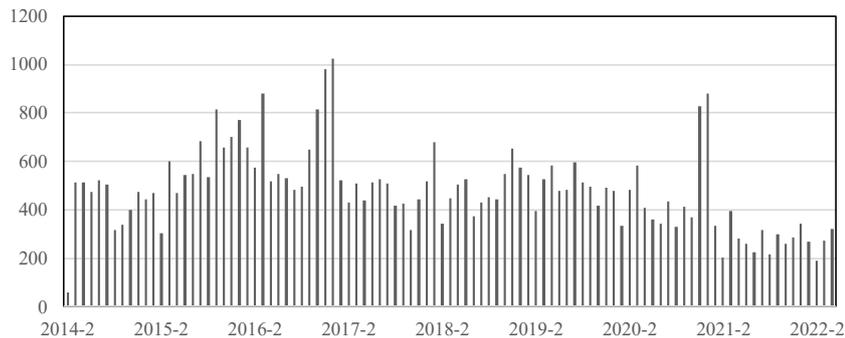


图 1 金融相关新闻文本条数变化情况

3.2 构建新闻文本情绪区间指数

为构建新闻文本情绪指数, 本文借鉴姜富伟等 (2021) 所构建的金融情绪词典来提取文章情绪. 该词典借鉴国外金融市场最具影响力的 LM 金融词典, 且扩充为更全面的中文金融词典, 能够更好地刻画新闻媒体文本情绪. 具体步骤如下:

第一步, 首先对爬取的新闻文本进行分词处理, 去除停用词语. 然后, 利用情感词典筛选出所有情感词语.

第二步, 由于否定词可以改变情感倾向, 本文将情感词、否定词组合为情感单元. 我们假设情感词只受其之前的词汇影响, 因此, 将前一个情感词之后到该情感词为止作为一个情感单元. 图 2 为情感单元的图形表示.

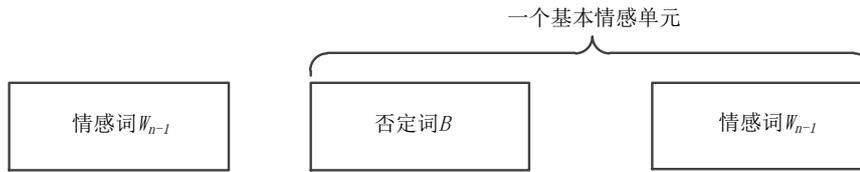


图 2 情感单元示意图

第三步, 赋予情感得分. 本文将积极词权重设为 1, 消极词权重设为 -1 . 考虑到双重否定表示肯定的情况, 否定词的权重为 $(-1)^n$, 其中, n 为否定词出现的次数. 那么, 第 j 篇文章的情绪值为

$$\text{sentiment}_j = N^{-1} \sum_{i=1}^N W_i (-1)^n, \quad (1)$$

其中, N 为全文中的情感单元总数, W_i 为情感词得分,

$$W_i = \begin{cases} 1, & \text{positive,} \\ -1, & \text{negative.} \end{cases}$$

第四步, 构建粉丝加权新闻情绪指标. 由于网络新闻媒体具有发布渠道多、传播速度快、影响力持久等特点, 很容易引起投资者关注和情绪波动, 从而引起股市价格回报的变动 (Barber and Odean, 2008). 但现有文献对新闻文本情绪指标的构建, 主要通过对当日内所有新闻文本情绪指标的简单平均获得, 认为不同新闻报道的影响力是相同的, 并没有考虑到不同新闻媒体的影响力存在差异, 从而导致新闻文本的传播范围存在差异, 所以简单平均的文本情绪指标难以真实反映主流的新闻媒体情绪. 为此, 少数文献研究将权重纳入情绪指标体系构建. 例如, 吕华揆等 (2021) 基于带权重的情感词典对新闻情感进行量化, Liu et al. (2018) 提出神经网络架构——层次互补注意网络 (HCAN), 采用两级注意机制以量化给定新闻中单词和句子的重要性. 但是, 他们的权重设定主要用于文本情绪提取, 而无直观的解释意义. 因此, 本文针对新闻媒体传播特性, 将新闻媒体影响力纳入文本情绪考虑中, 提出基于新闻文本传播范围进行加权, 构建加权新闻媒体情绪指数. 考虑到难以客观的衡量新闻文本信息的真实传播, 本文用新闻媒体在主流信息传播平台 (微博、今日头条和百度百家) 上的粉丝值进行替代加权. 具体公式为:

$$\text{Sentiment}^{\text{FW}} = \frac{\sum_{j=1}^J \text{sentiment}_j \times \ln(\text{fans}_j)}{\sum_{j=1}^J \ln(\text{fans}_j)}, \quad (2)$$

其中, $\text{Sentiment}^{\text{FW}}$ 表示粉丝加权的日度文本情绪指数, fans_j 表示文章 j 对应媒体的粉丝数, J 则表示当天的新闻文本数量. 进一步, 周度的新闻媒体情绪区间指数可表示为 $[\text{Sentiment}_L^{\text{FW}}, \text{Sentiment}_H^{\text{FW}}]$, 即一周内新闻媒体情绪指数的最低值和最高值. 图 3 描绘了不同新闻来源媒体的粉丝占比. 其中, 新华社、人民日报等主流媒体在主流平台上的粉丝超过 1 亿, 是影响力最大的新闻媒体; 粉丝数在 1000~5000 万的媒体有 25 家, 具有极大的社会影响力, 其余 100 多家粉丝数小于 1000 万的媒体, 传播范围较为局限, 影响力相对较低.

可以看出,由粉丝数量所表现的新闻媒体影响力存在较大的差异,导致新闻媒体报道对社会、经济金融的影响力度也存在较大差异。

图4描绘粉丝加权新闻媒体情绪区间指数的变化走势。由图4可知,新闻情绪值整体呈现正面趋势,但粉丝加权新闻媒体情绪区间指数的趋势与波动能够反映出我国股市的变动情况。例如,在2015年、2016年、2018年与2020年新冠疫情期间,股票市场处于熊市,市场情绪偏向悲观,所构建的粉丝加权新闻媒体情绪区间指数也存在显著的下降趋势。而在2020年上半年的短期牛市行情下,我国股市综合指数不断高涨,股市回报提升,所构建的粉丝加权新闻媒体情绪区间指数也呈现出上升趋势。直观上,本文构建的加权新闻媒体情绪区间指数能够很好地反映我国股市行情的变化。

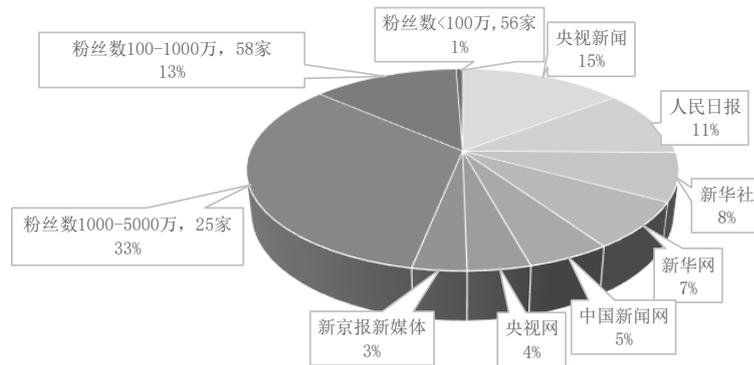


图3 媒体新闻来源及粉丝数量占比

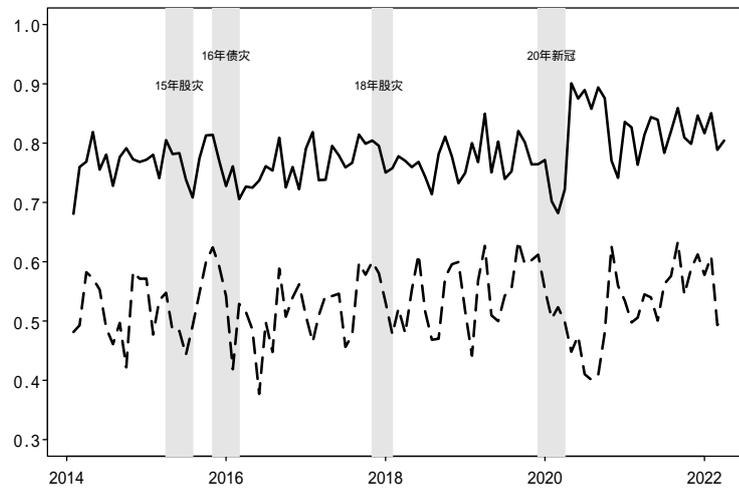


图4 加权新闻情绪区间指数变化图

4 区间数据模型估计与统计推断

本文采用 Han et al. (2016) 的区间运算法则,构建自回归条件区间模型 (ACIX) 对我国股市回报进行建模预测分析。

4.1 自回归条件区间模型

假设 $K_c(R)$ 为实数 R 中的非空紧的区间集合, 随机选取区间 $A, B \in K_c(R)$, 其中, $A = [A_L, A_H]$, $B = [B_L, B_H]$, A_L, B_L 和 A_H, B_H 分别为区间 A, B 的上界和下界. 在 $K_c(R)$ 上定义:

- 1) 区间加法运算 $A + B = [A_L + B_L, A_H + B_H]$;
- 2) 区间数乘运算 $\beta \cdot A = [\beta \cdot A_L, \beta \cdot A_H]$;
- 3) 区间差分运算 $A - B = [A_L - B_L, A_H - B_H]$.

基于上述区间运算法则, Han et al. (2012) 提出带外生变量的自回归条件区间模型 (ACIX). 假设 $\{Y_t : Y_t = [Y_{L,t}, Y_{H,t}]\}_{t=1}^T$ 为随机区间过程, 则 ACIX(p, q, s) 可以表示为:

$$Y_t = \alpha_0 + \beta_0 I_0 + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=1}^q \gamma_j u_{t-j} + \sum_{k=0}^s \delta'_k X_{t-k} + u_t, \tag{3}$$

其中, Y_{t-i} ($i = 1, \dots, p$) 为 Y_t 的滞后期项, X_t 为外生平稳区间向量过程, $u_t = [u_{L,t}, u_{H,t}]$ 是关于信息集 I_{t-1} 区间鞅差分序列过程, 满足 $E(u_t | I_{t-1}) = [0, 0]$, $I_0 = [-\frac{1}{2}, \frac{1}{2}]$ 为常值单位区间, $\alpha_0 + \beta_0 I_0 = [\alpha_0 - \frac{1}{2}\beta_0, \alpha_0 + \frac{1}{2}\beta_0]$ 为区间截距项. 这里, $\theta = (\alpha_0, \beta_i, \gamma_j, \delta_k)$ 为待估计参数, $i = 0, 1, \dots, p$; $j = 0, 1, \dots, q$; $k = 0, 1, \dots, s$.

4.2 模型估计与推断

ACIX 模型采用最小 D_K 距离估计, 其中, D_K 距离用于度量两个区间之间的距离. 对于任意 $A, B \in K_c(R)$, D_K 距离可表示为:

$$D_K(A, B) = \sqrt{\int_{S^0} (s_A(u) - s_B(u))(s_A(v) - s_B(v))dK(u, v)}, \tag{4}$$

其中, $K(\cdot)$ 为核函数, 函数 s 是从空间 $K_c(R)$ 到 Hilbert 空间中闭凸锥 $[C(S^0), \|\cdot\|_K]$ 的一个等距映射. 如果用 $\langle \cdot, \cdot \rangle_K$ 表示对应的内积, 那么有 $D_K^2(A, B) = \langle s_A - s_B, s_A - s_B \rangle_K$.

为了估计模型 (3) 中参数向量 $\theta = [\alpha_0, \beta_0, \dots, \beta_p, \gamma_1, \dots, \gamma_q, \theta'_1, \dots, \theta'_s]'$, 可以采用基于最小 D_K 距离的参数估计量 $\hat{\theta}$:

$$\begin{aligned} \hat{\theta} &= \arg \min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T q_t(\theta), \\ q_t(\theta) &= u_{t,K}^2(\theta) = D_K^2[u_t(\theta), 0], \end{aligned}$$

其中, $u_{t,K}(\theta) = Y_t - [\alpha_0 + \beta_0 I_0 + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=1}^q \gamma_j u_{t-j}(\theta) + \sum_{k=0}^s \delta'_k X_{t-k}]$.

尽管上述最小距离估计量 $\hat{\theta}$ 是相合的, 但是它依赖于核函数 K 的选择. 为此, 本文采用 Han et al. (2016) 提出的两阶段 D_K 距离估计方法. 该方法是在所有满足条件的核函数中最有效的 (Sun et al., 2018; He et al., 2021). 具体方法是:

第一阶段, 代入指定的核函数 K , 得到最优的核函数 \hat{K}^{opt} ,

$$\hat{K}^{\text{opt}} = \frac{1}{T} \sum_{t=1}^T \begin{bmatrix} \hat{u}_{L,t}^2(\hat{\theta}) & \hat{u}_{L,t}(\hat{\theta})\hat{u}_{H,t}(\hat{\theta}) \\ \hat{u}_{H,t}(\hat{\theta})\hat{u}_{L,t}(\hat{\theta}) & \hat{u}_{H,t}^2(\hat{\theta}) \end{bmatrix}. \tag{5}$$

第二阶段,重新代入最优核函数 $K = \hat{K}^{\text{opt}}$, 可得最优参数估计 $\hat{\theta}^{\text{opt}}$:

$$\hat{\theta}^{\text{opt}} = \arg \min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T D_{K^{\text{opt}}}^2[u_t(\theta), 0]. \quad (6)$$

ACIX 模型中参数估计的显著性可以通过区间的 Wald 统计量进行检验. 原假设定义为:

$$H_0 : R\theta = r,$$

其中, R 是一个 $l \times m$ 的满秩矩阵且 $l \leq m$, r 是一个 $l \times 1$ 的向量, m 是模型 ACIX(p, q, s) 中参数 θ 的维数. Wald 检验统计量可定义为:

$$W = [T(R\hat{\theta} - r)]' [R\hat{M}_T^{-1}(\hat{\theta})\hat{V}_T(\hat{\theta})\hat{M}_T^{-1}(\hat{\theta})R']^{-1} [(R\hat{\theta} - r)], \quad (7)$$

其中, $\hat{\theta}$ 是 D_K -距离估计量, $\hat{M}_T(\hat{\theta}) = -\frac{1}{T} \sum_{t=1}^T \frac{\partial^2 q_t(\hat{\theta})}{\partial \theta \partial \theta'}$, $\hat{V}_T(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^T \frac{\partial q_t(\hat{\theta})}{\partial \theta} \frac{\partial q_t(\hat{\theta})}{\partial \theta'}$. 可以证明, 当 $T \rightarrow \infty$ 时, $W \xrightarrow{D} \chi^2(l)$.

4.3 变量选择

为了选择模型最优滞后阶, 本文采用 Sun et al. (2020) 提出的区间 Mallows 准则估计量:

$$C_p = \frac{SS_{\text{ep}}}{\sigma^2} - 2T + 2p, \quad (8)$$

其中, $SS_{\text{ep}} = \sum_{t=1}^T u_{t,K}(\hat{\theta})^2$ 是 m 维模型基于 D_K 距离的区间值残差平方和, $\sigma^2 = E\|u_t\|_K^2$, T 为时间期数. 通过最小化估计量 C_p , 可以选择最优的模型滞后阶.

5 实证结果

本节将探究新闻媒体情绪是否对我国股票市场回报具有预测能力¹. 首先, 本文研究粉丝加权的新闻媒体情绪指数对我国主板市场股票回报的预测性能. 主板市场选取深证指数 (SZ) 2014/02/07–2022/04/29 的周度区间数据. 数据来源为 Wind 数据库. 然后, 参照申万一级行业分类指标, 本文选取医药生物、非银金融、计算机、电子、银行和农林牧渔六个分板块子市场, 检验新闻媒体情绪对子市场股票回报的预测能力. 子市场样本期为 2014/02/07–2022/04/29 的周度区间数据. 数据来源为 Wind 数据库. 接下来, 研究新闻媒体情绪在极端事件下的预测性能. 最后, 研究新闻媒体情绪区间指数的预测性能进行稳健性检验, 讨论不同预测步长对预测结果的影响.

表 1 给出粉丝加权情绪区间指数、深证指数 (SZ) 和电子 (DZ)、计算机 (JSJ)、农林牧渔 (NLMY)、医药生物 (YYSW)、银行 (YH)、非银金融 (FYJR) 六个子市场区间对数回报的描述性统计结果. 从表中可以看到, 粉丝加权新闻媒体情绪指数的区间上、下界均值分别是 0.781 与 0.530, 说明新闻媒体对国内股市的情绪总体偏积极. 每个区间序列上下界的 ADF 检验均显著拒绝原假设, 说明序列是平稳的, 符合自回归条件区间模型的建模要求

¹本文所使用的数据和代码请参见科学数据银行 (ScienceDB) 期刊社区, DOI: 10.57760/sciencedb.j00214.00017 和 CSTR: 31253.11.sciencedb.j00214.00017. 若使用文中数据信息, 请注明引文和数据出处.

表 1 描述性统计

	均值	最大值	最小值	标准差	偏度	峰度	Jarque-Bera	ADF
Sentiment _H ^{FW}	0.781	1	0.599	0.065	0.366	0.246	11.278***	-5.321***
Sentiment _L ^{FW}	0.53	0.73	-0.133	0.108	-1.953	7.345	1301.753***	-6.404***
Δ ln SZ _H	0.001	0.136	-0.176	0.03	-0.785	9.459	778.689***	-6.252***
Δ ln SZ _L	0.001	0.105	-0.202	0.037	-1.332	8.543	666.730***	-6.909***
Δ ln DZ _H	0.003	0.168	-0.232	0.05	-0.361	5.432	113.475***	-6.996***
Δ ln DZ _L	0.003	0.163	-0.279	0.052	-0.825	6.814	304.350***	-7.285***
Δ ln JSJ _H	0.001	0.143	-0.212	0.041	-0.5	6.745	264.844***	-6.362***
Δ ln JSJ _L	0.001	0.171	-0.257	0.047	-0.934	7.609	436.003***	-6.805***
Δ ln NLMY _H	0.002	0.135	-0.185	0.036	-0.408	6.844	272.230***	-6.454***
Δ ln NLMY _L	0.001	0.157	-0.235	0.039	-1.117	10.04	961.580***	-6.821***
Δ ln YYSW _H	0.001	0.114	-0.172	0.032	-0.934	7.009	344.696***	-6.582***
Δ ln YYSW _L	0.001	0.111	-0.192	0.037	-1.157	7.304	420.873***	-6.856***
Δ ln YH _H	0.001	0.168	-0.108	0.029	0.694	7.166	339.812***	-7.340***
Δ ln YH _L	0.001	0.142	-0.174	0.029	-0.159	9.296	700.500***	-7.668***
Δ ln FYJR _H	0.000	0.269	-0.26	0.048	0.636	9.065	676.832***	-6.854***
Δ ln FYJR _L	0.000	0.285	-0.267	0.049	0.118	10.821	1079.106***	-7.019***

注: *, **, *** 分别表示 10%、5% 和 1% 的显著性水平。基于 Hukuhara 差分, 其中 Δ ln SZ_H 表示深证指数区间上界的对数回报, Δ ln SZ_L 则表示深证指数区间下界的对数回报, 其余子市场标示与上述一致。ADF, 修正的 Dickey-Fuller 检验统计量; Jarque-Bera, Jarque-Bera 检验统计量。

(Han et al., 2016)。从而将区间文本情绪指标纳入区间自回归模型 (ACIX) 模型中是合理的。而且, 每个时间序列 Jarque-Bera 检验均显著拒绝原假设, 说明它们均不服从正态分布, 且峰度值均较大, 具有“尖峰厚尾”的特征。

5.1 主板市场预测能力检验

5.1.1 主板市场样本内回归结果

为了考虑主板市场预测能力, 本文构建粉丝加权新闻媒体情绪区间指数对主板市场股票回报的自回归条件区间模型:

$$\Delta Y_t = \alpha_0 + \beta_0 I_0 + \sum_{i=1}^p \beta_i \Delta Y_{t-i} + \delta_1 \text{Sentiment}_{t-1}^{\text{FW}} + u_t, \tag{9}$$

其中, $Y_t = [\ln Y_{L,t}, \ln Y_{H,t}]$, $\ln Y_{L,t}$ 和 $\ln Y_{H,t}$ 分别表示深圳成指每日最低价和最高价的对数, $\Delta Y_t = Y_t - Y_{t-1}$ 为 Hukuhara 差分后的区间序列, 即区间对数收益, $\text{Sentiment}_{t-1}^{\text{FW}}$ 为 $t-1$ 期的粉丝加权新闻媒体情绪区间指数。

作为比较, 本文构建简单平均新闻媒体情绪区间指数 Sentiment 对主板市场股指回报条件区间自回归模型:

$$\Delta Y_t = \alpha_0 + \beta_0 I_0 + \sum_{i=1}^p \beta_i \Delta Y_{t-i} + \delta_1 \text{Sentiment}_{t-1} + u_t, \tag{10}$$

其中, $Sentiment_{t-1}$ 是简单平均新闻情感区间指数的一阶滞后. 根据第 4.3 节的区间 Mallows 准则, 本文选取模型的最优滞后阶数为 $p = 2$.

表 2 给出主板市场样本内的区间数据模型回归结果, 样本内时期为 2014/02/07–2016/12/30. 可以看到, 新闻媒体情绪区间指数对市场回报的影响均不显著. 但是, 从主板市场全样本期的区间数据模型回归结果可以看到, 简单平均新闻媒体情绪区间指数与粉丝加权新闻媒体情绪区间指数对市场回报呈现出显著的负相关性, 表明滞后的新闻媒体情绪对股市回报变动有负向影响. 这与孙淑娜和孙谦 (2018) 研究结论一致, 他们发现投资者情绪与股市收益在短期内呈正效应, 但在一周后出现反转后变为负向影响. 这是因为证券机构公开披露股票信息, 导致股价回归到真实价值附近. 回归结果验证了粉丝加权新闻媒体情绪区间指数的有效性, 能较好地反映新闻媒体情绪对我国股市主板市场的影响.

表 2 主板市场样本内与全样本期的区间数据模型回归结果

(%)	截距	I_0	ΔY_{t-1}	ΔY_{t-2}	$Sentiment_{t-1}$
Panel A: 样本内					
Sentiment ^{FW}	-0.25	-0.14	29.81***	7.13	0.59
Sentiment	-0.23	-0.13	29.78***	7.17	0.56
Panel B: 全样本期					
Sentiment ^{FW}	1.07**	0.37*	28.77***	6.28**	-1.57**
Sentiment	1.01**	0.38*	28.81***	6.35**	-1.61**

注: *, **, *** 分别表示 10%、5% 和 1% 的显著性水平.

5.1.2 主板市场样本外预测结果

本小节采用滚动估计方法研究主板市场样本外回报的预测, 其中, 滚动窗口 $T = 150$. 首个滚动窗口为 2014/02/07–2016/12/30, 样本外预测时间段为 2017/01/06–2022/04/30. 本文构建如下粉丝加权新闻媒体情绪区间指数的主板市场回报预测模型:

$$Sentiment^{FW} : \Delta Y_t = \alpha_0 + \beta_0 I_0 + \sum_{i=1}^p \beta_i \Delta Y_{t-i} + \gamma_1 u_{t-1} + \delta_1 Sentiment_{t-1}^{FW} + u_t. \quad (11)$$

作为比较, 本文也构建基准区间自回归模型 (Base)、简单平均新闻媒体情绪区间指数预测模型 (Sentiment):

$$Base : \Delta Y_t = \alpha_0 + \beta_0 I_0 + \sum_{i=1}^p \beta_i \Delta Y_{t-i} + \gamma_1 u_{t-1} + u_t, \quad (12)$$

$$Sentiment : \Delta Y_t = \alpha_0 + \beta_0 I_0 + \sum_{i=1}^p \beta_i \Delta Y_{t-i} + \gamma_1 u_{t-1} + \delta_1 Sentiment_{t-1} + u_t, \quad (13)$$

根据第 4.3 节的区间 Mallows 准则, 上述模型的最优滞后阶数均为 $p = 2$. 参考 Sun et al. (2018)、He et al. (2021), 本文采用七个预测评价指标来评估不同模型下的预测性能.

$$1) \text{ 平均距离误差 (MDE): } \Omega_{MDE} = \frac{1}{T} \sum_t \sqrt{(|\hat{M}_t - M_t|^2 + |\hat{R}_t - R_t|^2)},$$

- 2) 非覆盖率: $\Omega_c = 1 - \frac{1}{T} \frac{w([L_t, H_t] \cap [\hat{L}_t, \hat{H}_t])}{w([L_t, H_t])}$,
- 3) 无效率: $\Omega_e = 1 - \frac{1}{T} \frac{w([L_t, H_t] \cap [\hat{L}_t, \hat{H}_t])}{w([L_t, H_t])}$,
- 4) 区间下界 L_t 均方根误差 RMSE: $\Omega_L = \sqrt{\frac{1}{T} \sum_t [\hat{L}_t - L_t]^2}$,
- 5) 区间上界 H_t 均方根误差 RMSE: $\Omega_H = \sqrt{\frac{1}{T} \sum_t [\hat{H}_t - H_t]^2}$,
- 6) 区间中点 M_t 均方根误差 RMSE: $\Omega_M = \sqrt{\frac{1}{T} \sum_t [\hat{M}_t - M_t]^2}$,
- 7) 区间范围 R_t 均方根误差 RMSE: $\Omega_R = \sqrt{\frac{1}{T} \sum_t [\hat{R}_t - R_t]^2}$,

其中, T 是数据序列长度, $w(\cdot)$ 是区间宽度, \cap 是集合交集, 均方根误差 RMSE 分别评估区间下界 L_t 、区间上界 H_t 、区间中点 M_t 、区间范围 R_t 的预测精度. 上述七个预测指标值越小, 说明对应模型具有更优异的预测效果.

表 3 给出主板市场样本外预测效果评价. 从表 3 可知, 模型 Sentiment^{FW} 的主板市场样本外预测表现最优. 这表明本文构建的粉丝加权新闻媒体情绪区间指数预测模型 Sentiment^{FW} 显著优于另外两种模型. 简单平均新闻媒体情绪区间指数预测模型比基准模型有更好的预测效果, 验证了新闻媒体情绪的确能影响我国股市回报. 但是, 由于简单平均新闻媒体情绪区间指数构造相对简单, 未能考虑新闻媒体的传播力与影响力, 所以它的预测效果明显劣于粉丝加权新闻媒体情绪区间指数.

图 5 画出不同模型对主板市场的一步滚动预测结果. 由图可知, 三个不同区间模型的预

表 3 主板市场样本外预测效果评价

模型	Ω_{MDE}	Ω_c	Ω_e	Ω_M	Ω_R	Ω_L	Ω_H
Base	0.0204	0.9633	0.9016	0.1039	0.0327	0.1092	0.1011
Sentiment	0.0178	0.9008	0.8019	0.0841	0.0313	0.0873	0.0838
Sentiment ^{FW}	0.0165	0.7670	0.7285	0.0696	0.0306	0.0694	0.0730

注: 指标数值越小表示效果越优异, 粗体表示比较模型中的最小值.

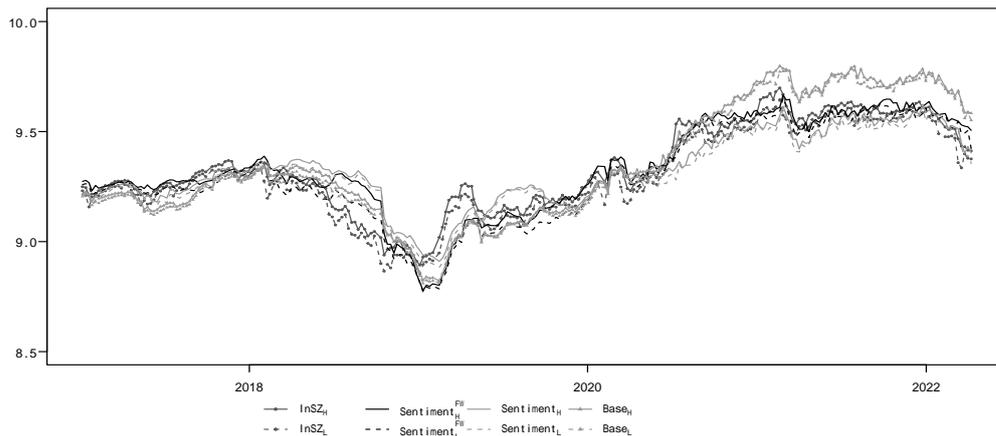


图 5 主板市场指数的一步预测结果

测结果走势与实际价格走势均较为吻合,说明它们对我国股市价格均存在较好的预测能力.这是因为区间数据模型能充分利用区间上界、区间下界和区间极差信息,从而能有很好的预测效果(杨威等,2016).另外,从模型间的预测结果走势来看,粉丝加权新闻媒体情绪区间指数预测模型 $\text{Sentiment}^{\text{FW}}$ 的预测走势与原始价格变化更贴近,具有明显优异的预测效果.尤其是在2020年1月至2022年4月期间,模型 $\text{Sentiment}^{\text{FW}}$ 的预测结果走势与真实走势十分贴近,说明在新冠疫情期间我国股市投资者深受新闻媒体情绪的影响,因而股票市场价格走势受到主流媒体的引导.

为了检验粉丝加权新闻媒体情绪区间指数是否能显著提升模型的预测性能,本文采用修正的 Diebold-Mariano 检验方法(Diebold and Mariano 1995; Harvey et al., 1997)进行显著性检验.该检验方法的基本思路是:假设模型 i 和模型 j 的预测误差分别为 e_{it} 和 e_{jt} ,平方预测误差分别为 $L(\hat{Y}_{it}) = e_{it}^2$, $L(\hat{Y}_{jt}) = e_{jt}^2$, 损失差分序列为 $d_t = L(\hat{Y}_{it}) - L(\hat{Y}_{jt})$. 如果序列 d_t 是协方差平稳且具有短记忆特征,那么修正的 Diebold-Mariano 检验统计量可表示为:

$$[T^{-1}(T+1-2h+T^{-1}h(h-1))]^{\frac{1}{2}} \bar{d} / V_h(\bar{d})^{\frac{1}{2}},$$

其中, $V_h(\bar{d}) = (\gamma_0 + 2 \sum_{k=1}^{h-1} \gamma_k) / T$, $\bar{d} = \sum_{t=1}^T d_t$, γ_k 是损失差分序列 d_t 的 k 阶自协方差, h 为样本外的预测阶数.

考虑到区间型数据的特点,本文分别从区间下界 L_t 、区间上界 H_t 、区间中点 M_t 、区间范围 R_t 四个维度分别进行修正的 Diebold-Mariano 检验.从表4可以看出,对于所有评价指标,修正的 Diebold-Mariano 统计检验大多具有1%以上的显著性水平.这说明粉丝加权新闻媒体情绪区间指数模型 $\text{Sentiment}^{\text{FW}}$ 预测效果在不同评价指标下均显著优于 Base 模型和 Sentiment 模型,证实新闻媒体情绪对中国股市具有引导性,而粉丝加权新闻媒体情绪指数能有效地对我国股市回报进行预测.

表4 主板市场样本外预测效果比较(修正的 Diebold-Mariano 统计检验)

	统计量		统计量
$L_t : \frac{\text{Base}}{\text{Sentiment}^{\text{FW}}}$	8.94***	$M_t : \frac{\text{Base}}{\text{Sentiment}^{\text{FW}}}$	8.68***
$L_t : \frac{\text{Sentiment}}{\text{Sentiment}^{\text{FW}}}$	3.84***	$M_t : \frac{\text{Sentiment}}{\text{Sentiment}^{\text{FW}}}$	3.71***
$H_t : \frac{\text{Base}}{\text{Sentiment}^{\text{FW}}}$	7.64***	$R_t : \frac{\text{Base}}{\text{Sentiment}^{\text{FW}}}$	1.50*
$H_t : \frac{\text{Sentiment}}{\text{Sentiment}^{\text{FW}}}$	3.08***	$R_t : \frac{\text{Sentiment}}{\text{Sentiment}^{\text{FW}}}$	0.50

注: $\frac{A}{B}$ 表示模型 A 与模型 B 预测效果优的比较.修正 Diebold-Mariano 检验原假设 H_0 : 模型 A 的预测效果要优于模型 B. *, **, *** 分别表示在 10%、5% 和 1% 显著性水平显著.

5.2 子板块市场预测能力检验

5.2.1 子板块市场样本内回归结果

为了研究新闻媒体情绪对我国不同子市场回报的预测性,本文参照申万证券一级行业分级标准,依照市场流动性与产业分类,选取医药生物、非银金融、计算机、电子、银行和农林牧渔共6个子行业市场.从产业分类来看,农林牧渔、电子板块分属一、二产业,医药生物板

块在第一、二产业中均有占比, 剩下的非银金融、计算机、银行板块属于第三产业. 从流动性的角度来看, 农林牧渔、电子、银行的流通性相对较弱, 非银金融、计算机、医药生物的流动性较强. 本文后续对 6 个子市场分别建立区间自回归分析, 模型同 (9) 式.

表 5 给出子板块市场样本内的区间数据模型回归结果, 样本期为 2014/02/07–2016/12/30. 表中结果显示, 新闻媒体情绪对各子市场回报的影响并不显著. 但是从子板块市场全样本期的区间数据模型回归结果可以发现, 粉丝加权新闻媒体情绪区间指数对医药生物、证券板块的回报影响显著为负, 而对计算机、电子、银行和农林牧渔市场指数的回归系数并不显著. 这说明医药生物、证券市场回报在全样本期内总体受新闻媒体情绪的负向影响, 因为该些市场的散户投资者较多, 市场情绪主导较明显, 导致股市收益反转现象凸显. 在计算机和电子板块市场, 全样本期内市场回报受新闻情绪影响不显著, 可能原因是市场回报波动较大, 新闻媒体情绪对股市回报的影响受股市结构性变动影响, 导致全样本下影响抵消. 在银行板块, 企业多属于高市值且为国有控股, 受到严格的政策约束, 机构投资者多, 股票价格波动幅度小, 且价格变动受政策影响较大, 所以不容易受到新闻媒体情绪的影响. 另外, 农林牧渔板块的企业大多属于第一产业, 企业市值与产能、经营效益密切相关, 因而难以受到除基本面信息以外的其他市场扰动因素的影响. 因此, 新闻媒体情绪对农林牧渔板块指数的影响不显著.

表 5 分板块子市场样本内与全样本期的区间数据模型回归结果

(%)	医药生物	非银金融	计算机	电子	银行	农林牧渔
Panel A: 样本内						
截距	-0.24	-0.17	-0.92	-1.43	0.10	-0.33
I_0	-0.16	-0.12	-0.38	-0.61	-0.05	-0.23
ΔY_{t-1}	28.08***	40.06***	34.06***	33.80***	21.90***	30.75***
ΔY_{t-2}	11.00***	12.74*	10.02**	4.26	4.35	-2.38
$\text{Sentiment}_{t-1}^{\text{FW}}$	0.64	0.50	1.72	2.51	0.21	1.00
Panel B: 全样本期						
截距	0.93*	1.45*	0.8	0.57	0.49	0.82
I_0	0.29	0.52	0.28	0.15	0.15	0.25
ΔY_{t-1}	32.02***	38.21***	33.00***	35.37***	26.60***	29.85***
ΔY_{t-2}	8.03***	6.43	6.76**	9.62***	5	4.22
$\text{Sentiment}_{t-1}^{\text{FW}}$	-1.32*	-2.19*	-1.19	-0.68	-0.66	-1.1

注: 表中值为回归系数, *, **, *** 分别表示在 10%、5% 和 1% 显著水平下显著.

5.2.2 子板块市场样本外预测结果

限于篇幅, 本小节重点讨论新闻媒体情绪指数对医药生物、证券、计算机和半导体 4 个子行业市场回报的预测性能. 类似地, 本文建立区间自回归模型 Base 和简单平均新闻媒体情绪区间指数预测模型 Sentiment (模型 (12) 与 (13)), 新闻媒体情绪区间指数预测模型 $\text{Sentiment}^{\text{FW}}$ (模型 (11)). 采用滚动估计方法对不同行业市场的样本外收益率进行一步预测. 其中, 滚动窗口与首个滚动窗口等设置同第 5.1.2 节.

表 6 给出子市场的模型预测效果的评价结果. 由表中结果可知, 相比两个基准模型, 粉

丝加权新闻媒体情绪区间指数预测模型 $\text{Sentiment}^{\text{FW}}$ 的所有评价指标更小, 说明粉丝加权新闻媒体情绪区间指数对子市场回报的预测效果更为优异. 图 6 ~ 图 9 描绘出不同子行业的一步滚动预测结果, 直观地证实粉丝加权新闻媒体情绪区间指数预测模型的预测结果更贴近真实价格区间变化. 这说明粉丝加权新闻媒体情绪区间指数对不同子行业市场仍然具有优异的预测性能.

为了检验预测效果, 本文采用修正 Diebold-Mariano 检验进行验证. 由表 7 中结果可知, 相比于模型, 新闻媒体情绪区间指标预测模型的预测效果显著提高, 表明新闻媒体情绪能显著提高子行业市场回报的预测效果. 另外, 相比于简单平均新闻媒体情绪区间预测模型, 粉丝加权的新闻媒体情绪区间模型的预测能力在大多数市场下均有显著提高. 这表明本文所构建的粉丝加权情绪区间指数在我国股市分板块子市场回报预测中的优势十分明显.

表 6 分板块子市场样本外预测效果评价

市场	模型	Ω_{MDE}	Ω_c	Ω_e	Ω_M	Ω_R	Ω_L	Ω_H
医药生物	Base	0.0227	0.9000	0.8796	0.1439	0.0365	0.1458	0.1444
	Sentiment	0.0205	0.8555	0.8312	0.1102	0.0339	0.1101	0.1128
	$\text{Sentiment}^{\text{FW}}$	0.0191	0.8632	0.8308	0.0818	0.0386	0.0888	0.079
非银金融	Base	0.0257	0.8999	0.8732	0.1760	0.0486	0.1875	0.1672
	Sentiment	0.0216	0.6871	0.7731	0.1132	0.0471	0.1040	0.1262
	$\text{Sentiment}^{\text{FW}}$	0.0196	0.7780	0.7759	0.0906	0.0434	0.0910	0.0953
计算机	Base	0.0239	0.8867	0.8632	0.1405	0.0455	0.1277	0.1556
	Sentiment	0.0256	0.9314	0.9185	0.1585	0.0436	0.1516	0.1681
	$\text{Sentiment}^{\text{FW}}$	0.0238	0.9422	0.9062	0.1331	0.0457	0.1218	0.1471
电子	Base	0.0227	0.8840	0.7989	0.1204	0.0483	0.1266	0.1188
	Sentiment	0.0231	0.8786	0.8236	0.1269	0.0487	0.1296	0.1289
	$\text{Sentiment}^{\text{FW}}$	0.0203	0.7605	0.788	0.0966	0.0373	0.0936	0.1030

注: 表中指标数值越小表示预测效果越优异, 粗体表示比较模型中的最小值.

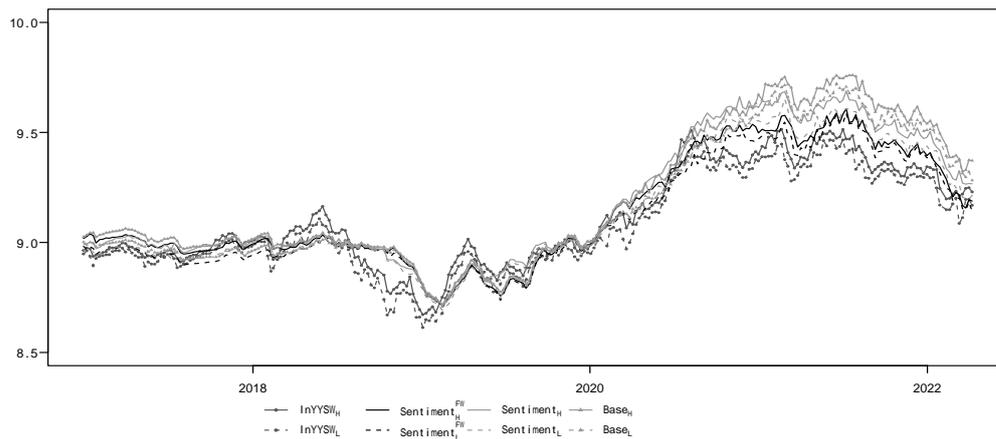


图 6 医药生物行业指数预测结果

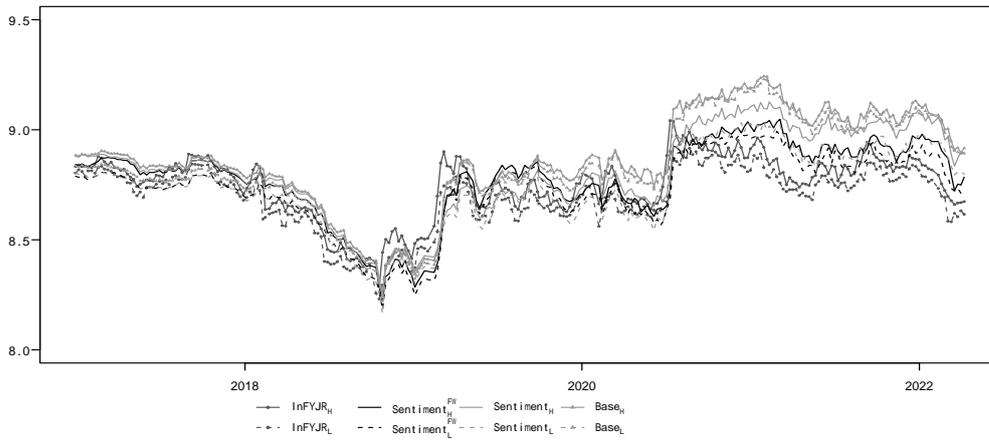


图7 非银行业指数预测结果

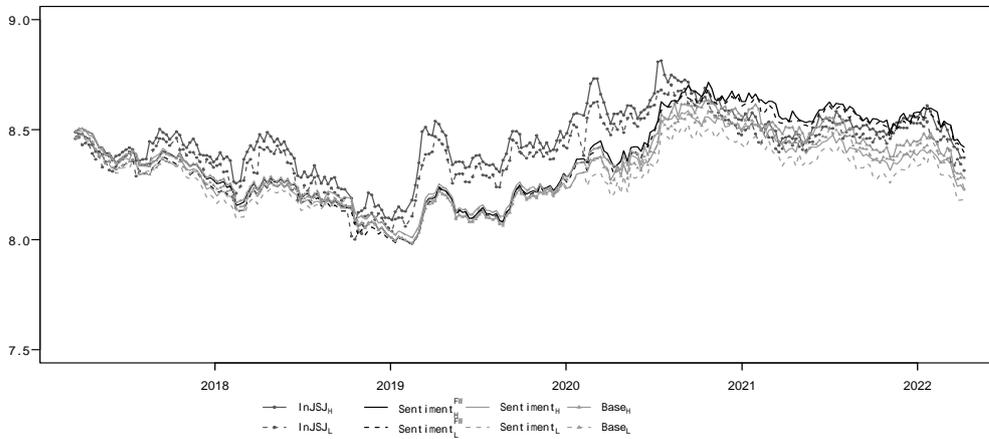


图8 计算机行业指数预测结果

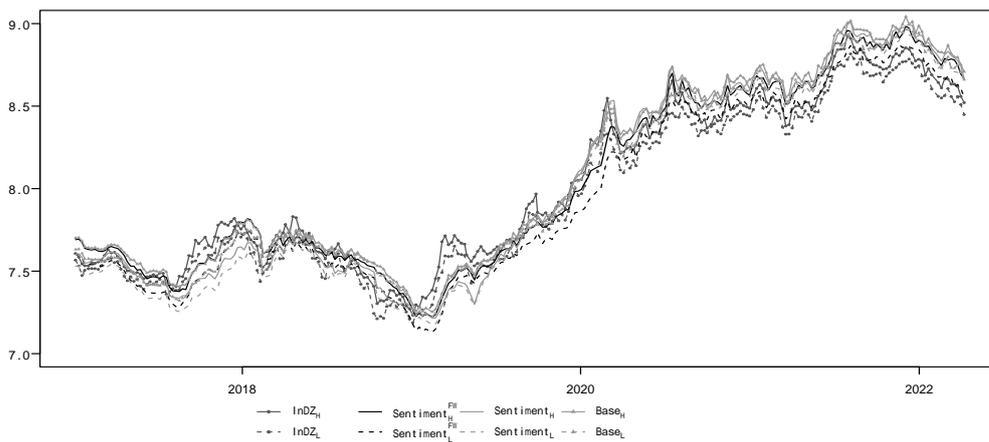


图9 电子行业指数预测结果

表7 分板块市场预测效果比较 (修正的 Diebold-Mariano 统计检验)

	医药生物	非银金融	计算机	电子
$L_t : \frac{\text{Base}}{\text{Sentiment}^{\text{FW}}}$	9.965***	12.145***	2.456***	7.552***
$L_t : \frac{\text{Sentiment}}{\text{Sentiment}^{\text{FW}}}$	7.391***	4.174***	6.948***	9.059***
$H_t : \frac{\text{Base}}{\text{Sentiment}^{\text{FW}}}$	9.711***	10.412***	5.694***	5.298***
$H_t : \frac{\text{Sentiment}}{\text{Sentiment}^{\text{FW}}}$	8.303***	7.264***	7.249***	6.784***
$M_t : \frac{\text{Base}}{\text{Sentiment}^{\text{FW}}}$	9.985***	11.457***	4.042***	6.451***
$M_t : \frac{\text{Sentiment}}{\text{Sentiment}^{\text{FW}}}$	8.330***	6.346***	7.297***	8.047***
$R_t : \frac{\text{Base}}{\text{Sentiment}^{\text{FW}}}$	-2.326	5.162***	-0.197	4.611***
$R_t : \frac{\text{Sentiment}}{\text{Sentiment}^{\text{FW}}}$	-4.019	1.763**	-1.668	4.228***

注: $\frac{A}{B}$ 表示模型 A 与模型 B 预测效果的比较. 修正 Diebold-Mariano 检验原假设 H_0 : 模型 A 的预测效果优于模型 B. *, **, *** 分别表示在 10%, 5%, 1% 显著性水平显著.

5.3 极端事件下模型预测能力

我国股市中个人投资者是非理性行为的主要实施者, 一方面受限于自身的有限注意力与认知偏差, 另一方面有舆论媒体和市场炒作的推波助澜, 容易引发大规模群体性狂热与恐慌, 导致证券市场的“蝴蝶效应”, 推动股价背离其内在价值 (于琴和张兵等, 2020). 当面临经济衰退时, 个人投资者情绪受群体效应的影响, 容易产生非理性行为, 从而对证券市场价格造成显著影响. Huang et al. (2015) 指出, 投资者情绪在经济衰退期对市场回报的预测能力更好. 本小节以新冠疫情事件为例, 研究所构建模型在极端事件下的预测能力. 时间区间选取为 2019/01/30–2021/01/30, 选择 2020/1/23 为时间断点, 滚动窗口 $T = 50$ (一年内的股指交易周数).

表 8~9 给出粉丝加权文本情绪区间模型在新冠疫情前后的预测评价结果与预测效果比较结果. 从结果来看, 在新冠疫情暴发后, 除医药生物板块以外, 本文模型对主板市场、非银金融、计算机、电子板块的预测效果普遍比暴发前更好, 印证了 Huang et al. (2015) 指出的投资者情绪在经济衰退期对市场回报的预测能力更好的结论. 对于医药生物板块, 文本构建的粉丝加权新闻情绪指数区间模型的预测效果在新冠疫情后反而变弱, 说明在新冠疫情期间新闻文本对医药生物市场的影响减弱. 可能原因在于, 医药生物企业在新冠疫情期间是主要的药物供应方, 企业收益受到疫情发展情况的影响, 并且我国政府在疫情期间一直严格管控药物的供给与生产, 对医药生物企业存在直接的政策影响, 因此投资者对医药生物企业的投资决策较少受到新闻媒体情绪的影响.

5.4 预测效果的稳健性检验

为了检验模型 $\text{Sentiment}^{\text{FW}}$ 预测效果的稳健性, 本小节采用 2 步、4 步、6 步三种预测步长 (分别代表短、中、长三种预测周期), 对主板市场、医药生物、非银金融、计算机、电子市场的模型预测结果进行检验. 限于篇幅, 表 10 报告综合性评价指标 Ω_{MDE} 的结果. 由表 10 可知, 在多周期预测的情形下, 除了非银金融与计算机板块的 4 步与 6 步预测结果, 模型

表 8 新冠疫情前后样本外模型预测评价结果

		Ω_{MDE}	Ω_c	Ω_e	Ω_M	Ω_R	Ω_L	Ω_H
主板市场	新冠疫情前	0.0445	0.7332	0.8363	0.1004	0.0308	0.0965	0.1064
	新冠疫情后	0.0374	0.8831	0.726	0.0494	0.0354	0.0554	0.0494
医药生物	新冠疫情前	0.0394	0.9497	0.699	0.0578	0.0369	0.0496	0.07
	新冠疫情后	0.0445	0.7951	0.7982	0.0855	0.0372	0.0892	0.0857
非银金融	新冠疫情前	0.0503	0.908	0.8283	0.1091	0.0519	0.1065	0.1175
	新冠疫情后	0.0456	0.8701	0.6929	0.0808	0.0579	0.0833	0.0882
计算机	新冠疫情前	0.0719	1	1	0.2152	0.0583	0.1886	0.2425
	新冠疫情后	0.0564	0.9318	0.8945	0.1399	0.0529	0.1267	0.1564
电子	新冠疫情前	0.0531	0.8619	0.8488	0.1311	0.0402	0.1286	0.1365
	新冠疫情后	0.0447	0.7878	0.7977	0.0899	0.0275	0.0856	0.096

注: 表中指标数值越小表示预测效果越优异, 粗体表示比较模型中的最小值。

表 9 新冠疫情前后市场预测效果比较 (修正的 Diebold-Mariano 统计检验)

	主板市场	医药生物	非银金融	计算机	电子
L_t : 新冠疫情前/新冠疫情后	3.647***	-4.083	1.788**	6.627***	3.544***
H_t : 新冠疫情前/新冠疫情后	3.057***	-1.587	1.509*	8.203***	2.471***
M_t : 新冠疫情前/新冠疫情后	3.365***	-2.881	1.839**	7.865***	2.967***
R_t : 新冠疫情前/新冠疫情后	-1.297	-0.054	-0.714	0.817	2.668***

注: 新冠疫情前/新冠疫情后表示断点前 (2019/01/30-2020/01/23) 与断点后 (2020/01/30-2021/01/30) 预测效果的比较. 修正 Diebold-Mariano 检验原假设 H_0 : 模型 A 的预测效果优于模型 B. *, **, *** 分别表示在 10%、5% 和 1% 显著性水平显著。

Sentiment^{FW} 的预测效果在不同市场中表现最优. 这说明在多步预测下, 本文所构建的粉丝加权新闻媒体情绪区间指数仍然具有优异的预测性能。

运用修正的 Diebold-Mariano 检验, 表 11 给出模型 Sentiment^{FW} 多步预测效果比较结果. 从表 11 可以发现:

1) 在主板市场, 1 步预测的效果均显著优于多步预测效果, 而且随着步长的增加, 新闻媒体情绪指数的预测性能下降. 这说明新闻媒体情绪主要在短期对我国主板股市有较好的预测性能, 但它的影响随着时间逐渐减弱, 反映出我国股市总体节奏快, 能快速地消化新闻情绪等信息, 股市整体的有效性较高。

2) 受行业特征及股市有效性的影响, 新闻媒体情绪对不同子市场的影响存在差异性. 例如, 在医药生物市场, 1 步预测相比于多步预测的 Diebold-Mariano 检验结果均显著, 但 6 步预测结果依次优于 4 步预测与 2 步预测结果. 这说明该板块对新闻媒体情绪能及时反应, 但是在短期影响减弱而在长期影响逐渐加强. 而在非银金融市场, 新闻媒体情绪指数的最优预测步长为 2 步, 且其他步长预测效果均弱于 1 步预测结果. 这说明新闻媒体情绪对证券市场的影响集中在中短期, 而且我国非银金融市场对来自于新闻媒体情绪的反应存在迟滞性. 在计算机板块市场, 最优预测步长为 6 步, 但多步预测结果均优于 1 步预测, 而且预测效果随着步长的增加而提高. 这说明我国计算机市场受到新闻媒体情绪影响会逐渐增强. 电子板块市场的最优预测步长为 1 步, 说明新闻情绪对电子板块市场的影响主要体现在前期. 这可能

表 10 主板市场与子市场多步预测结果评价

		2 步预测	4 步预测	6 步预测
主板市场	Base	0.0206	0.0282	0.0314
	Sentiment	0.0204	0.0254	0.0288
	Sentiment ^{FW}	0.0196	0.0209	0.0259
医药生物	Base	0.0242	0.0226	0.0205
	Sentiment	0.0257	0.0259	0.0337
	Sentiment ^{FW}	0.0200	0.0223	0.0192
非银金融	Base	0.0191	0.0240	0.0221
	Sentiment	0.0251	0.0204	0.0269
	Sentiment ^{FW}	0.0185	0.0222	0.0226
计算机	Base	0.0242	0.0221	0.0217
	Sentiment	0.0273	0.0220	0.0238
	Sentiment ^{FW}	0.0235	0.0233	0.0230
电子	Base	0.0268	0.0283	0.0456
	Sentiment	0.0370	0.0383	0.0468
	Sentiment ^{FW}	0.0259	0.0263	0.0404

是因为该板块市场投资者大多属于短期投资或散户投资者,投资者会根据新闻媒体信息迅速调整自身投资组合,从而影响市场股价。

6 结论

投资者情绪是影响股市价格变动的重要因素,如何合理测度投资者情绪并对股市回报进行预测是研究者关注的重点问题。本文从文本挖掘新闻媒体情绪的角度出发,构建新的加权新闻媒体情绪指数,并应用区间数据条件自回归模型,研究新闻媒体情绪对我国股市回报的可预测性能。首先,本文在中国国务院网站爬取金融相关新闻文本,基于金融情感词典分析文本情绪,并运用粉丝加权方法构建加权新闻媒体情绪区间指数;其次,本文构建区间数据预测模型,研究粉丝加权新闻情感区间指数对我国股市主板市场、细分子市场回报的预测能力,并与基准模型、简单平均新闻情感指数模型进行对比;最后,本文讨论了粉丝加权新闻媒体情绪区间指数对我国股市回报的最优预测步长。

基于实证结果,可得如下结论: 1) 无论在主板市场还是分板块子市场中,本文构建的粉丝加权新闻情绪区间指数的预测性能显著优于简单平均新闻情绪指数与区间自回归两个基准模型,粉丝加权新闻情绪区间指数能够提升预测精度。进一步说明我国股市回报明显受到新闻媒体情绪的影响,而且与新闻的影响力息息相关。新闻媒体的粉丝越多,影响力越高,从而新闻文本情绪对股市回报的影响越明显。2) 在分板块子市场中,新闻媒体情绪在样本内对各子市场回报的影响并不显著。但在全样本期内,粉丝加权新闻媒体情绪指数仅对医药生物、非银金融两个板块子市场回报显著的负影响,而对其余子市场回报的影响并不显著。这反映出医药生物与非银金融市场的散户投资者较多,容易受到市场情绪的引导,市场并非完全有效,从而易于遭受市场风险。因此,在金融市场上需要严厉的监管,谨防外部风险传导致使市场整体风险的增加。3) 新闻媒体情绪对主板市场股市回报的预测效果随着预测步长的增加

表 11 Sentiment^{FW} 模型多步预测效果比较 (修正的 Diebold-Mariano 统计检验)

	主板市场	医药生物	非银金融	计算机	电子
L_t : 2 步预测/1 步预测	6.14***	4.47***	-5.24	1.36*	2.06**
H_t : 2 步预测/1 步预测	6.47***	6.36***	-4.36	0.29	-0.47
M_t : 2 步预测/1 步预测	6.64***	4.85***	-1.4	0	12.08***
R_t : 2 步预测/1 步预测	1.02	-4.92	2.33**	-2.2	1.44*
L_t : 4 步预测/1 步预测	8.86***	6.66***	8.94***	-0.83	0.23
H_t : 4 步预测/1 步预测	7.20***	8.88***	1.93**	-3.4	1.13
M_t : 4 步预测/1 步预测	8.30***	6.88***	5.03***	-3.7	12.01***
R_t : 4 步预测/1 步预测	2.28**	-2.67	0.6	-0.71	5.56***
L_t : 6 步预测/1 步预测	10.53***	1.61*	3.59***	-0.39	10.15***
H_t : 6 步预测/1 步预测	11.19***	2.70***	3.75***	-3.54	8.72***
M_t : 6 步预测/1 步预测	10.85***	0.86	3.54***	-3.71	14.65***
R_t : 6 步预测/1 步预测	12.03***	-0.59	3.64***	-4.29	5.43***
L_t : 4 步预测/2 步预测	6.45***	7.92***	12.94***	-2.39	-1.71
H_t : 4 步预测/2 步预测	1.98**	11.24***	7.22***	-5.58	2.11**
M_t : 4 步预测/2 步预测	4.45***	7.76***	7.83***	-4.67	8.68***
R_t : 4 步预测/2 步预测	2.06**	2.39***	-2.91	2.98	4.61***
L_t : 6 步预测/2 步预测	8.64***	-5.69	7.55***	-1.62	9.18***
H_t : 6 步预测/2 步预测	9.09***	-8.63	12.14***	-5.04	8.21***
M_t : 6 步预测/2 步预测	8.77***	-8.44	5.99***	-3.94	-3.47
R_t : 4 步预测/2 步预测	10.83***	5.33***	2.56***	-2.05	4.52***
L_t : 6 步预测/4 步预测	10.01***	-12.23	-1.44	0.89	9.74***
H_t : 6 步预测/4 步预测	11.08***	-13.54	1.81**	-1.28	9.01***
M_t : 6 步预测/4 步预测	10.72***	-12.7	-1.12	-0.31	-5.17
R_t : 6 步预测/4 步预测	10.05***	3.17***	3.42***	-4.41	-0.63
最优步长	1	1	2	6	1

注: $\frac{A}{B}$ 表示模型 A 与模型 B 预测效果的比较. 修正 Diebold-Mariano 检验原假设 H_0 : 模型 A 的预测效果优于模型 B. *, **, *** 分别表示在 10%、5% 和 1% 显著性水平显著.

而减弱, 最优预测步长为 1 步. 但是, 新闻媒体情绪指数对不同板块子市场的股市回报预测效果具有异质性. 医药生物与电子板块市场在短期内受到新闻媒体情绪影响 (最优预测步长均为 1 步), 非银金融行业则在中短期内被新闻媒体情绪影响 (最优预测步长为 2 步), 而计算机板块在中长期内 (最优预测步长分别为 6 步) 能被新闻媒体情绪指数较好地预测, 反映出它们的市场有效性不足.

本文编制的粉丝加权新闻媒体情绪区间指数, 为新闻文本情绪指标的构建提供了新思路. 而且, 本文实证研究新闻媒体情绪对我国股市具有显著影响, 新闻文本情绪能有效地预测股市回报. 最后, 本文利用区间数据模型对我国股市回报进行预测研究, 为区间数据模型的应用提供新的预测实践. 实证结果不仅有助于为金融投资者和政策制定者提供决策参考, 还为我国股市的风险防控提供启示意义.

参 考 文 献

- 部慧, 解峥, 李佳鸿, 吴俊杰, (2018). 基于股评的投资者情绪对股票市场的影响 [J]. 管理科学学报, 21(4): 86-101.
- Bu H, Xie Z, Li J H, Wu J J, (2018). Investor Sentiment Extracted from Internet Stock Message Boards and Its Effect on Chinese Stock Market[J]. Journal of Management Sciences in China, 21(4): 86-101.
- 陈炜, 徐慧琳, 汪寿阳, 孙少龙, (2023). 基于误差修正与分解的区间值股价时间序列预测研究 [J]. 系统工程理论与实践, 43(2): 383-397.
- Chen W, Xu H L, Wang S Y, Sun S L, (2023). Error Correction and Decomposition Method for Forecast of Interval-valued Stock Price Time Series[J]. Systems Engineering — Theory & Practice, 43(2): 383-397.
- 池丽旭, 张广胜, 庄新田, 宋大雷, (2012). 投资者情绪指标与股票市场——基于扩展卡尔曼滤波方法的研究[J]. 管理工程学报, 26(3): 122-128.
- Chi L X, Zhang G S, Zhuang X T, Song D L, (2012). A Study on Investor Sentiment Index and Chinese Stock Market Based on the Extended Kalman Filter Method[J]. Journal of Industrial Engineering and Engineering Management, 26(3): 122-128.
- 范小云, 王业东, 王道平, 郭文璇, 胡煊翊, (2022). 不同来源金融文本信息含量的异质性分析——基于混合式文本情绪测度方法 [J]. 管理世界, 38(10): 78-101.
- Fan X Y, Wang Y D, Wang D P, Guo W X, Hu X Y, (2022). Heterogeneity Analysis of Information Content for Financial Text from Different Sources: A Hybrid Text Sentiment Measurement Method[J]. Management World, 38(10): 78-101.
- 顾洪梅, 张嫚玲, (2022). 新闻情绪风险与股票收益 [J]. 中央财经大学学报, (7): 37-47.
- Gu H M, Zhang M L, (2022). News Sentiment Beta and Stock Returns[J]. Journal of Central University of Finance & Economics, (7): 37-47.
- 黄润鹏, 左文明, 毕凌燕, (2015). 基于微博情绪信息的股票市场预测 [J]. 管理工程学报, 29(1): 47-52.
- Huang R P, Zuo W M, Bi L Y, (2015). Predicting the Stock Market Based on Microblog Mood[J]. Journal of Industrial Engineering and Engineering Management, 29(1): 47-52.
- 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, 21(4): 1323-1344.
- 金雪军, 祝宇, 杨晓兰, (2013). 网络媒体对股票市场的影响——以东方财富网股吧为例的实证研究 [J]. 新闻与传播研究, 20(12): 36-51.
- Jin X J, Zhu Y, Yang X L, (2013). Effects of Online Media on Stock Market: An Empirical Study on Eastmoney. Com[J]. Journalism & Communication, 20(12): 36-51.
- 林昱, 常晋源, 黄雁勇, (2022). 融合经验模态分解与深度时序模型的股价预测 [J]. 系统工程理论与实践, 42(6): 1663-1677.
- Lin Y, Chang J Y, Huang Y Y, (2022). On the Prediction of the Stock Price Based on Empirical Mode Decomposition and Deep Time Series Model[J]. Systems Engineering — Theory & Practice, 42(6): 1663-1677.
- 刘维奇, 刘新新, (2014). 个人和机构投资者情绪与股票收益——基于上证 A 股市场的研究 [J]. 管理科学学报, 17(3): 70-87.
- Liu W Q, Liu X X, (2014). Individual/Institutional Investor Sentiment and Stock Returns: Study

- Based on Shanghai A-share Market[J]. *Journal of Management Sciences in China*, 17(3): 70–87.
- 刘志峰, 张子沅, 戴鹏飞, 刘文华, (2023). 碳市场与股票市场间的崩盘风险溢出效应研究: 新冠疫情、投资者情绪与经济政策不确定性 [J]. *系统工程理论与实践*, 43(3): 740–754.
- Liu Z F, Zhang Z F, Dai P F, Liu W H, (2023). A Study on the Spillover Effect of Crash Risk Between Carbon and Stock Markets: COVID-19, Investor Sentiment and Economic Policy Uncertainty[J]. *Systems Engineering — Theory & Practice*, 43(3): 740–754.
- 卢米雪, (2022). 投资者情绪的测量及其对股市波动率的影响效应研究 [J]. *宏观经济研究*, (9): 106–119.
- Lu M X, (2022). The Measurement of Investor Sentiment and Its Effect on Stock Market Volatility[J]. *Scientific Management Research*, (9): 106–119.
- 陆沁晔, 陈昊, (2021). 媒体报道、投资者情绪与股价波动 [J]. *财经问题研究*, (3): 60–67.
- Lu Q Y, Chen H, (2021). Media Coverage, Investor Sentiment and Stock Price Volatility[J]. *Research on Financial and Economic Issues*, (3): 60–67.
- 鹿坪, 冷军, (2017). 投资者情绪与盈余管理——基于应计盈余管理与真实盈余管理的实证研究 [J]. *财经问题研究*, (2): 88–96.
- Lu P, Leng J, (2017). Investor Sentiment and Surplus Management — An Empirical Study Based on Accrual Surplus Management and Real Surplus Management[J]. *Research on Financial and Economic Issues*, (2): 88–96.
- 吕华揆, 刘政昊, 钱宇星, 洪旭东, (2021). 异质性财经新闻与股市关系研究 [J]. *数据分析与知识发现*, 5(1): 99–111.
- Lü H K, Liu Z H, Qian Y X, Hong X D, (2021). Relationship Between Financial News and Stock Market Fluctuations[J]. *Data Analysis and Knowledge Discovery*, 5(1): 99–111.
- 钱宇, 李子饶, 李强, 袁华, (2020). 在线社区支持倾向对股市收益和波动的影响 [J]. *管理科学学报*, 23(2): 141–155.
- Qian Y, Li Z R, Li Q, Yuan H, (2020). Impact of Online Community Support Tendencies on Returns and Volatility in Chinese Stock Market[J]. *Journal of Management Sciences in China*, 23(2): 141–155.
- 石善冲, 朱颖楠, 赵志刚, 康凯立, 熊熊, (2018). 基于微信文本挖掘的投资者情绪与股票市场表现 [J]. *系统工程理论与实践*, 38(6): 1404–1412.
- Shi S C, Zhu Y N, Zhao Z G, Kang K L, Xiong X, (2018). The Investor Sentiment Mined from WeChat Text and Stock Market Performance[J]. *Systems Engineering — Theory & Practice*, 38(6): 1404–1412.
- 孙少龙, 魏云捷, 黎建强, (2022). 基于在线外汇新闻情感挖掘的汇率预测研究 [J]. *计量经济学报*, 2(2): 441–464.
- Sun S L, Wei Y J, Lai K K, (2022). Exchange Rate Forecasting with Online Forex News Sentiment Mining[J]. *China Journal of Econometrics*, 2(2): 441–464.
- 孙彦林, 陈守东, 刘洋, (2019). 基于股市和汇市成交量信息视角的股价波动预测 [J]. *系统工程理论与实践*, 39(4): 935–945.
- Sun Y L, Chen S D, Liu Y, (2019). Forecast of Stock Price Fluctuation Based on the Perspective of Volume Information in Stock and Foreign Exchange Market[J]. *Systems Engineering — Theory & Practice*, 39(4): 935–945.
- 孙淑娜, 孙谦, (2018). 投资者关注和股市表现——基于雪球关注度的研究 [J]. *管理科学学报*, 21(6): 60–71.
- Sun S N, Sun Q, (2018). Investor Attention and Market Performance: Evidence Based on “Xueqiu Attention”[J]. *Journal of Management Sciences in China*, 21(6): 60–71.
- 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*, (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 Finance Research*, (3): 180–193.
- 唐国豪, 姜富伟, 张定胜, (2016). 金融市场文本情绪研究进展 [J]. *经济学动态*, (11): 137–147.
- Tang G H, Jiang F W, Zhang D S, (2016). *Advances in Text Sentiment Research in Financial Markets*[J]. *Economic Perspectives*, (11): 137–147.
- 王靖一, 黄益平, (2018). 金融科技媒体情绪的刻画与对网贷市场的影响 [J]. *经济学 (季刊)*, 17(4): 1623–1650.
- Wang J Y, Huang Y P, (2018). *Characterizing the Media Sentiment on FinTech and Its Impact on Online Lending Market*[J]. *China Economic Quarterly*, 17(4): 1623–1650.
- 王美今, 孙建军, (2004). 中国股市收益、收益波动与投资者情绪 [J]. *经济研究*, (10): 75–83.
- Wang M J, Sun J J, (2004). *Stock Market Returns Volatility and the Role of Investor Sentiment in China*[J]. *Economic Research Journal*, (10): 75–83.
- 武静, 赵二龙, 孙少龙, 汪寿阳, (2023). 多源异构数据驱动的后疫情时期旅游需求预测方法研究 [J]. *计量经济学报*, 3(2): 350–366.
- Wu J, Zhao E L, Sun S L, Wang S Y, (2023). *Multi-source Heterogeneous Data-driven Tourism Demand Forecasting Approach amid Post-COVID-19 Era*[J]. *China Journal of Econometrics*, 3(2): 350–366.
- 许雪晨, 田侃, (2021). 一种基于金融文本情感分析的股票指数预测新方法 [J]. *数量经济技术经济研究*, 38(12): 124–145.
- Xu X C, Tian K, (2021). *A Novel Financial Text Sentiment Analysis-Based Approach for Stock Index Prediction*[J]. *The Journal of Quantitative & Technical Economics*, 38(12): 124–145.
- 杨威, 韩艾, 汪寿阳, (2016). 基于区间型数据的金融时间序列预测研究 [J]. *系统工程学报*, 31(6): 816–830.
- Yang W, Han A, Wang S Y, (2016). *Forecasting Research of Financial Time Series Based on Interval Data*[J]. *Journal of Systems Engineering*, 31(6): 816–830.
- 姚加权, 冯绪, 王赞钧, 季荣嵘, 张维, (2021). 语调、情绪及市场影响: 基于金融情绪词典 [J]. *管理科学学报*, 24(5): 26–46.
- Yao J Q, Feng X, Wang Z J, Ji R R, Zhang W, (2021). *Tone, Sentiment and Market Impacts: The Construction of Chinese Sentiment Dictionary in Finance*[J]. *Journal of Management Sciences in China*, 24(5): 26–46.
- 姚加权, 张锟澎, 罗平, (2020). 金融学文本大数据挖掘方法与研究进展 [J]. *经济学动态*, (4): 143–158.
- Yao J Q, Zhang K P, Luo P, (2020). *Text Mining in Financial Big Data and Its Research Progress*[J]. *Economic Perspectives*, (4): 143–158.
- 易志高, 茅宁, (2009). 中国股市投资者情绪测量研究: CICSI 的构建 [J]. *金融研究*, (11): 174–184.
- Yi Z G, Mao N, (2009). *A Study on Measuring Investor Sentiment in Chinese Stock Market: Construction of CICSI*[J]. *Journal of Finance Research*, (11): 174–184.
- 于琴, 张兵, (2020). 股市的媒体强化效应: “强者恒强”还是“盛极而衰”[J]. *山西财经大学学报*, 42(6): 45–58.
- Yu Q, Zhang B, (2020). *Media Reinforcement Effect of Stock Market: “Remaining Strong” or “Declining after Prosperity”*[J]. *Journal of Shanxi University of Finance and Economics*, 42(6): 45–58.
- 于志军, 杨善林, (2013). 基于误差校正的 GARCH 股票价格预测模型 [J]. *中国管理科学*, 21(S1): 341–345.
- Yu Z J, Yang S L, (2013). *A Model for Stock Price Forecasting Based on Error Correction*[J]. *Chinese Journal of Management Science*, 21(S1): 341–345.
- 余佩琨, 钟瑞军, (2009). 个人投资者情绪能预测市场收益率吗 [J]. *南开管理评论*, 12(1): 96–101.
- Yu P K, Zhong R J, (2009). *Can Individual Investor Sentiment Predict Market Returns?*[J]. *Nankai*

- Business Review, 12(1): 96–101.
- 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 G Y, (2021). Media's Emotional Contagion and Analyst Optimistic Bias: Evidence Based on the Technique of Machine Learning[J]. Management World, 37(1): 170–185.
- 周文凯, 杨威, (2020). 基于区间型金融时间序列数据的宏观经济预测研究 [J]. 经济问题, (3): 35–41.
- Zhou W K, Yang W, (2020). Research on Macroeconomic Forecasting Based on Interval-valued Financial Time Series Data[J]. On Economic Problems, (3): 35–41.
- 周战强, (2004). 行为金融: 理论与应用 [M]. 北京: 清华大学出版社.
- Zhou Z Q, (2004). Behavioral Finance: Theory and Applications[M]. Beijing: Tsinghua University Press.
- 朱伟骅, 张宗新, (2008). 投资者情绪、市场波动与股市泡沫 [J]. 经济理论与经济管理, (2): 45–50.
- Zhu W H, Zhang Z X, (2008). Investor Sentiment Market Volatility and Stock Market Bubble[J]. Economic Theory and Business Management, (2): 45–50.
- Antweiler W, Frank M Z, (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards[J]. The Journal of Finance, 59: 1259–1294.
- Arroyo J, Espínola R, Maté C, (2011). Different Approaches to Forecast Interval Time Series: A Comparison in Finance[J]. Computational Economics, 37: 169–191.
- Baker M, Wurgler J, (2006). Investor Sentiment and the Cross-section of Stock Returns[J]. The Journal of Finance, 61: 1645–1680.
- Barber B M, Odean T, (2008). All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors[J]. Review of Financial Studies, 21: 785–818.
- Billard L, Diday E, (2000). Regression Analysis for Interval-valued Data[M]// Data Analysis, Classification and Related Methods. Belgium: Springer.
- Billard L, Diday E, (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis[J]. Journal of the American Statistical Association, 98: 470–487.
- Blanco-Fernández A, Corral N, González-Rodríguez G, (2011). Estimation of a Flexible Simple Linear Model for Interval Data Based on Set Arithmetic[J]. Computational Statistics & Data Analysis, 55: 2568–2578.
- Bollen J, Mao H, Zeng X, (2010). Twitter Mood Predicts the Stock Market[J]. Journal of Computer Science, 2: 1–8.
- Chan S W K, Chong M W C, (2017). Sentiment Analysis in Financial Texts[J]. Decision Support Systems, 94: 53–64.
- Diebold F X, Mariano R S, (1995). Comparing Predictive Accuracy[J]. Journal of Business & Economic Statistics, 13: 253–263.
- Fama E F, (1970). Efficient Capital Markets: A Review of Theory and Empirical Work[J]. The Journal of Finance, 25: 383–417.
- Feldman R, Govindaraj S, Livnat J, Segal B, (2010). Management's Tone Change, Post Earnings Announcement Drift and Accruals[J]. Review of Accounting Studies, 15: 915–953.
- Gil M Á, González-Rodríguez G, Colubi A, Montenegro M, (2007). Testing Linear Independence in Linear Models with Interval-valued Data[J]. Computational Statistics & Data Analysis, 51: 3002–3015.
- Han A, Hong Y, Wang S, Yun X, (2016). A Vector Autoregressive Moving Average Model for Interval-valued Time Series Data[J]. Essays in Honor of Aman Ullah (Advances in Econometrics), 36: 417–460.

- Han A, Lai K K, Wang S, Xu S, (2012). An Interval Method for Studying the Relationship Between the Australian Dollar Exchange Rate and the Gold Price[J]. *Journal of Systems Science and Complexity*, 25: 121–132.
- Harvey D, Leybourne S, Newbold P, (1997). Testing The Equality of Prediction Mean Squared Errors[J]. *International Journal of Forecasting*, 13: 281–291.
- He L T, Hu C, Casey K M, (2009). Prediction of Variability in Mortgage Rates: Interval Computing Solutions[J]. *The Journal of Risk Finance*, 10: 142–154.
- He Y, Han A, Hong Y, Sun Y, Wang S, (2021). Forecasting Crude Oil Price Intervals and Return Volatility Via Autoregressive Conditional Interval Models[J]. *Econometric Reviews*, 40: 584–606.
- Hu C, (2010). A Note on Probabilistic Confidence of the Stock Market ILS Interval Forecasts[J]. *The Journal of Risk Finance*, 11: 410–415.
- Huang D, Jiang F, Tu J, Zhou G, (2015). Investor Sentiment Aligned: A Powerful Predictor of Stock Returns[J]. *Review of Financial Studies*, 28: 791–837.
- Kim S H, Kim D, (2014). Investor Sentiment from Internet Message Postings and the Predictability of Stock Returns[J]. *Journal of Economic Behavior & Organization*, 107: 708–729.
- Li F, (2008). Annual Report Readability, Current Earnings, and Earnings Persistence[J]. *Journal of Accounting and Economics*, 45: 221–247.
- Liu Q, Cheng X, Su S, Zhu S, (2018). Hierarchical Complementary Attention Network for Predicting Stock Price Movements with News[C]// Torino Italy: ACM.
- Loughran T, McDonald B, (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks[J]. *The Journal of Finance*, 66: 35–65.
- Moore R E, (1966). *Interval Analysis*[M]. Englewood Cliffs: Prentice-Hall.
- Neto E A L, De Carvalho F D A T, (2008). Centre and Range Method for Fitting a Linear Regression Model to Symbolic Interval Data[J]. *Computational Statistics & Data Analysis*, 52: 1500–1515.
- Neto E A L, De Carvalho F D A T, (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables[J]. *Computational Statistics & Data Analysis*, 54: 333–347.
- Oliveira N, Cortez P, Areal N, (2017). The Impact of Microblogging Data for Stock Market Prediction: Using Twitter to Predict Returns, Volatility, Trading Volume and Survey Sentiment Indices[J]. *Expert Systems with Applications*, 73: 125–144.
- Price S M, Doran J S, Peterson D R, Bliss B A, (2012). Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone[J]. *Journal of Banking & Finance*, 36: 992–1011.
- Qiao K, Sun Y, Wang S, (2019). Market Inefficiencies Associated with Pricing Oil Stocks During Shocks[J]. *Energy Economics*, 81: 661–671.
- Shefrin H, (2002). *Beyond Greed and Fear*[M]. New York: Oxford University Press.
- Sun Y, Bao Q, Zheng J, Wang S, (2020). Assessing the Price Dynamics of Onshore and Offshore RMB Markets: An ITS Model Approach[J]. *China Economic Review*, 62: 101476.
- Sun Y, Han A, Hong Y, Wang S, (2018). Threshold Autoregressive Models for Interval-valued Time Series Data[J]. *Journal of Econometrics*, 206: 414–446.
- Yang W, Han A, (2015). A New Approach for Forecasting the Price Range with Financial Interval-valued Time Series Data[J]. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems Part B: Mechanical Engineering*, 1(2): 021004.
- Yang W, Han A, Hong Y, Wang S, (2016). Analysis of Crisis Impact on Crude Oil Prices: A New Approach with Interval Time Series Modelling[J]. *Quantitative Finance*, 16: 1917–1928.