



# 投资者文本情绪、文本 异质信念与市场崩盘风险

李少育<sup>1</sup>, 朱春辉<sup>2</sup>, 宁凯旋<sup>3</sup>, 于志豪<sup>3</sup>

1 华南师范大学 国际商学院, 广州 510631

2 华南师范大学 马克思主义学院, 广州 510631

3 西南财经大学 证券与期货学院, 成都 611130

**摘要:** 区别于个股崩盘风险, 在投资者情绪和异质信念急骤变化的情景下, 市场崩盘风险更具有毁灭性, 防范市场崩盘风险已成为政府和监管部门关注的热点问题。然而, 当前学者重点关注个股崩盘风险及其影响因素, 对投资者情绪和异质信念影响市场崩盘风险的研究较为匮乏。

利用2015年1月1日至2019年8月27日东方财富网上证综指股吧评论大数据, 通过机器学习优化方法, 提取针对上证市场走势的投资者情绪和异质信念直接指标, 即投资者文本情绪和文本异质信念。检验投资者情绪和异质信念直接指标与间接指标的相关性, 分析投资者文本情绪和文本异质信念的相对优越性。检验投资者情绪和异质信念直接指标与市场惯性、波动率、流动性等市场微观结构变量的紧密关系, 论证投资者文本情绪和文本异质信念对市场崩盘风险的影响和内在机制。

研究表明, ①投资者文本情绪和文本异质信念比传统间接指标具有更多的私有信息和更好的市场拟合度。②投资者文本情绪与累计收益率正相关, 与收益率波动率和非流动性负相关; 投资者文本异质信念与累计收益率负相关, 与收益率波动率和非流动性正相关。③投资者文本情绪和文本异质信念利用微观结构变量的变化作用于交易型开放指数基金, 进而影响市场崩盘风险。当投资者文本情绪低且文本异质信念高时, 市场惯性效应变弱, 波动率和流动性成本升高, 投资者增加交易型开放指数基金的赎回行为, 促使基金为持有更多现金而被动卖出大量市场权重股, 向下崩盘风险的可能性上升。反之, 向上崩盘风险的可能性上升。

研究结果丰富了关于市场崩盘风险影响因素和机制的研究, 抑制羊群行为和过度波动、提升市场流动性是管控投资者情绪和异质信念的重要途径, 加强对主动型机构投资者的培育, 限制被动型基金的交易规模 and 市场份额, 引入基于交易型开放指数基金的金融衍生品, 以切断发生市场崩盘风险的机制, 对防范市场系统性风险具有重要的政策启示。

**关键词:** 投资者文本情绪; 投资者文本异质信念; 上证综指; 市场崩盘风险; 机器学习

**中图分类号:** F830.9

**文献标识码:** A

**doi:** 10.3969/j.issn.1672-0334.2023.04.011

**文章编号:** 1672-0334(2023)04-0147-17

**收稿日期:** 2021-05-13 **修返日期:** 2022-05-07

**基金项目:** 国家自然科学基金(72271099); 教育部人文社会科学研究项目(22YJC790062, 22YJC790191)

**作者简介:** 李少育, 经济学博士, 华南师范大学国际商学院教授, 研究方向为金融风险、衍生品定价、资本市场和金融大数据分析等, 代表性学术成果为“Sentiment-aware jump forecasting”, 发表在2021年第7期《Knowledge-Based Systems》, E-mail: [syli@ibc.scnu.edu.cn](mailto:syli@ibc.scnu.edu.cn)

朱春辉, 经济学博士, 华南师范大学马克思主义学院讲师, 研究方向为数字经济、绿色与可持续发展和国有企业改革等, 代表性学术成果为“Government-led urbanization and natural gas demand in China”, 发表在2021年第9期《Renewable & Sustainable Energy Reviews》, E-mail: [zhuchunhui@m.scnu.edu.cn](mailto:zhuchunhui@m.scnu.edu.cn)

宁凯旋, 西南财经大学证券与期货学院硕士研究生, 研究方向为大数据分析和金融风险管理等, E-mail: [kaixuan\\_ning@126.com](mailto:kaixuan_ning@126.com)

于志豪, 西南财经大学证券与期货学院硕士研究生, 研究方向为金融风险管理等, E-mail: [a347044107@163.com](mailto:a347044107@163.com)

## 引言

市场崩盘风险是指股市整体发生极端暴涨暴跌的系统性风险,资本市场系统性风险的影响因素及机制是资本市场监管和宏观调控关注的焦点问题,防范化解金融风险特别是防止发生系统性金融风险,是金融工作的根本性任务。投资者情绪和异质信念与市场崩盘风险紧密相关,例如,受到新冠疫情的影响,投资者出现恐慌情绪,美国股市在2020年3月9日、12日和16日触发3次熔断,标普500指数单日内下跌幅度均超过7%,演化为一场罕见的金融危机。因此,基于情绪驱动的市场崩盘风险应当引起高度重视。

近年来,崩盘风险是学术研究的热点问题,重点在于揭示个股崩盘风险或与个股相关的崩盘风险的来源,主要包括管理层自利、机构投资者监督、监管层限薪政策、投资者网络搜索能力、卖空限制、管理层语境和个股语调情绪等。然而,鲜有研究针对更具毁灭性的股票市场崩盘风险,探讨投资者情绪和异质信念影响市场崩盘风险的研究也较为匮乏。

本研究利用个体投资者针对市场总体趋势的评论文本数据,结合机器学习方法和文本分析技术,构建投资者文本情绪和文本异质信念作为投资者情绪和异质信念的直接指标,揭示投资者文本情绪和文本异质信念利用微观结构变量的变化作用于交易型开放指数基金(exchange traded fund, ETF)进而影响市场崩盘风险的机制,把投资机构影响崩盘风险的机制进一步拓展到考察ETF行为层面。

## 1 相关研究述评

### 1.1 投资者文本情绪和文本异质信念

国内外学术界对投资者情绪和异质信念的定义还未完全统一,缺乏精确的测量指标,已有研究沿用间接指标测量投资者情绪和异质信念。最近的研究正尝试采用数据挖掘技术,从投资者网络文本中提取更为客观、高频和精准的投资者文本情绪和文本异质信念测量投资者情绪和异质信念。

已有研究多以市场指标作为测量投资者情绪或异质信念的间接指标。LEE et al.<sup>[1]</sup>以封闭式基金折价率测量投资者情绪;WURGLER et al.<sup>[2]</sup>和HUANG et al.<sup>[3]</sup>利用封闭式基金折价率、市场交易量等6个市场指标测量投资者情绪,并采用主成分分析法或最小二乘法提取投资者情绪。

近年来许多研究开始利用互联网上的投资者评价数据信息,通过文本分析构建投资者情绪和异质信念直接指标。ANTWEILER et al.<sup>[4]</sup>针对网络媒体信息进行文本分析,并构建网络媒体情绪;董大勇等<sup>[5]</sup>和LI et al.<sup>[6]</sup>利用东方财富网股吧数据进行文本分析,测量投资者情绪和异质信念。针对情绪和异质信念的文本分析方法包括以下两类:

(1) 词典方法。先对句子进行分词,然后统计各个词汇的个数,最后在情感字典中查找这些单词对应的情感值,计算出总体的情感和异质信念。尹海

员等<sup>[7]</sup>以“金融界”网站股票帖为样本,以Python的类库SnowNLP中自带的情感语料库作为词典,利用词库匹配法对所有样本进行情感分类,分类准确率为83.200%;JIANG et al.<sup>[8]</sup>利用美国上市公司财报、电视和电话会议文本,基于金融字典的文本分类建立了经理人情绪指数;姚加权等<sup>[9]</sup>结合深度学习算法和词典建构金融领域中文情绪词典,并合成上市公司的年报语调和社交媒体情绪指标,发现能有效预测上市公司股票价格特征和风险。

(2) 机器学习方法。通过输入部分语句及其情感标签,训练出一个语句情感分类器,从而预测新语句的情感。杨晓兰等<sup>[10]</sup>以文本挖掘工具Weka的KNN算法将东方财富网的投资者评论分为积极、消极和中立3类,准确率为72%;LI et al.<sup>[11]</sup>结合中文金融词典、机器学习和深度学习的方法,对市场指数股吧的评论信息进行文本分析,获得市场汇总的投资者文本情绪和文本异质信念,研究发现单纯使用词典法的精度只有74%,而结合机器学习或者深度学习的准确率可达到81%;唐国豪等<sup>[12]</sup>的研究表明,由于股吧评论中语言表达较为丰富,社交语句较多,机器学习对情感分析更为精准。

已有研究主要基于个股的评论文本,鲜有从市场层面的文本分析出发提取个体投资者针对市场总体趋势的投资者文本情绪和文本异质信念。因此,针对市场总体的投资者评论文本,利用机器学习方法进行文本分析,有助于研究市场层面的投资者情绪和异质信念问题,也有利于深入探讨其对资本市场的总体影响。

### 1.2 崩盘风险

对崩盘风险的研究重点分析影响个股崩盘风险的因素,主要从信息不对称和投资者行为两个角度展开。在信息不对称方面,HUTTON et al.<sup>[13]</sup>和赵璨等<sup>[14]</sup>认为管理层在信息披露过程中降低公司透明度;夏常源等<sup>[15]</sup>发现股票市场缺少机构投资者监督增加了个股崩盘风险;BAI et al.<sup>[16]</sup>的研究表明监管层的限薪政策促进了个股崩盘风险。

在投资者行为方面,XU et al.<sup>[17]</sup>发现投资者网络搜索能力降低会引起个股崩盘风险的上升;DENG et al.<sup>[18-19]</sup>的研究表明,投资者受到卖空限制能显著影响个股崩盘风险;杨七中等<sup>[20]</sup>认为投资者受管理层语境的影响将改变个股崩盘风险。

然而,已有针对市场崩盘风险来源的研究还未有效开展,针对投资者情绪和异质信念影响市场崩盘风险的研究较为稀少。

### 1.3 投资者文本情绪和文本异质信念对市场微观结构和市场崩盘风险的影响

投资者文本情绪和文本异质信念会深刻改变惯性效应、波动率和流动性等市场微观结构变量。段江娇等<sup>[21]</sup>的研究表明,文本情绪能够正向影响股价的未来收益率和波动率,文本异质信念对交易量有正向影响;林红新等<sup>[22]</sup>发现中国股市的投资者乐观情绪会促进反转效应的产生,悲观情绪会抵制反转

效应;唐也然等<sup>[23]</sup>认为中国股市的惯性效应相当一部分来自于投资者情绪,大部分投资者持乐观态度时,其他投资者会盲目跟风。已有研究聚焦于针对个股层面的投资者文本情绪和文本异质信念与市场微观结构的联系,未进一步探讨市场微观结构变化如何作用于市场崩盘风险。

投资者情绪和异质信念影响证券市场定价的稳定性和投资者的交易行为,引起市场定价的系统性偏离和市场崩盘风险。DE LONG et al.<sup>[24]</sup>认为噪音交易者的非理性情绪会造成资本市场价格的系统性风险;陈国进等<sup>[25]</sup>的研究表明,投资者异质信念对股价向下崩盘风险有正向关系,投资者异质信念越大,崩盘风险越高;孙鲲鹏等<sup>[26]</sup>发现投资者情绪越乐观,未来的崩盘风险会越严重。已有研究较为关注投资者情绪和异质信念对股票向下崩盘风险的作用,却鲜有从市场层面揭示投资者情绪和异质信念对向上崩盘风险的影响。

投资者文本情绪和文本异质信念可通过多种途径影响个股的崩盘风险。尹海员等<sup>[7]</sup>认为噪音交易是投资者情绪影响股票收益的重要驱动力;孙鲲鹏等<sup>[26]</sup>的研究表明,机构投资者持股加剧了乐观情绪在股吧的传播,进而加大了崩盘风险;关静怡等<sup>[27]</sup>发现股吧评论分歧会通过证券分析师跟踪人数和研究报告数量影响股价的崩盘风险。已有研究鲜有从总体市场层面上分析投资者文本情绪和文本异质信念影响市场崩盘风险的机制。同时,中国资本市场的买空卖空限制条件已经逐步放松,投资者的乐观和悲观私人信息很容易通过融资融券、股指期货和股指期货交易得到迅速释放。因此,投资者文本情绪上涨(趋向乐观)和文本异质信念上升对向下崩盘风险的作用方向还需要深入探讨。

#### 1.4 评述

综上所述,已有研究还有不完善之处。本研究认为分析向上的市场崩盘风险也非常重要,中国资本市场允许融资融券、股指期货、股指期货等双向交易,市场股指的骤然大幅度上升对采取做空的投资者来说也是一种崩盘风险。因此,本研究将使用多种监督式机器学习方法,针对市场总体走势获得更高分类精确度的投资者文本情绪和文本异质信念指标,揭示投资者文本情绪和文本异质信念如何通过市场微观结构的变化和ETF买卖机制影响市场崩盘风险;通过多种机器学习的文本分析方法,利用上证综指股吧评论构建针对股票整体市场的投资者文本情绪和文本异质信念指标,并应用于研究它们对市场崩盘风险(向上和向下)的效应及机制。

## 2 理论分析和研究假设

### 2.1 投资者文本情绪和文本异质信念与市场微观结构的关系

近年来,行为金融学的投资者有限关注理论成为研究投资者文本情绪和文本异质信念与市场微观结构关系的重要理论基础。由于投资者的注意力有限,

投资者倾向于买入近期被广泛关注的股票。因此,当投资者文本情绪趋向乐观,大量投资者在同一时间内以同方向交易同一支股票,出现羊群行为,从而惯性效应得到显著增强。当投资者文本情绪较为悲观时,反转效应出现的可能性增加。当投资者的消极情绪变得更浓厚时,投资者参与股票交易的热情变低,市场的成交量显著下滑,市场流动性变得较小。然而,当投资者对市场的分歧(异质信念)较大时,股票的交易变得频繁,市场波动率和流动性增大,惯性效应持续的可能性变小。因此,本研究提出假设。

H<sub>1</sub> 投资者文本情绪与市场的惯性效应和流动性正相关,投资者文本异质信念与市场波动率和流动性正相关,但与市场的惯性效应负相关。

### 2.2 投资者文本情绪和文本异质信念反映市场微观结构的变化并影响市场崩盘风险

行为金融学中的噪音交易者理论表明,噪音交易者的非理性情绪是资本市场发生系统性风险的重要来源。丁慧等<sup>[28]</sup>的研究表明,当投资者异质信念较小时,市场崩盘风险变低。当投资者的意见分歧较小而投资者乐观情绪高涨时,波动率被抑制,市场的上涨惯性将会持续下去,此时向下的崩盘风险较小,而向上的崩盘风险较高。当投资者情绪较低而异质信念较高时,悲观投资者占据上风,波动率上升而流动性较低,向下的崩盘风险骤增。综上,投资者文本情绪和文本异质信念的综合效应支配着市场崩盘风险的走向。因此,本研究提出假设。

H<sub>2</sub> 投资者文本情绪趋向乐观而文本异质信念较低,则向下的市场崩盘风险变小,向上的市场崩盘风险变大;投资者文本情绪趋向悲观而文本异质信念较高,则向下的市场崩盘风险变大,向上的市场崩盘风险变小。

### 2.3 投资者文本情绪和文本异质信念影响市场崩盘风险的机制

对于市场崩盘风险的影响机制分析,现有理论通常从噪音交易者行为的路径出发,还有研究发现机构投资者的羊群行为对股票市场崩盘风险具有显著作用(机构投资者行为路径)。当前,股票市场指数的最大跟踪者是机构投资者中的特殊类别,即ETF。个体投资者会通过交易ETF的持有份额,促进ETF更加频繁地交易换手,或者ETF通过调仓来改变现金持有水平,从而影响大盘指数。基于这一路径分析,个体投资者的文本情绪和文本异质信念很可能通过个体投资者的噪音交易(即改变ETF的换手率)和调整对ETF的投资权重(即影响ETF的现金持有)作用于总体的股票市场崩盘风险。因此,本研究提出假设。

H<sub>3</sub> 投资者文本情绪趋向乐观而文本异质信念较低,则ETF的换手率上升,向下的市场崩盘风险变小,向上的市场崩盘风险变大;投资者文本情绪趋向悲观而文本异质信念较高,则ETF的换手率下降,向下的市场崩盘风险变大,向上的市场崩盘风险变小。

H<sub>4</sub> 投资者文本情绪趋向乐观而文本异质信念较



低,则ETF的现金持有减少,向下的市场崩盘风险变小,向上的市场崩盘风险变大;投资者文本情绪趋向悲观而文本异质信念较高,则ETF的现金持有增加,向下的市场崩盘风险变大,向上的市场崩盘风险变小。

以上理论分析给出了投资者文本情绪和文本异质信念影响市场崩盘风险的机制,即投资者文本情绪或文本异质信念体现了市场微观结构的变化,促进ETF的变动,进而改变市场崩盘风险发生的可能性。下文将依据该逻辑实证检验投资者文本情绪和文本异质信念对市场崩盘风险的影响效应和机制。

### 3 崩盘风险、投资者文本情绪和文本异质信念的指标构建与分析

#### 3.1 数据来源

本研究尝试从市场指数股吧评论中提取文本信息来测量表征个体投资者总体意见并驱动市场交易的投资者文本情绪和文本异质信念。根据LI et al.<sup>[6]</sup>的研究,中国股市个体投资者持有股票比例约占总市值的20%,却贡献了80%的市场总体交易量;LIU et al.<sup>[29]</sup>发现大量上市公司是国有控股,代表政府的股东很少频繁交易股票,而个体投资者持有88%的市场自由流通股。鉴于股评数据的来源用户众多,而机构投资者数量有限,绝大部分参与评论的用户是中小投资者,因此,基于市场指数股吧的投资者评论很可能包含个体投资者对市场的总体情绪信息。根据艾丽丝和艾瑞专业网站的数据,东方财富网在中国金融互联网平台中排名第一,日均点击量超过1亿次,并且具有较为丰富和连续的股评文本数据,特别是市场指数股吧中的上证指数吧。因此,本研究爬取了2015年1月1日至2019年8月27日东方财富网上证指数吧的投资者评论数据,共约397万条股评文本。同时,本研究从国泰安数据库获取上证综指的数据,以上证综指的崩盘风险测量市场崩盘风险,样本区间从2015年1月1日到2019年8月30日,共1115个交易日频率和56个月度频率的数据。为规避新冠疫情影响,并考虑到数据的获得情况,将样本截止时间确定为2019年8月底。在稳健性检验时,本研究考虑了疫情期间的延长样本数据。

#### 3.2 市场崩盘风险

本研究采用两类方法构建市场崩盘风险指标,分别为MARIN et al.<sup>[30]</sup>的极值测量方法以及KIM et al.<sup>[31]</sup>的负收益偏态系数和收益上下波动率比率的方法。第1类方法的优点是能够测量向下和向上的崩盘风险,并被主要用于实证分析,分别构建月度和日度频率的极值测量指标。第2类方法能够测量日度频率的负收益偏态系数和收益上下波动率比率,用于实证分析结果的稳健性检验。

(1) MARIN et al.<sup>[30]</sup>的极值指标。本研究利用上证综指收益率的极值测量市场崩盘风险,从月度和日度频率的收益率序列中分别计算月度和日度的极值指标。因上证综指是市场指数,所以利用市场超额

收益率的负极值和正极值分别表示系统性向下和向上崩盘风险。本研究设定二元变量刻画市场是否崩盘,即

$$Fcd_t = \begin{cases} 1, & r_t - \bar{r}_t \leq -L\delta_t \\ 0, & \text{其他} \end{cases} \quad (1)$$

$$Fcu_t = \begin{cases} 1, & r_t - \bar{r}_t \geq L\delta_t \\ 0, & \text{其他} \end{cases} \quad (2)$$

其中, $t$ 为月; $Fcd_t$ 为 $t$ 月的向下市场崩盘风险; $Fcu_t$ 为 $t$ 月的向上市场崩盘风险; $r_t$ 为 $t$ 月的上证综指超额收益率, $r_t = r_{m,t} - r_{f,t}$ , $r_{m,t}$ 为 $t$ 月上证综指收益率(不计股利复利), $r_{f,t}$ 为 $t$ 月无风险收益率; $\bar{r}_t$ 为 $r_t$ 在 $(t-12)$ 月至 $(t-1)$ 月之间的均值; $\delta_t$ 为 $r_t$ 在 $(t-12)$ 月至 $(t-1)$ 月之间的标准差; $L$ 为崩盘事件临界阈值中标准差的倍数,参考陈国进等<sup>[25]</sup>的研究,分别取值为1.650、2和2.250。(1)式和(2)式中,当 $r_t$ 变小,并低于 $\bar{r}_t$ 等于 $L\delta_t$ 时,出现月度向下崩盘;当 $r_t$ 变大,并高于 $\bar{r}_t$ 等于 $L\delta_t$ 时,出现月度向上崩盘。为便于下文实证分析, $L = 1.650$ 时 $Fcd_t = DF_1$ , $Fcu_t = UF_1$ ; $L = 2$ , $Fcd_t = DF_2$ , $Fcu_t = UF_2$ ; $L = 2.250$ , $Fcd_t = DF_3$ , $Fcu_t = UF_3$ 。日度频率的极值指标采用类似的方法, $i$ 为日, $Fcd_i$ 为 $i$ 日的向下市场崩盘风险, $Fcu_i$ 为 $i$ 日的向上市场崩盘风险。

(2) KIM et al.<sup>[31]</sup>的负收益偏态系数和收益上下波动率比率指标。本研究利用上证综指日度收益率滚动计算季度窗口的负收益偏态系数( $Ncw$ )和收益上下波动率比率( $Dul$ ),构建向下崩盘风险的日度频率,即采用 $i$ 日至 $(i+59)$ 日共60天内上证指数日度收益率计算负收益偏态系数和收益上下波动率,该值越大,崩盘风险越大,这两个指标为季度窗口滚动的日频率数据。

### 3.3 投资者文本情绪和文本异质信念

#### 3.3.1 投资者股评文本的情感分析

计算投资者文本情绪和文本异质信念需要先对股评文本进行情感分析。本研究在Python语言环境下调用Request和BeautifulSoup库进行动态网页爬虫,从上证指数吧详情页获取的内容包括每条帖子的标题和具体时间。部慧等<sup>[32]</sup>发现,中文股评语法繁复,适用的词典极为缺乏,所以本研究采用金融领域最常用的基于分类器建模的机器学习方法,以更有效地实现对庞大文本信息的分类。运用监督式机器学习方法构建情感分析分类模型,实现对股评文本数据的情感分类,图1给出文本分析的整个过程,包括文本清洗和预处理、创建训练样本、训练机器学习算法、使用最佳算法预测分类4个基本步骤。

(1) 文本清洗和预处理。该步骤对抓取的397万条股评文本进行清洗和预处理,转换为特征向量后传入机器学习模型。首先清洗数据,初步筛选去除重复贴和空白行等,以Python语言的Jieba库进行分词,再去除如“的”“个”等停用词。然后预处理,将所有原始的语料文本转为特征向量形式,构建机器学习算法需要的数据格式。

(2) 创建训练样本。从清洗后的全部股评文本中随机抽取5000条股评,首先,手动分类为积极评论

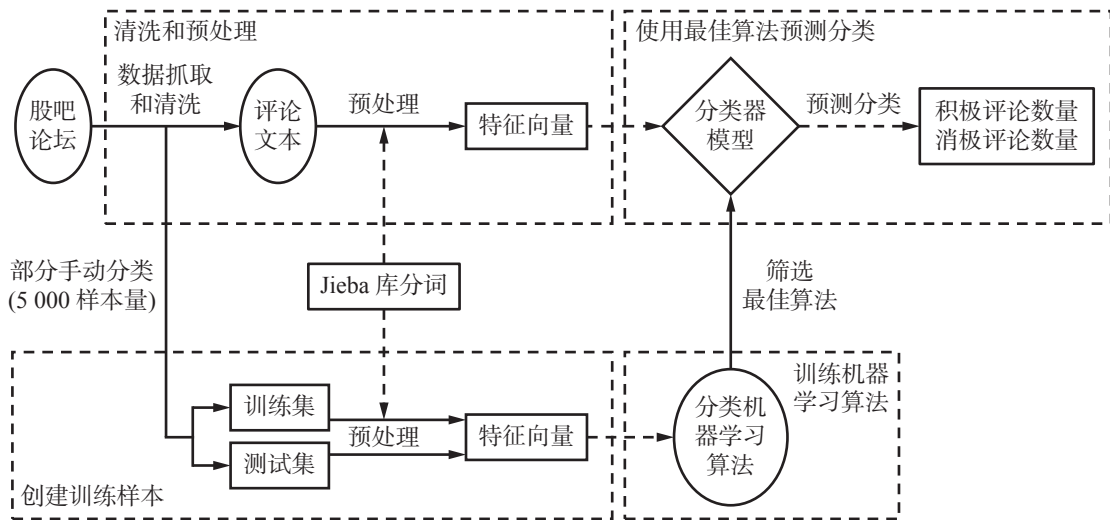


图1 机器学习与投资者评论文本的分类

Figure 1 Machine Learning and Classification of Investor Review Texts

表1 各种机器学习分类器的效果

Table 1 Effects of Various Machine Learning Classifiers

模型	准确度/%	积极组 精确度/%	消极组 精确度/%	积极组 召回率/%	消极组 召回率/%	积极组 F1/%	消极组 F1/%
伯努利贝叶斯	88.506	93.128	86.221	76.904	96.230	84.217	90.908
多项式贝叶斯	89.045	92.011	87.419	79.343	95.412	85.214	91.206
逻辑回归	86.710	90.809	84.715	74.218	95.009	81.708	89.603
线性支持向量机	87.605	89.707	86.608	78.105	94.023	83.526	90.111
非线性支持向量机	86.815	87.308	86.505	78.418	92.414	82.605	89.409
决策树	82.503	81.914	82.811	72.109	89.410	76.703	86.004
随机森林	82.409	82.724	82.212	70.615	90.204	76.217	86.033

组和消极评论组,用于训练机器学习模型。其次,对已分类的样本进行随机化排列,划分为4份样本集,每份1250条股评,其中3份作为模型训练样本集,1份作为测试样本集。再次,在已分类的5000条股评中统计每个词的信息量,因一个词可能同时出现在积极评论组和消极评论组,如“高点”“逆势”等,所以,分别统计每个词的积极和消极卡方统计量,并加总得到每个词的整体信息量,以测量该词的信息丰富程度。最后,选取前800个信息量高的关键词用于机器学习模型的降维。

(3) 训练机器学习算法。首先,根据NLTK中Scikit-Learn接口,导入7种分类机器学习算法,分别为伯努利贝叶斯、多项式贝叶斯、逻辑回归、支持向量机(线性、非线性)、决策树和随机森林分类器。其次,将800个关键词和已转换为特征向量的3份训练样本集导入模型算法进行训练,并以1份测试样本集用于算法准确度检验。对于每个分类机器学习算法,根据准确度、精确度、召回率和F1得分4个指标进行评估。实践中通常认为F1得分是综合反映整体的

指标,优于其他指标,因此,当拟合每个单独的分类器模型时,通过选择参数以最大化F1分数。

(4) 使用最佳算法预测分类,得到积极评论数量和消极评论数量。通过F1分数筛选最佳分类器模型,表1给出本研究应用于样本的各种机器学习分类器的实际效果。基于多项式贝叶斯估计分类器在各项指标上综合表现最好,准确度达89.045%,积极组F1得分为85.214%,消极组F1得分为91.206%。因此,本研究选择此分类器进行全部文本的情感分类,模型的结果输出为投资者积极评论数量和消极评论数量。

### 3.3.2 投资者文本情绪和文本异质信念

本研究先用日度积极评论数量和消极评论数量测量日度文本情绪和文本异质信念,使用ANTWEILER et al.<sup>[4]</sup>的股吧帖子分类指标计算日度文本情绪,该值越大表明投资者情绪越高涨;采用日度积极评论数量与消极评论数量差值的绝对值除以总的有效评论数量计算日度文本异质信念,该值越大表明投资者异质信念越大。具体计算公式为

$$BI_i = \ln \frac{1 + Pos_i}{1 + Neg_i} \quad (3)$$

$$HB_i = \frac{|Pos_i - Neg_i|}{Pos_i + Neg_i} \quad (4)$$

其中,  $BI_i$  为  $i$  日投资者文本情绪,  $HB_i$  为  $i$  日投资者文本异质信念,  $Pos_i$  为  $i$  日投资者积极评论数量,  $Neg_i$  为  $i$  日投资者消极评论数量。

由于单日的情绪或单日的异质信念指标并不能准确代表投资者一段时间的情绪和异质信念, 所以本研究基于移动平均法从日度数据构建月度投资者文本情绪和文本异质信念。计算过程如下:

第1步, 计算月窗口滚动的日度文本情绪和文本异质信念, 通过移动平均向后 20 个交易日数据获得,  $Bim_i$  为采用算数平均法计算的日度投资者文本情绪,  $Hbm_i$  为采用算数平均法计算的日度投资者文本异质信念,  $Biw_i$  为采用加权平均法计算的日度投资者文本情绪,  $Hbw_i$  为采用加权平均法计算的日度投资者文本异质信念。

第2步, 根据第1步日度指标, 提取每个月的月末数据, 形成月度文本情绪和文本异质信念,  $Bim_i$  为基于算数平均法计算的日度数据获取的月度投资者文本情绪,  $Hbm_i$  为基于算数平均法计算的日度数据获取的月度投资者文本异质信念,  $Biw_i$  为基于加权平均法计算的日度数据获取的月度投资者文本情绪,  $Hbw_i$  为基于加权平均法计算的日度数据获取的月度投资者文本异质信念。

#### 3.4 投资者情绪和异质信念的间接指标、市场微观结构和ETF变动

为了探讨投资者文本情绪和文本异质信念的动态特征, 分析其与投资者情绪和异质信念间接指标的关系和优势, 本研究引入测量投资者情绪的间接代理变量, 包括借鉴魏星集等<sup>[33]</sup>的研究构建的投资者情绪指数 ( $Isi$ )、借鉴易志高等<sup>[34]</sup>的研究构建的综合情绪指数 ( $Cic$ ); 引入测量投资者异质信念的间接代理变量, 参考陈国进等<sup>[35]</sup>的研究利用上证综指构建月度换手率 ( $Tur$ )。

为进一步分析投资者文本情绪和文本异质信念对市场微观结构的影响, 本研究采用惯性效应 ( $CR$ )、市场波动率 ( $Sig$ )、流动性 ( $Ami$ ) 表征市场微观结构。具体地, 用过去 12 个月的市场累积收益率测量惯性效应, 用过去 12 个月收益率的历史波动率测量市场波动率, 用市场收益率绝对值除以成交金额计算的非流动性测量流动性, 非流动性越大, 流动性越小。

本研究尝试分别从ETF的噪声交易和投资权重探讨投资者文本情绪和文本异质信念影响市场崩盘风险的作用机制, 以ETF的换手率 ( $Tuf$ ) 测量噪声交易, 以ETF的现金占净比 ( $Cas$ ) 测量ETF调整持股比例后的现金持有水平 (即投资权重)。其中, 换手率为投资者对ETF的噪声交易行为, 换手率越高则噪声交易行为越多; 现金占净比为基金投资组合中现金占总市值的比重, 现金持有为ETF持股行为, 当基金持有更大比例的股票时, 现金占净比将缩小。另外,

为了控制宏观周期因素, 本研究还引入以年份 ( $Yea$ ) 构建的哑变量代理年度控制变量。

#### 3.5 描述性统计分析

本研究探讨投资者文本情绪和文本异质信念的动态特征, 分析其与投资者情绪和异质信念间接指标的关系。表2给出投资者文本情绪和文本异质信念及相关变量的描述性统计结果,  $Bim_i$  与  $Biw_i$ 、 $Hbm_i$  与  $Hbw_i$  的差别较小, 平均看, 投资者对股市的情绪是负面的, 但是从偏度看, 投资者在总体负面的情况下更偏向乐观, 大部分情况下投资者异质信念较小。投资者文本情绪与投资者情绪间接指标  $Isi_i$  和  $Cic_i$  在偏度和峰度上存在显著的差异, 反映了不同的投资者情绪信息。投资者文本异质信念与投资者异质信念间接指标  $Tur$  也存在显著的差异, 表明投资者文本异质信念指标可能提供了额外的异质信念信息。 $UF$  的均值比  $DF$  的均值大, 表明向上崩盘风险发生的可能性比向下崩盘风险发生的可能性大。

进一步分析投资者文本情绪和文本异质信念与股票市场的关联和动态趋势, 图2给出  $Biw_i$  和  $Hbw_i$  与上证综指月收益率 ( $Ret$ ) 和月度崩盘风险的走势。由图2的 (a) 和 (b) 可知,  $Biw_i$  高涨伴随着后续的低潮,  $Hbw_i$  也有类似的动态特征。对  $Biw_i$  和  $Hbw_i$  分别进行一阶自回归, 估计结果表明, 两者均存在显著的均值回复系数, 均值回复速度均约为 0.580,  $Biw_i$  的长期均值为 -0.860,  $Hbw_i$  的长期均值为 0.390, 表明投资者文本情绪和文本异质信念的动态走势存在均值回复现象。同时还发现, 上证综指月收益率与  $Biw_i$  的走势方向大致相同, 与  $Hbw_i$  的走势正好相反, 意味着投资者文本情绪可能与收益率存在正相关关系, 投资者文本异质信念与收益率存在负相关关系。以上分析表明, 投资者文本情绪和文本异质信念符合市场的直觉和理论预期。同样的, 由图2的 (c) 和 (d) 可知, 在  $Biw_i$  下降并进入低谷、 $Hbw_i$  上升并进入高峰时, 向下崩盘风险就会出现, 这也符合理论预期。由图2的 (e) 和 (f) 可知, 在  $Biw_i$  上升并进入高峰、 $Hbw_i$  下降并进入低谷时, 向上崩盘风险就会出现。图2的描述性分析为后继的实证分析提供了经验参考。

#### 3.6 投资者文本情绪和文本异质信念的优越性

投资者文本情绪和文本异质信念指标能够更准确地解释投资者情绪和投资者异质信念对股票市场的影响, 本研究利用线性回归的拟合优度大小检验投资者文本情绪和文本异质信念对股票市场收益率解释的优越性。表3给出投资者文本情绪和文本异质信念对上证综指收益率的回归结果, 投资者文本情绪和文本异质信念的回归系数比间接指标的回归系数大得多, 并且与  $Ret$  的相关性更为显著;  $Biw_i$  和  $Hbw_i$  的回归结果类似。说明研究市场崩盘风险时使用投资者文本情绪和文本异质信念更恰当。

为了进一步说明基于机器学习的投资者文本情绪和文本异质信念与间接指标的联系和区别, 本研究利用  $Isi_i$  和  $Cic_i$  分别对月度投资者文本情绪指标进行回归, 利用  $Tur_i$  对月度投资者文本异质信念进行回



表2 描述性统计结果  
Table 2 Results for Descriptive Statistics

	均值	中位数	标准差	最小值	最大值	偏度	峰度
$Bim_t$	-0.860	-0.855	0.126	-1.093	-0.526	0.283	2.658
$Biw_t$	-0.870	-0.888	0.129	-1.079	-0.519	0.442	2.798
$Hbm_t$	0.402	0.401	0.052	0.257	0.494	-0.374	2.774
$Hbw_t$	0.405	0.414	0.054	0.253	0.491	-0.548	2.970
$DF_1$	0.339	0	0.478	0	1	0.679	1.461
$DF_2$	0.304	0	0.464	0	1	0.854	1.730
$DF_3$	0.286	0	0.456	0	1	0.949	1.900
$UF_1$	0.393	0	0.493	0	1	0.439	1.193
$UF_2$	0.357	0	0.483	0	1	0.596	1.356
$UF_3$	0.321	0	0.471	0	1	0.765	1.585
$Isi_t$	64.124	57.490	21.070	47.080	137.890	2.528	8.802
$Cic_t$	42.974	42.830	4.368	33.300	51.920	-0.318	2.641
$Tur_t$	0.145	0.115	0.103	0.045	0.498	1.837	5.865
$CR_t$	0.087	0.032	0.291	-0.391	0.870	0.979	3.598
$Sig_t$	0.062	0.060	0.032	0.019	0.121	0.434	1.881
$Ami_t$	0.407	0.330	0.256	0.097	1.526	2.111	8.869

注:样本量为56。

归,分析投资者情绪和异质信念间接指标解释投资者情绪和异质信念直接指标的程度,以及投资者情绪和异质信念直接指标提供私有信息的比重,回归结果见表4。

由表4可知,  $Isi_t$  和  $Cic_t$  与投资者文本情绪均在1%水平上显著正相关,  $R^2$  分别为0.163、0.136、0.109和0.112,表明投资者文本情绪具有间接指标所没有的私有信息,而且这部分未解释的文本情绪占比超过80%。因此,在研究投资者情绪时,仅利用间接指标无法准确反映投资者文本情绪的完整信息。同时,已有研究表明,换手率越高,异质信念越大。但是,由表4可知,换手率与文本异质信念显著负相关,说明异质信念的间接指标存在较大噪音,甚至使该指标存在相反的影响效应,从而扭曲了异质信念的真实作用;文本异质信念的  $R^2$  只有0.117和0.099,表明间接指标对直接指标存在较低的解释度,文本异质信念存在大量私有信息。表4的结果表明,在研究投资者情绪和异质信念时,只使用间接指标很可能导致投资者情绪和异质信念对市场崩盘风险的影响效应估计结果存在较大偏差,从而突显了利用投资者文本情绪和文本异质信念测量投资者情绪和异质信念的优越性。

4 实证研究

本研究利用月度指标分析投资者文本情绪和文

本异质信念与惯性效应、市场波动率、流动性的关系,探讨投资者文本情绪和文本异质信念对市场崩盘风险的效应及机制。同时,为了确保不同频率数据及外生冲击条件下实证结果的可靠性,本研究还利用日度指标替代月度指标以及新冠疫情的延长样本数据进行稳健性检验。

4.1 投资者文本情绪和文本异质信念反映市场微观结构的变化

根据2.2部分的理论分析及投资者文本情绪和文本异质信念与市场收益的动态趋势,本研究认为投资者文本情绪和文本异质信念可能与市场微观结构存在紧密联系。为进一步分析投资者文本情绪和文本异质信念背后反映的逻辑,本研究构建模型为

$$BI_t/HB_t = c + \beta_1 CR_t + \beta_2 Sig_t + \beta_3 Ami_t + \varepsilon_t^1 \quad (5)$$

其中,  $c$  为截距项,  $\beta_1 \sim \beta_3$  为回归系数,  $\varepsilon_t^1$  为残差项。

表5给出基于(5)式的回归结果,累积收益率与投资者文本情绪的回归系数显著正相关,与投资者文本异质信念的回归系数显著负相关;历史波动率和非流动性与投资者文本情绪的回归系数显著负相关,与投资者文本异质信念的回归系数显著正相关。以上结果表明,投资者文本情绪与市场的惯性效应和流动性正相关,投资者文本异质信念与市场波动率正相关,与惯性效应和流动性负相关,  $H_1$  得到部分验证。针对表5的结果,本研究的解释是:首先,惯

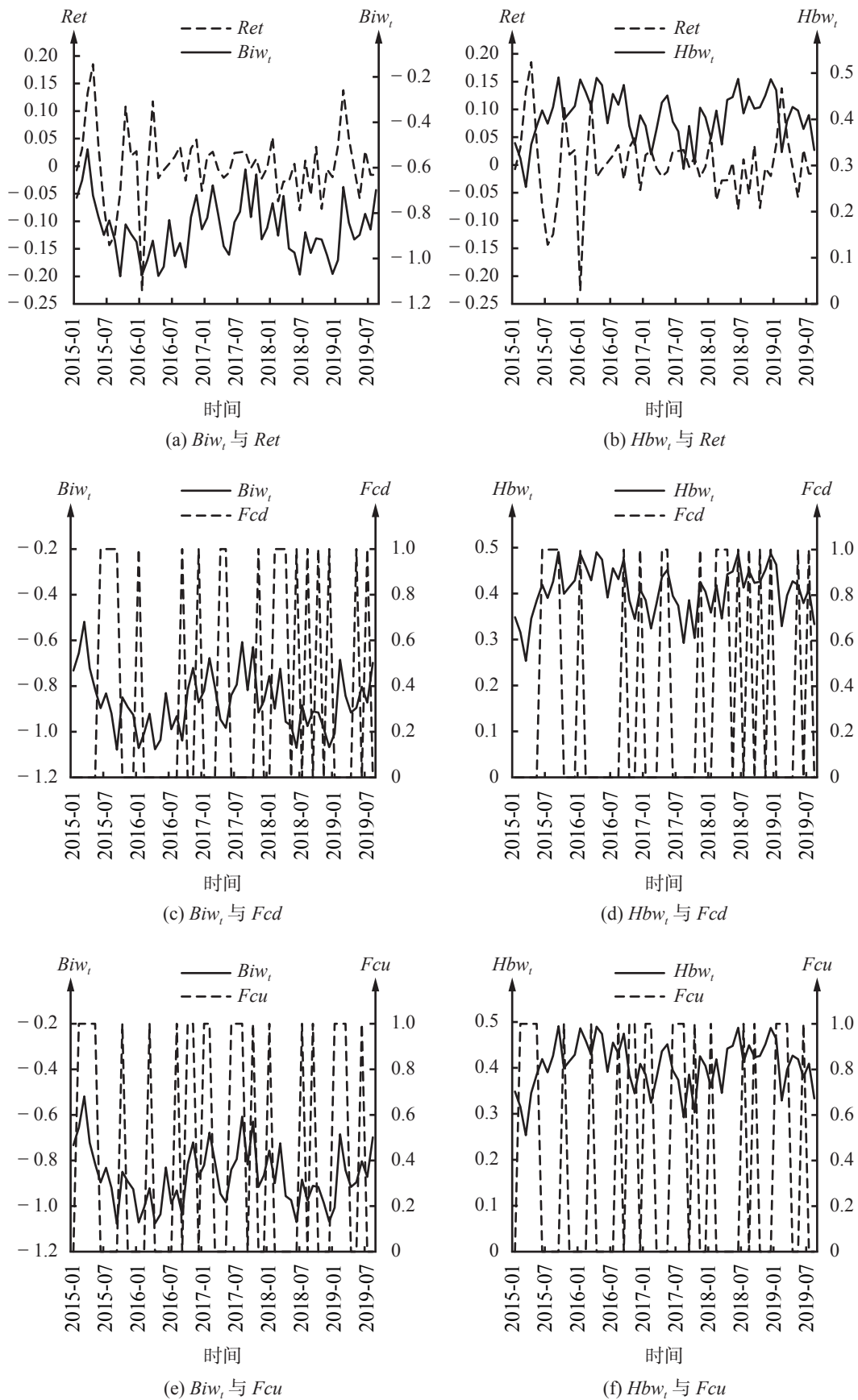


图2  $Biw_t$  和  $Hbw_t$  与上证综指月收益率和月度市场崩盘风险的走势

Figure 2  $Biw_t$  and  $Hbw_t$  and Trend of Monthly Returns and Market Crash Risks of Shanghai Composite Stock Index



表3 投资者文本情绪和文本异质信念的优越性检验结果

Table 3 Test Results for Advantage of Investor Textual Sentiment and Textual Heterogeneous Belief

	$Bim_t$	$Hbm_t$	$Isi_t$	$Cic_t$	$Tur_t$
$Ret$	0.286*** (0.060)	-0.681*** (0.145)	0.001** (0.0005)	0.003 (0.002)	0.070 (0.087)
截距项	0.247*** (0.052)	0.274*** (0.059)	-0.067** (0.027)	-0.120 (0.088)	-0.010 (0.015)
样本数	56	56	56	56	56
$R^2$	0.294	0.289	0.114	0.034	0.012

注:括号内数据为稳健性标准差,\*\*\*为在1%水平上显著,\*\*为在5%水平上显著,下同。

表4 投资者情绪和异质信念的直接指标与其间接指标的关系

Table 4 Relationship between Direct and Indirect Variables of Investor Sentiment and Heterogeneous Belief

	$Bim_t$		$Biw_t$		$Hbm_t$	$Hbw_t$
	(1)	(2)	(3)	(4)		
$Isi_t$	0.003*** (0.001)		0.002*** (0.006)			
$Cic_t$		0.011*** (0.003)		0.011*** (0.006)		
$Tur_t$					-0.186*** (0.006)	-0.184*** (0.008)
截距项	-1.022*** (0.00002)	-1.341*** (0.0001)	-1.012*** (0.0004)	-1.328*** (0.00008)	0.429*** (0.0002)	0.432*** (0.0006)
样本数	56	56	56	56	56	56
$R^2$	0.163	0.136	0.109	0.112	0.117	0.099

表5 投资者文本情绪和文本异质信念对市场微观结构的影响

Table 5 Impacts of Investor Textual Sentiment and Textual Heterogeneous Belief on Market Microstructure

	$Bim_t$	$Biw_t$	$Hbm_t$	$Hbw_t$
$CR_t$	0.117** (0.047)	0.124** (0.052)	-0.049** (0.020)	-0.052** (0.022)
$Sig_t$	-0.802* (0.426)	-1.067** (0.478)	0.340* (0.179)	0.445** (0.200)
$Ami_t$	-0.249*** (0.053)	-0.189*** (0.059)	0.102*** (0.022)	0.078*** (0.025)
截距项	-0.719*** (0.034)	-0.737*** (0.038)	0.343*** (0.014)	0.351*** (0.016)
样本数	56	56	56	56
$R^2$	0.430	0.326	0.422	0.321

注:\*为在10%水平上显著,下同。

性效应的增强反映了投资者情绪升高和异质信念减少。因为市场处于惯性上升趋势,过去一段时间收

益率较高,未来仍有很大的可能获得较高收益率。许多市场投资者的情绪受到其他投资者的影响而上升,从而出现了非理性从众心理和羊群行为,此时投资者的分歧较小,进而异质信念也较弱。其次,市场波动率变小,市场投资者分歧较小,投资者心理波动水平也会较低,从而情绪水平较高;最后,市场流动性水平较低体现了投资者分歧较大,投资者的交易意愿较低,从而情绪水平也越低。

#### 4.2 投资者文本情绪和文本异质信念对市场崩盘风险的影响

依据投资者文本情绪和文本异质信念与市场微观结构的逻辑关系,本研究探讨投资者文本情绪和文本异质信念对市场崩盘风险的影响效应。采用逻辑回归模型,具体模型为

$$\Pr[FC_t = 1] = \gamma + \alpha_1 BI_t / HB_t + \alpha_2 BI_{t-1} / HB_{t-1} + Yea + \varepsilon_t^2 \quad (6)$$

其中, $\gamma$ 为截距项, $\alpha_1$ 和 $\alpha_2$ 为回归系数, $\varepsilon_t^2$ 为残差项。由于当月收益率以收盘价计算,所以当月产生的投资者评论以及计算的投资者文本情绪和文本异质信念先于月尾收盘价,同时投资者文本情绪和文本异

质信念存在均值回复的趋势,即前后两期之间存在负的相关性,因此,本研究把当月和上月的文本情绪或文本异质信念同时纳入模型中。故月度数据样本数从56减少为55。由于上月与当月变量之间的相关性,本研究选择使用上月的文本情绪和当月的文本异质信念分析投资者情绪和异质信念对崩盘风险的影响。

表6和表7给出基于(6)式的回归结果,分别探讨文本情绪和文本异质信念对向下和向上崩盘风险的影响。由表6可知,上月的投资者文本情绪高涨使发生向下崩盘风险的可能性增加,即投资者情绪越趋向积极,向下崩盘风险越可能发生。当月的异质信念越低,出现向下崩盘风险的可能性越小。综上所述,投资者文本情绪趋向乐观而文本异质信念较低,向下的市场崩盘风险变小;投资者文本情绪趋向悲观而文本异质信念较高,向下的市场崩盘风险变大。表7的回归结果与表6的相反,上月的投资者文本情绪高涨使当月发生向上崩盘风险的可能性减少,当月的异质信念越高,出现向上崩盘风险的可能性

越小。因此,  $H_2$  得到验证。

基于市场微观结构的分析,当投资者乐观情绪高涨而投资者的异质信念较小时,市场波动率被抑制,市场的上涨惯性将会持续下去,此时向下崩盘风险较小,而向上崩盘风险较高;当投资者情绪较低而异质信念较高时,悲观投资者占据上风,市场波动率上升而流动性较低,市场的向上崩盘风险变小,而向下崩盘风险骤增。

#### 4.3 投资者文本情绪和文本异质信念影响市场崩盘风险的作用机制

本研究尝试利用 Pearson 相关性检验,分别从ETF的换手率和现金持有探讨投资者文本情绪和文本异质信念影响市场崩盘风险的作用机制。表8给出ETF的换手率和现金持有对市场崩盘风险影响的检验结果,因为表5已经给出投资者文本情绪和文本异质信念与市场微观结构变量的相关关系,表8省略了市场微观结构变量与ETF变量的相关关系,后续总结分析可依据表8中投资者文本情绪和文本异质信念与ETF变量的相关关系,直接推断市场微观结

表6 月度投资者文本情绪和文本异质信念对向下崩盘风险的影响

Table 6 Effects of Monthly Investor Textual Sentiment and Textual Heterogeneous Belief on Downward Crash Risks

	$DF_1$		$DF_2$		$DF_3$	
	(1)	(2)	(3)	(4)	(5)	(6)
$Bim_t$	-18.476*** (6.066)		-16.229*** (5.675)		-13.115*** (4.988)	
$Bim_{t-1}$	10.913** (4.542)		9.823** (4.528)		9.845** (4.500)	
$Biw_t$		-14.306*** (4.793)		-13.709*** (4.891)		-12.614*** (4.789)
$Biw_{t-1}$		7.202** (3.382)		6.571* (3.498)		7.946** (3.726)
$Yea$	控制	控制	控制	控制	控制	控制
截距项	-7.337* (4.188)	-6.912* (3.799)	-6.303 (4.164)	-6.903* (3.938)	-3.685 (3.774)	-4.855 (3.805)
$Hbm_t$	44.682*** (14.755)		39.142*** (13.802)		31.357*** (12.041)	
$Hbm_{t-1}$	-25.399** (10.622)		-22.764** (10.542)		-22.562** (10.429)	
$Hbw_t$		35.208*** (11.917)		33.710*** (12.144)		30.941*** (11.856)
$Hbm_{t-1}$		-16.775** (7.990)		-15.257* (8.237)		-18.427** (8.737)
$Yea$	控制	控制	控制	控制	控制	控制
截距项	-8.567* (4.776)	-8.212* (4.392)	-7.359 (4.732)	-8.180* (4.545)	-4.381 (4.266)	-5.863 (4.386)
样本数	55	55	55	55	55	55

表7 月度投资者文本情绪和文本异质信念对向上崩盘风险的影响

Table 7 Effects of Monthly Investor Textual Sentiment and Textual Heterogeneous Belief on Upward Crash Risks

	$UF_1$		$UF_2$		$UF_3$	
	(1)	(2)	(3)	(4)	(5)	(6)
$Bim_t$	16.732*** (4.878)		12.123*** (3.832)		12.136*** (3.865)	
$Bim_{t-1}$	-7.470** (3.539)		-6.776** (3.278)		-5.310 (3.241)	
$Biw_t$		16.557*** (4.571)		11.440*** (3.456)		11.144*** (3.499)
$Biw_{t-1}$		-6.778** (3.303)		-6.064* (3.171)		-4.411 (3.117)
$Yea$	控制	控制	控制	控制	控制	控制
截距项	7.489** (3.723)	8.086** (3.805)	3.771 (3.232)	3.864 (3.123)	4.939 (3.257)	4.969 (3.157)
$Hbm_t$	-39.801*** (11.643)		-28.571*** (9.101)		-28.599*** (9.175)	
$Hbm_{t-1}$	17.379** (8.345)		15.741** (7.784)		12.157 (7.706)	
$Hbw_t$		-39.186*** (10.949)		-26.574*** (8.172)		-25.973*** (8.272)
$Hbm_{t-1}$		15.431** (7.762)		13.777* (7.531)		9.813 (7.415)
$Yea$	控制	控制	控制	控制	控制	控制
截距项	8.538** (4.189)	9.232** (4.300)	4.333 (3.636)	4.390 (3.508)	5.681 (3.661)	5.678 (3.546)
样本数	55	55	55	55	55	55

表8 作用机制检验结果

Table 8 Testing Results for Influential Mechanisms

	$DF_1$	$DF_2$	$DF_3$	$UF_1$	$UF_2$	$UF_3$	$Bim_t$	$Biw_t$	$Hbm_t$	$Hbw_t$
$Tuf_t$	0.057	0.039	0.049	-0.210	-0.244	-0.196	0.068	0.065	-0.076	-0.075
$Cas_t$	0.332**	0.290**	0.294**	-0.283**	-0.194	-0.199	-0.369**	-0.288**	0.369**	0.299**

构变量与ETF变量的关系。

由表8可知,换手率与投资者文本情绪和文本异质信念都没有显著的相关关系,且与市场崩盘风险也没有显著的相关关系,这表明没有证据支持投资者文本情绪和文本异质信念通过噪声交易的途径影响市场崩盘风险。 $H_3$ 没有得到验证。现金持有与投资者文本情绪具有显著的负相关关系,与投资者文本异质信念具有显著的正相关关系,与向下崩盘风险显著正相关,与向上崩盘风险呈弱负相关关系。这意味着,投资者文本情绪越偏向乐观,文本异质信念越小,则现金占净比越小,即基金配置更多的资金用于购买权重股票,从而向下的崩盘风险减小,向上

的崩盘风险增加;投资者文本情绪趋向悲观而文本异质信念较高,则ETF的现金持有增加,向下的市场崩盘风险上升,向上的市场崩盘风险减少。因此, $H_4$ 得到验证。

为了更清晰地说明投资者文本情绪和文本异质信念通过作用于ETF的现金占净比影响市场崩盘风险,本研究用图3综合比较投资者文本情绪、文本异质信念、现金占净比与崩盘风险的动态趋势。由图3的(a)和(b)可知,在2015年4月至2016年1月,投资者情绪开始显著下降,异质信念呈波动上升趋势。因此,ETF需要卖出更多的股票来准备更多的现金,以应对流动性紧张的状况。由图3的(c)可知,市场



股指向下崩盘风险的概率增加,在2019年1月至2019年4月,投资者情绪进入高峰,异质信念处于低谷,现金占净比显著下降。因此,ETF无需持有更多现金,ETF配置了更大比例的股票,导致市场股指遭

受向上崩盘风险的概率增加。

在上述实证检验基础上,本研究进一步发现,投资者文本情绪或文本异质信念反映了市场微观结构变化,通过驱动ETF行为变化影响市场崩盘风险。具体分析如下:

(1)上月的投资者文本情绪较高,根据均值回复原理,当月文本情绪出现反转的可能性越大。从而,当月投资者文本情绪会较低,惯性效应会较弱,市场波动率和流动性成本也会较高。投资者对股票市场增长的信心越弱,上证综指ETF投资者赎回行为就越多,基金就会被动卖出更多股票,以持有更多现金以备赎回,导致当月发生向下崩盘风险的可能性较高,而向上崩盘风险较低。

(2)上月的投资者分歧小,文本异质信念在月份之间通常会出现反转的趋势,则当月的分歧就会变大,意味着投资者中积极评论的人数与消极评论的人数之间存在较大的差距。在投资者文本情绪较弱而文本异质信念较大时,惯性效应变弱,市场波动率和流动性成本增高,从而投资者增加ETF的赎回行为,基金公司被动卖出大量股票,进而向下的崩盘风险出现概率增大。文本异质信念对向上崩盘风险的作用与向下崩盘风险的情形类似。

(3)上月的投资者分歧较大,文本异质信念在月份之间通常会出现反转的趋势,则当月的分歧就会变小,意味着惯性效应较强,市场波动率和流动性成本较小,市场处于上行高位,从而向上崩盘风险的概率增加。在投资者文本情绪较高和文本异质信念较小时,积极评论的人数高于消极评论的人数,上证综指升高,从而ETF申购行为增加,基金公司会增加权重股的配置,进而向上崩盘风险增加。

4.4 稳健性检验

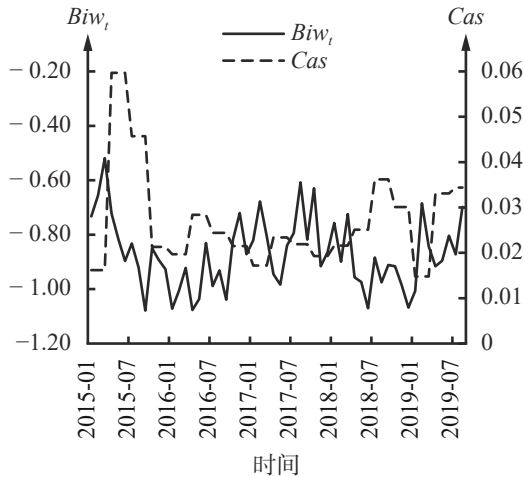
本研究利用日度指标替代月度指标以及新冠疫情的延长样本数据进行稳健性检验。利用日度指标有助于消除因数据频率过于低频而导致实证结果不稳健的忧虑,而延长样本数据的时间序列至新冠疫情期间,能够进一步说明外生冲击干扰下实证结果的可靠性。

投资者文本情绪和文本异质信念具有均值回复的特点,本研究使用与月度指标回归类似的逻辑回归模型进行稳健性检验。采用月窗口滚动的日度投资者文本情绪和文本异质信念分别对市场崩盘风险进行回归。具体模型为

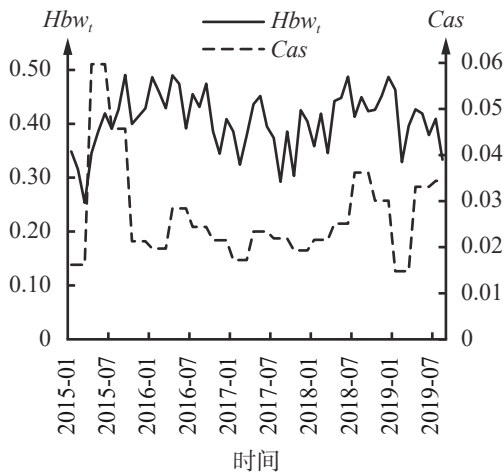
$$Pr[FC_i = 1] = \phi + \theta_1 BI_i / HB_i + \theta_2 BI_{i-1} / HB_{i-1} + \varepsilon_i^3 \quad (7)$$

其中,  $FC_i$  分别为  $Fcd_i$  和  $Fcu_i$ ,  $\phi$  为截距项,  $\theta_1$  和  $\theta_2$  为回归系数,  $\varepsilon_i^3$  为残差项。因模型中存在一阶滞后项,故日度数据样本从1115减少为1114。

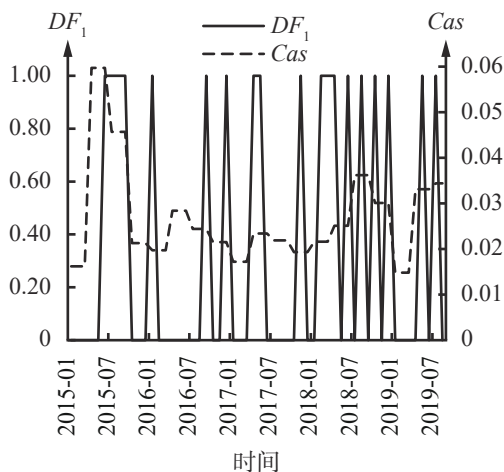
表9给出基于(7)式的回归结果,与表6的月度数据估计结果相比,日度投资者文本情绪和文本异质信念对向下崩盘风险的影响与月度数据的影响基本一致。日度投资者文本情绪和文本异质信念对向上崩盘风险的影响与表7的月度数据也基本一致,不



(a) 投资者文本情绪与现金净占比



(b) 投资者文本异质信念与现金净占比



(c) 向下崩盘风险与现金净占比

图3 解释机制分析

Figure 3 Explanation Mechanism Analysis

表9 日度投资者文本情绪和文本异质信念对向下崩盘风险的影响

Table 9 Effects of Daily Investor Textual Sentiment and Textual Heterogeneous Belief on Downward Crash Risks

	$DF_1$		$DF_2$		$DF_3$	
	(1)	(2)	(3)	(4)	(5)	(6)
$Bim_i$	-71.328*** (9.573)		-81.046*** (11.274)		-81.886*** (12.486)	
$Bim_{i-1}$	73.834*** (9.641)		82.329*** (11.315)		82.725*** (12.517)	
$Biw_i$		-121.610*** (12.264)		-137.421*** (15.585)		-137.442*** (17.171)
$Biw_{i-1}$		11.756*** (1.173)		13.086*** (1.471)		13.011*** (1.611)
截距项	-1.066 (0.880)	-3.083*** (1.079)	-2.613** (1.078)	-5.588*** (1.412)	-3.293*** (1.231)	-6.654*** (1.635)
$Hbm_i$	159.048*** (22.393)		179.561*** (26.279)		182.149*** (29.180)	
$Hbm_{i-1}$	-164.573*** (22.443)		-181.992*** (26.229)		-183.515*** (29.086)	
$Hbw_i$		306.062*** (31.657)		352.604*** (41.125)		356.515*** (45.934)
$Hbw_{i-1}$		-29.345*** (2.992)		-33.272*** (3.831)		-33.436*** (4.253)
截距项	-0.945 (0.972)	-3.879*** (1.240)	-2.657** (1.193)	-6.988*** (1.663)	-3.382** (1.366)	-8.275*** (1.942)
样本数	1 114	1 114	1 114	1 114	1 114	1 114

再重复列表。以上结果说明本研究结果稳健。

本研究采用负收益偏态系数和收益上下波动率比率作为崩盘风险的替代变量,进行稳健性检验。检验模型为

$$Ncw_i = m^1 + k^1 BI_i / HB_i + \varepsilon_i^4 \quad (8)$$

$$Dul_i = m^2 + k^2 BI_i / HB_i + \varepsilon_i^5 \quad (9)$$

其中,  $m^1$  和  $m^2$  为截距项,  $k^1$  和  $k^2$  为回归系数,  $\varepsilon_i^4$  和  $\varepsilon_i^5$  为残差项。因滞后1日的解释变量不显著,故未放入模型设定中。

表10给出基于(8)式和(9)式的回归结果,所有解释变量的回归系数均在1%水平上显著,并且投资者文本情绪的回归系数为正,文本异质信念的回归系数为负,与前文结果一致。

FORSTER et al.<sup>[36]</sup>认为新冠疫情的实际暴发是从2019年9月开始的,因此本研究将样本区间拓展至2020年5月22日,包括了疫情期间的数据样本。表11给出新冠疫情下延长样本数据的稳健性检验结果,与表10的模型设定和变量相同,结果表明回归系数的显著性和作用与表10结果一致,说明本研究结果稳健。

## 5 结论

预防股票市场崩盘风险是防范金融市场系统性风险的重要内容,裘江南等<sup>[37]</sup>和陈康等<sup>[38]</sup>关于市场情绪的研究结果表明,在市场环境的骤变过程中,投资者文本情绪和文本异质信念能够深刻地影响股票市场崩盘风险。本研究基于2015年至2019年东方财富网股吧关于上证市场的投资者评论文本信息,通过机器学习的文本分析,构建投资者文本情绪和文本异质信念指标,探讨投资者文本情绪和文本异质信念如何影响市场崩盘风险。研究结果表明,①投资者文本情绪和文本异质信念与惯性效应、市场波动率和流动性因素存在显著的关联,这些因素能够显著反映投资者心理、情绪和异质信念变化。市场的惯性效应增强,波动率较小,伴随着投资者文本情绪升高和文本异质信念变小,而市场非流动性水平升高则伴随着投资者文本情绪降低且文本异质信念增大。②根据月度和日度数据的实证检验发现,在投资者文本情绪较高和文本异质信念较低时,向下的市场崩盘风险较低,向上的市场崩盘风险较高;在投资者文本情绪较低和文本异质信念较高时,向下的市场崩盘风险较高,向上的市场崩盘风险较低。

表 10 采用崩盘风险替代变量的稳健性检验结果  
Table 10 Robust Test Results for Using Substitution Variables of Crash Risks

	$Ncw_i$				$Dul_i$			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Bi_m$	1.648*** (0.187)				0.357*** (0.116)			
$Bi_w$		0.147*** (0.017)				0.032*** (0.011)		
$Hb_m$			-3.946*** (0.448)				-0.821*** (0.278)	
$Hb_m$				-0.351*** (0.041)				-0.073*** (0.025)
截距项	1.823*** (0.163)	1.735*** (0.156)	1.990*** (0.182)	1.886*** (0.174)	0.380*** (0.101)	0.363*** (0.097)	0.403*** (0.113)	0.380*** (0.108)
样本数	1 114	1 114	1 114	1 114	1 114	1 114	1 114	1 114
$R^2$	0.065	0.063	0.065	0.062	0.008	0.008	0.008	0.007

表 11 新冠疫情下延长样本数据的稳健性检验结果  
Table 11 Robust Test Results for Using Extended Sample Under the COVID-19

	$Ncw_i$				$Dul_i$			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Bim_i$	1.573*** (0.197)				0.316*** (0.105)			
$Biw_i$		1.475*** (0.190)				0.304*** (0.100)		
$Hbm_i$			-3.753*** (0.471)				-0.740*** (0.250)	
$Hbm_i$				-3.507*** (0.453)				-0.704*** (0.240)
截距项	1.790*** (0.170)	1.706*** (0.163)	1.943*** (0.189)	1.846*** (0.182)	0.354*** (0.090)	0.343*** (0.086)	0.379*** (0.100)	0.365*** (0.096)
样本数	1 291	1 291	1 291	1 291	1 291	1 291	1 291	1 291
$R^2$	0.047	0.045	0.047	0.044	0.007	0.007	0.007	0.007

③投资者文本情绪和文本异质信念作用于ETF的申购和赎回行为,导致基金被动减持或增持股票,进而影响市场指数走势及崩盘风险发生的可能性。当投资者文本情绪低且文本异质信念高时,市场的惯性效应变弱,市场波动率和流动性成本升高,个体投资者对ETF的赎回行为增加,促使ETF需要持有更多现金而被动大量卖出权重股,市场股指加速下降;当投资者文本情绪高且文本异质信念较小时,个体投资者申购ETF需求增加,促使ETF被动大量买入权重股,市场股指加速上升。

本研究可能的边际贡献在于:①本研究结果为投资者文本情绪和文本异质信念的测量及效应提供更

直接和准确的证据。本研究基于优化后的机器学习方法获得的情绪分类结果准确度达到89.045%,比已有研究具有一定优势。②本研究发现投资者文本情绪和文本异质信念与市场微观结构紧密相关,揭示了投资者文本情绪和文本异质信念作用于市场的渠道。③本研究发现投资者文本情绪利用市场微观结构的变化作用于ETF进而影响市场崩盘风险的机制。因此,本研究把许年行等<sup>[39]</sup>发现的投资机构影响崩盘风险的机制进一步拓展到考察ETF行为层面。

本研究的政策启示在于:①本研究发现,高涨的投资者文本情绪和较低的文本异质信念会显著提升股市的向下崩盘风险,从而为投资者情绪和异质信



念与市场崩盘风险的理论关系提供了直接证据,并验证了投资者情绪和异质信念通过ETF的申购和赎回行为作用于市场崩盘风险的机制。这说明机构投资者并没有成为市场的稳定器,尤其是被动投资机构ETF,反而会受到个体投资者针对市场走势的情绪和异质信念的影响,对市场过度反应,加剧市场崩盘风险。因此,相关监管部门应进一步加强对主动型机构投资者的培育,限制被动型基金的发行和交易规模,同时加紧拓宽基于ETF的金融衍生品市场,引导机构投资者利用对冲工具防范投资者情绪和异质信念变化驱动的被动买卖行为(踩踏行为)。<sup>②</sup>本研究认为,现实中向上崩盘风险往往会被忽略,随着融资融券交易的放开,市场杠杆交易者显著增多,向上和向下的崩盘风险都需要引起重视。尽管当前资本市场受到许多限制,如涨跌停板限制,但是随着未来监管和交易限制放松,市场暴涨或暴跌将导致投资者面临前所未有的极端风险。因此,稳定个体投资者情绪和降低其异质信念越发重要。相关监管部门应当进一步加强市场信息的披露机制,减少各类投资者的信息不对称,及时且多渠道地公开各类经济政策,并通过各类互联网工具加强与投资者的沟通和交流。<sup>③</sup>与张普等<sup>[40]</sup>的研究发现类似,投资者文本情绪和文本异质信念与市场本身的先期趋势、波动率、流动性息息相关,及时抑制投资者的羊群行为、降低市场的过度波动、提升市场流动性是管控个体投资者情绪和异质信念的重要途径。

本研究的不足之处在于:①在分析针对市场走势的个体投资者情绪和异质信念如何作用于市场崩盘风险的问题中,本研究重点关注情绪和异质信念影响机构投资者行为的渠道,特别是ETF行为的变化,而理论中揭示的个体投资者本身的噪声行为渠道还有待挖掘。②基于数据的可得性,本研究对ETF行为机制的检验只能基于可行的季度频率ETF持仓数据。在未来数据可获取的情况下,可以从更高频率的角度检验该行为机制。③投资者文本情绪和文本异质信念指标是建立在基于机器学习的文本分析基础上的,尽管该方法已经获得了较高的精度,但是未来研究还可以基于神经网络的深度学习方法进一步深化该类文本分析指标。

#### 参考文献:

- [1] LEE C M C, SHLEIFER A, THALER R H. Investor sentiment and the closed-end fund puzzle. *The Journal of Finance*, 1991, 46(1): 75-109.
- [2] WURGLER J A, BAKER M P, WURGLER J. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 2006, 61(4): 1645-1680.
- [3] HUANG D S, JIANG F W, TU J, et al. Investor sentiment aligned: a powerful predictor of stock returns. *The Review of Financial Studies*, 2015, 28(3): 791-837.
- [4] ANTWEILER W, FRANK M Z. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 2004, 59(3): 1259-1294.
- [5] 董大勇,肖作平. 证券信息交流家乡偏误及其对股票价格的影响:来自股票论坛的证据. *管理世界*, 2011, 27(1): 52-61.  
DONG Dayong, XIAO Zuoping. The home bias in the exchange of the stock information and its impact on stock prices: evidences from the stock forum. *Journal of Management World*, 2011, 27(1): 52-61.
- [6] LI J, CHEN Y, SHEN Y, et al. *Measuring China's stock market sentiment*. Durham, NC: Duke University, 2019.
- [7] 尹海员,吴兴颖. 投资者高频情绪对股票日内收益率的预测作用. *中国工业经济*, 2019(8): 80-98.  
YIN Haiyuan, WU Xingying. Predictive effect of high-frequency investor sentiment on the intraday stocks return. *China Industrial Economics*, 2019(8): 80-98.
- [8] JIANG F W, LEE J, MARTIN X M, et al. Manager sentiment and stock returns. *Journal of Financial Economics*, 2019, 132(1): 126-149.
- [9] 姚加权,冯绪,王赞钧,等. 语调、情绪及市场影响:基于金融情绪词典. *管理科学学报*, 2021, 24(5): 26-46.  
YAO Jiaquan, FENG Xu, WANG Zanjun, et al. Tone, sentiment and market impacts: the construction of Chinese sentiment dictionary in finance. *Journal of Management Sciences in China*, 2021, 24(5): 26-46.
- [10] 杨晓兰,沈翰彬,祝宇. 本地偏好、投资者情绪与股票收益率:来自网络论坛的经验证据. *金融研究*, 2016(12): 143-158.  
YANG Xiaolan, SHEN Hanbin, ZHU Yu. The effect of local bias in investor attention and investor sentiment on stock markets: evidence from online forum. *Journal of Financial Research*, 2016(12): 143-158.
- [11] LI S Y, NING K X, ZHANG T. Sentiment-aware jump forecasting. *Knowledge-based Systems*, 2021, 228: 107292-1-107292-20.
- [12] 唐国豪,姜富伟,张定胜. 金融市场文本情绪研究进展. *经济学动态*, 2016(11): 137-147.  
TANG Guohao, JIANG Fuwei, ZHANG Dingsheng. The progress of research on textual sentiment in financial markets. *Economic Perspectives*, 2016(11): 137-147.
- [13] HUTTON A P, MARCUS A J, TEHRANIAN H. Opaque financial reports,  $R^2$ , and crash risk. *Journal of Financial Economics*, 2009, 94(1): 67-86.
- [14] 赵臻,陈仕华,曹伟. “互联网+”信息披露:实质性陈述还是策略性炒作:基于股价崩盘风险的证据. *中国工业经济*, 2020(3): 174-192.  
ZHAO Can, CHEN Shihua, CAO Wei. “Internet plus” information disclosure: substantive statement or strategic manipulation: evidence based on the risk of stock price crash. *China Industrial Economics*, 2020(3): 174-192.
- [15] 夏常源,王靖懿,傅代国. 保险资金持股与股价崩盘风险:市场“稳定器”还是崩盘“加速器”? *经济管理*, 2020, 42(4): 158-174.  
XIA Changyuan, WANG Jingyi, FU Daiguo. Insurance funds and corporate stock price crash risk: market “stabilizer” or crash “accelerator”? *Business and Management Journal*, 2020, 42(4): 158-174.
- [16] BAI M, WANG R X, YU C F, et al. Limits on executive pay and stock price crash risk: evidence from a quasi-natural experiment. *Pacific-Basin Finance Journal*, 2019, 55: 206-221.
- [17] XU Y X, XUAN Y H, ZHENG G P. Internet searching and stock

- price crash risk: evidence from a quasi-natural experiment. *Journal of Financial Economics*, 2021, 141(1): 255–275.
- [18] DENG X H, GAO L. The monitoring of short selling: evidence from China. *Research in International Business and Finance*, 2018, 43(1): 68–78.
- [19] DENG X H, GAO L, KIM J B. Short-sale constraints and stock price crash risk: causal evidence from a natural experiment. *Journal of Corporate Finance*, 2020, 60: 101498-1–101498-16.
- [20] 杨七中, 章贵桥, 马蓓丽. 管理层语意与未来股价崩盘风险: 基于投资者情绪的中介效应分析. *中南财经政法大学学报*, 2020(1): 26–36.  
YANG Qizhong, ZHANG Guiqiao, MA Beili. Management meanings and stock price crash risk: the mediating role of investors sentiments. *Journal of Zhongnan University of Economics and Law*, 2020(1): 26–36.
- [21] 段江娇, 刘红忠, 曾剑平. 中国股票网络论坛的信息含量分析. *金融研究*, 2017(10): 178–192.  
DUAN Jiangjiao, LIU Hongzhong, ZENG Jianping. Analysis on the information content of China's internet stock message boards. *Journal of Financial Research*, 2017(10): 178–192.
- [22] 林红新, 宋高雅, 花俊国. 投资者文本情绪对动量效应与反转效应的影响. *金融理论探索*, 2020(1): 8–18.  
LIN Hongxin, SONG Gaoya, HUA Junguo. Research on the impact of investor emotion on momentum effect and inversion effect. *Exploration of Financial Theory*, 2020(1): 8–18.
- [23] 唐也然, 龙文, 石勇. 投资者非理性行为是动量效应产生的原因吗? 来自互联网金融论坛的证据. *证券市场导报*, 2020(4): 62–70.  
TANG Yeran, LONG Wen, SHI Yong. Is irrational behaviour of investors the cause of momentum effect? Evidence from online financial forum. *Securities Market Herald*, 2020(4): 62–70.
- [24] DE LONG J B, SHLEIFER A, SUMMERS L H, et al. Noise trader risk in financial markets. *Journal of Political Economy*, 1990, 98(4): 703–738.
- [25] 陈国进, 张贻军. 异质信念、卖空限制与我国股市的暴跌现象研究. *金融研究*, 2009(4): 80–91.  
CHEN Guojin, ZHANG Yijun. Short sale restriction, heterogeneous beliefs and stock price crash in China. *Journal of Financial Research*, 2009(4): 80–91.
- [26] 孙鲲鹏, 肖星. 互联网社交媒体对投资者情绪传染与股价崩盘风险的影响机制. *技术经济*, 2018, 37(6): 93–102.  
SUN Kunpeng, XIAO Xing. Influence of social media in internet on sentiment contagion and stock price crash. *Journal of Technology Economics*, 2018, 37(6): 93–102.
- [27] 关静怡, 朱恒, 刘娥平. 股吧评论、分析师跟踪与股价崩盘风险: 关于模糊信息的信息含量分析. *证券市场导报*, 2020(3): 58–68.  
GUAN Jingyi, ZHU Heng, LIU Eping. Stock BBS reviews, analyst coverage and stock price crash risk: analysis on the information content of ambiguous information. *Securities Market Herald*, 2020(3): 58–68.
- [28] 丁慧, 吕长江, 陈运佳. 投资者信息能力: 意见分歧与股价崩盘风险: 来自社交媒体“上证e互动”的证据. *管理世界*, 2018, 34(9): 161–171.  
DING Hui, LYU Changjiang, CHEN Yunjia. Investor information capability, difference of opinion and stock price crash risk: evidence from SSE e-interaction. *Journal of Management World*, 2018, 34(9): 161–171.
- [29] LIU J N, STAMBAUGH R F, YUAN Y. Size and value in China. *Journal of Financial Economics*, 2019, 134(1): 48–69.
- [30] MARIN J M, OLIVIER J P. The dog that did not bark: insider trading and crashes. *The Journal of Finance*, 2008, 63(5): 2429–2476.
- [31] KIM J B, LI Y H, ZHANG L D. CFOs versus CEOs: equity incentives and crashes. *Journal of Financial Economics*, 2011, 101(3): 713–730.
- [32] 部慧, 解峥, 李佳鸿, 等. 基于股评的投资者情绪对股票市场的影响. *管理科学学报*, 2018, 21(4): 86–101.  
BU Hui, XIE Zheng, LI Jiahong, et al. Investor sentiment extracted from internet stock message boards and its effect on Chinese stock market. *Journal of Management Sciences in China*, 2018, 21(4): 86–101.
- [33] 魏星集, 夏维力, 孙彤彤. 基于BW模型的A股市场投资者情绪测度研究. *管理观察*, 2014(33): 71–73, 76.  
WEI Xingji, XIA Weili, SUN Tongtong. The study of measuring investor sentiment in China's A stock market based on the BW model. *Management Observer*, 2014(33): 71–73, 76.
- [34] 易志高, 茅宁. 中国股市投资者文本情绪测量研究: CICSII的构建. *金融研究*, 2009(11): 174–184.  
YI Zhigao, MAO Ning. Research on the measurement of investor sentiment in Chinese stock market: the CICSII's construction. *Journal of Financial Research*, 2009(11): 174–184.
- [35] 陈国进, 张贻军, 王景. 再售期权、通胀幻觉与中国股市泡沫的影响因素分析. *经济研究*, 2009, 44(5): 106–117.  
CHEN Guojin, ZHANG Yijun, WANG Jing. Resale option, inflation illusion and the Chinese stock market bubbles. *Economic Research Journal*, 2009, 44(5): 106–117.
- [36] FORSTER P, FORSTER L, RENFREW C, et al. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 2020, 117(17): 9241–9243.
- [37] 裘江南, 葛一迪. 社交媒体情绪对信息行为的影响: 基于两类灾害事件的比较研究. *管理科学*, 2020, 33(1): 3–15.  
QIU Jiangnan, GE Yidi. Influence of emotions in social media on information behavior in two types of typical disasters. *Journal of Management Science*, 2020, 33(1): 3–15.
- [38] 陈康, 江嘉骏, 刘琦, 等. 空气质量、投资者情绪与股票收益率. *管理科学*, 2018, 31(6): 145–160.  
CHEN Kang, JIANG Jiajun, LIU Qi, et al. Air quality, investor sentiment, and stock returns. *Journal of Management Science*, 2018, 31(6): 145–160.
- [39] 许年行, 于上尧, 伊志宏. 机构投资者羊群行为与股价崩盘风险. *管理世界*, 2013, 29(7): 31–43.  
XU Nianhang, YU Shangyao, YI Zhihong. Institutional investor herding and stock price crash risk. *Journal of Management World*, 2013, 29(7): 31–43.
- [40] 张普, 蒋月娥, 倪文辉. 羊群行为与股价波动的关系: 来自理论层面的证据. *管理科学*, 2021, 34(2): 144–158.  
ZHANG Pu, JIANG Yue'e, NI Wenhui. Relationship between herding behavior and stocks' price volatility: evidence from a theoretical perspective. *Journal of Management Science*, 2021, 34(2): 144–158.

## Investor Textual Sentiment, Textual Heterogeneous Belief and Market Crash Risks

LI Shaoyu<sup>1</sup>, ZHU Chunhui<sup>2</sup>, NING Kaixuan<sup>3</sup>, YU Zhihao<sup>3</sup>

1 International Business College, South China Normal University, Guangzhou 510631, China

2 School of Marxism, South China Normal University, Guangzhou 510631, China

3 School of Securities and Futures, Southwestern University of Finance and Economics, Chengdu 611130, China

**Abstract:** Different from individual stock crash risks, and in the circumstance of fundamental change of investor sentiment and heterogeneous belief, market crash risks are more disastrous. Defending market crash risks have become a hot issue that government and regulators focus on. However, few research has studied the impacts of investor sentiment and heterogeneous belief on market crash risks and their mechanisms.

This study adopts machine learning method to construct investor textual sentiment and textual heterogeneous belief (direct variables representing the investor sentiment and heterogeneous belief) from the reviews and comments on market tendency between January 1<sup>st</sup>, 2015 and August 27<sup>th</sup>, 2019 of the Shanghai Composite Stock Index bar in Eastmoney website. The advantage of investor textual sentiment and textual heterogeneous belief is tested by examining the correlation between the direct variables and indirect variables. The correlations between the direct variables and market microstructure variables (cumulative index return, volatility risk, and illiquidity) are examined, and they are used to check the effect and internal mechanism of the investor textual sentiment and textual heterogeneous belief on the crash risks of the Shanghai Stock Market.

It reveals that firstly, the investor textual sentiment and textual heterogeneous belief estimated by machine learning and textual analysis have more private information and better fit the market than traditional indirect variables. Second, the investor textual sentiment correlates positively to the cumulative index return, but negatively to index volatility and illiquidity; The investor textual heterogeneous belief correlates negatively to the cumulative index return, but positively to index volatility and illiquidity. Third, the investor textual sentiment and textual heterogeneous belief use the changes of microstructure variables to affect the Exchange Trade Fund (ETF) and then the risks of collapse. When the investor sentiment is low and heterogeneous belief is high, the market inertia effect becomes weaker, volatility risk and liquidity costs increase, which leads to increasing redemption of ETFs. Therefore, funds hold more cash and passively sell off massive weighted stocks, and the possibility of downward crash increases. On the contrary, the risk of upward crash will increase.

The findings enrich the former research on the factors and mechanisms of market crash risks, and provide important policy enlightenments on avoiding market systematic risks. That is, constraining the shepherd behavior and excessive volatile and increasing market liquidity are the main channels to control the investor sentiment and heterogeneous belief. Cultivating active institutional investors, restraining the trading scale and market share of negative fund investors, and introducing financial derivative of ETF can help cut down the mechanism of generating market crash risks.

**Keywords:** investor textual sentiment; investor textual heterogeneous belief; Shanghai Composite Stock Index; market crash risks; machine learning

**Received Date:** May 13<sup>th</sup>, 2021      **Accepted Date:** May 7<sup>th</sup>, 2022

**Funded Project:** Supported by the National Natural Science Foundation of China (72271099) and the Social Science and Humanities Research Project of Ministry of Education of China (22YJC790062, 22YJC790191)

**Biography:** LI Shaoyu, doctor in economics, is a professor in the International Business College at South China Normal University. His research interests cover financial risk management, derivative pricing, capital market, and financial big data analysis. His representative paper titled "Sentiment-aware jump forecasting" was published in the *Knowledge-Based Systems* (Issue 7, 2021). E-mail: [syli@ibc.scnu.edu.cn](mailto:syli@ibc.scnu.edu.cn)

ZHU Chunhui, doctor in economics, is a lecturer in the School of Marxism at South China Normal University. Her research interests cover digital economics, green and sustainable development, and state-own firm reform. Her representative paper titled "Government-led urbanization and natural gas demand in China" was published in the *Renewable & Sustainable Energy Reviews* (Issue 9, 2021). E-mail: [zhuchunhui@m.scnu.edu.cn](mailto:zhuchunhui@m.scnu.edu.cn)

NING Kaixuan is a master degree candidate in the School of Securities and Futures at Southwestern University of Finance and Economics. Her research interests include big data analysis and financial risk management. E-mail: [kaixuan\\_ning@126.com](mailto:kaixuan_ning@126.com)

YU Zhihao is a master degree candidate in the School of Securities and Futures at Southwestern University of Finance and Economics. His research interest focuses on financial risk management. E-mail: [a347044107@163.com](mailto:a347044107@163.com) □

(责任编辑:刘思宏)