

文本大数据视角下投资者关注的统计测度与互动特征研究

吴华凤

景德镇陶瓷大学经济与法学院

DOI:10.32629/jmsr.v5i2.21039

[摘要] 针对A股散户主导下投资者关注分化的问题,本文以2020—2024年沪深交易所互动平台128.6万条文本数据为样本,通过LDA算法识别价值、投机、政策三大关注主题,结合HP滤波、VAR模型、脉冲响应和方差分解等方法,多维度测度投资者关注特征并量化主题互动关系。研究发现:A股投资者关注呈现“投机短期主导、价值稳步上升、政策事件驱动”的特征;全面注册制对引导关注向价值转型作用显著;不同投资者类型、市场板块及企业特征下关注主题分布差异明显;政策主题是主导引导变量,对价值主题具有长期正向影响,对投机主题具有短期负向影响,价值与投机主题之间存在双向负向替代关系。多重稳健性检验表明上述结论成立。本文为缓解关注结构性失衡、增强资本市场稳定性提供微观依据,也拓展了文本大数据在金融行为研究中的应用场景。

[关键词] 投资者关注; 文本大数据; 互动特征; LDA; VAR模型

中图分类号: F830.59 **文献标识码:** A

Statistical Measurement and Interactive Characteristics of Investor Attention from the Perspective of Text Big Data

Huafeng Wu

School of Economics, Management and Law, Jingdezhen Ceramic University

[Abstract] Aiming at the divergence of investor attention in the retail-investor-dominated A-share market, this paper uses 1.286 million text entries from the interactive platforms of the Shanghai and Shenzhen Stock Exchanges during 2020—2024 as the sample. The LDA algorithm is used to identify value, speculation and policy themes, and HP filtering, VAR modelling, impulse response and variance decomposition are applied to measure attention characteristics and theme interactions. The results show that investor attention is characterized by short-term dominance of speculation, steady growth of value attention and event-driven policy attention. The comprehensive registration-based system significantly guides attention toward value. Policy attention plays a leading role, with a long-term positive impact on value attention and a short-term negative impact on speculative attention, while value and speculation show a bidirectional negative substitution relationship. Robustness checks support these conclusions. The study provides micro evidence for improving market stability and expands financial applications of text big data.

[Key words] Investor Attention; Text Big Data; Interactive Characteristics; LDA; VAR Model

1 引言

1.1 研究背景及意义

2023年末以来,A股市场在低位震荡中呈现投资者关注分化与短期化并存的特征。散户主导的交易结构使市场关注在题材炒作、政策事件与公司基本面之间快速切换,进而影响资金流向、价格发现与市场稳定。有限关注理论认为,投资者受认知资源约束,难以同时处理全部信息,往往会选择性关注部分信息。在长期投资、价值投资和注册制改革持续推进的背景下,识别投资者关注主题并揭示其替代、引导与传导机制,具有重要理论意

义。实践上,关注结构失衡可能加剧短期交易和非理性波动,削弱政策引导与信息披露效果。本文基于交易所互动文本,构建可连续更新的投资者关注测度框架,以期对市场预期管理、风险监测和上市公司投资者关系管理提供微观证据。

1.2 研究内容与方法

本文的研究内容包括四个方面:第一,构建覆盖上交所E互动、深交所互动易和投资者关系活动记录的标准化语料库;第二,利用LDA模型识别价值、投机、政策三类核心关注主题,并形成月度主题关注占比;第三,从时间变化、投资者类型、市场板

块、市值规模和账面市值比等维度考察关注结构差异；第四，运用VAR模型、脉冲响应和方差分解分析三类主题之间的动态互动关系。方法上，文本采集采用Python爬虫和requests、selenium库，文本预处理采用jieba分词、词性过滤、停用词剔除和TF-IDF筛选，计量分析采用描述性统计、HP滤波、ADF检验、信息准则选阶、AR根检验、脉冲响应和方差分解。

2 文献综述

2.1 投资者关注的测度方法研究

投资者关注研究最早建立在有限理性和有限注意力框架之上。Simon的有限理性思想为后续投资者注意力研究奠定了认知基础^[1]。早期研究主要使用换手率、交易量、分析师跟踪人数等变量间接度量关注，但这些变量容易受到流动性、市场行情和机构覆盖偏差影响。Peng和Xiong从类别学习角度指出，投资者会把有限注意力集中在市场类别、风格和显著信息上^[2]。随着网络行为数据的出现，Da等将搜索指数引入投资者关注测度，使关注研究从传统市场代理变量扩展到互联网搜索行为^[3]。Boucaud等认为，投资者在形成盈利预期时存在“预期黏性”，对企业盈利信息反应不足，导致过去盈利能力较强的股票在未来仍能获得较高收益，从而形成盈利能力异象^[4]。国内研究也利用百度指数、交易数据和公告数据考察关注与收益、公告效应之间的关系^{[5][6]}。

近年来，文本挖掘为投资者关注研究提供了新的切入点。交易所互动平台上的提问与回复直接反映投资者关注内容，较传统搜索指数更接近投资者对上市公司经营、政策变化和市场热点的真实关切。相关研究已开始使用互动平台数据讨论投资者互动与企业创新、信息披露等问题^[7]，但多侧重于互动强度或单一主题影响，对关注内容的结构分解仍不充分。

2.2 投资者关注的市场影响研究

投资者关注对资本市场的影响主要体现在价格形成和公司治理两个方面。在价格形成方面，关注会改变信息进入价格的速度，影响收益反应、动量收益和市场异象。Bouchaud等研究了注意力黏性与动量策略收益之间的关系，说明关注持续性会影响资产价格调整过程。国内研究发现，投资者关注能够解释中国股市收益变化和公告反应差异^[8]。在公司治理方面，投资者对企业创新、环境信息和社会责任的关注能够形成外部监督压力，推动上市公司改善信息披露和经营决策。

2.3 研究缺口

现有研究仍存在两点不足。第一，较多文献将关注作为整体变量处理，忽视价值、投机、政策等关注主题之间的异质性，难以揭示投资者关注背后的行为逻辑。第二，已有研究多考察关注对收益或企业行为的单向影响，缺少对不同关注主题之间动态互动关系的量化分析。基于此，本文利用文本大数据将投资者关注进行结构化分解，并通过VAR模型估计主题间的相互影响，以补充现有研究。

3 研究设计

3.1 语料库构建与预处理

本文以上交所E互动、深交所互动易和上市公司投资者关系活动记录为数据来源，覆盖2020年1月至2024年6月A股上市公司互动文本。经Python爬虫抓取后，共获得128.6万条文本数据，数据规模约5G。清洗环节剔除重复记录、无效问答、乱码文本和非投资者关系内容，并统一公司代码、日期、市场板块和投资者类型字段。文本处理环节以月度为单位汇总语料，依次完成jieba分词、词性过滤、停用词剔除和同义词归并，再用TF-IDF算法筛选出896个关键词，为主题提取和序列构造奠定基础。

3.2 统计测度方法

本文采用LDA模型提取投资者关注主题。综合困惑度、主题一致性与人工解释性，最终确定3类主题：价值主题、投机主题和政策主题。价值主题涵盖业绩、估值、分红、治理、研发等基本词汇；投机主题对应涨停、热点、概念、短线、题材等情绪交易词汇；政策主题包括注册制、监管、信息披露、规则、改革等制度词汇。各月主题关注占比由文本主题概率加权平均并标准化得到，差异化分析从投资者类型、市场板块、市值规模和账面市值比展开。

为考察三大主题的动态互动关系，本文构建VAR(p)模型。V、S、P分别表示第t期价值、投机和政策主题关注占比；p为最优滞后阶数； α_1 、 α_2 、 α_3 为常数项； $\beta_{i,j}$ 为滞后系数； ε_1 、 ε_2 、 ε_3 为随机扰动项。原文第3页公式截图已替换为以下可编辑公式：

$$\begin{cases} V_t = \alpha_1 + \sum_{i=1}^p \beta_{11,i} V_{t-i} + \sum_{i=1}^p \beta_{12,i} S_{t-i} + \sum_{i=1}^p \beta_{13,i} P_{t-i} + \varepsilon_{1t} \\ S_t = \alpha_2 + \sum_{i=1}^p \beta_{21,i} V_{t-i} + \sum_{i=1}^p \beta_{22,i} S_{t-i} + \sum_{i=1}^p \beta_{23,i} P_{t-i} + \varepsilon_{2t} \\ P_t = \alpha_3 + \sum_{i=1}^p \beta_{31,i} V_{t-i} + \sum_{i=1}^p \beta_{32,i} S_{t-i} + \sum_{i=1}^p \beta_{33,i} P_{t-i} + \varepsilon_{3t} \end{cases}$$

3.3 稳健性检验设计

为保证结论可靠，本文设置三类稳健性检验。第一，替换核心解释变量，将LDA主题概率占比替换为主题关键词词频占比。第二，调整样本区间，剔除2020年1—6月疫情初期异常观测值后重新估计模型。第三，改变模型形式，在Johansen协整检验基础上采用VECM模型检验长期均衡关系和短期调整效应。

4 实证结果与分析

4.1 投资者关注的结构性特征

表1 三大关注主题占比的描述性统计(单位：%)

关注主题	均值	标准差	最大值	最小值	2020年	2021年	2022年	2023年	2024年上半年
价值主题	32.3	3.1	38.9	25.7	27.5	29.8	31.2	33.7	36.8
投机主题	38.7	4.8	52.1	30.2	41.2	40.5	42.3	37.6	35.1
政策主题	29.0	2.3	34.6	24.8	31.3	29.7	26.5	28.7	28.1

三大关注主题的描述性统计如表1所示。投机主题平均占比为38.7%，在三大主题中最高，说明A股投资者短期热点和交易情

绪关注仍较强；价值主题均值为32.3%，且从2020年的27.5%升至2024年上半年的36.8%，表现出持续上升趋势；政策主题均值为29.0%，标准差最小，表明其整体波动较稳定但受重大制度事件影响。

HP滤波结果见表2。价值主题趋势项斜率为0.23，且在1%水平显著；投机主题趋势项斜率为-0.18，也在1%水平显著；政策主题斜率不显著。周期项上，投机主题振幅达到±5.3%，波动幅度明显高于价值主题和政策主题，说明投机关注更容易受短期行情和热点影响。

表2 三大主题趋势与周期特征量化结果

关注主题	趋势项斜率	P 值	周期项振幅	周期长度(月)	趋势特征
价值主题	0.23	0.008	±2.1%	12	显著上升
投机主题	-0.18	0.005	±5.3%	6	显著下降
政策主题	-0.03	0.621	±3.2%	不固定	基本平稳

关键制度节点前后比较如表3所示。创业板和科创板设立后，价值关注均有所下降，投机和政策关注上升，反映新兴板块初期市场预期更容易集中于短期机会和政策安排。全面注册制实施后，价值主题占比上升4.2个百分点，投机主题下降5.8个百分点，说明注册制改革对投资者关注向价值导向转移具有明显促进作用。

表3 关键节点前后三大主题占比变化(单位：%)

关键节点	时期	价值主题	投机主题	政策主题	变化幅度(价值/投机/政策)
创业板成立	成立前	35.2	34.8	29.0	-1.7/+1.3/+1.4
创业板成立	成立后	33.5	36.1	30.4	—
科创板成立	成立前	31.8	37.2	31.0	-3.9/+1.8/+2.1
科创板成立	成立后	27.9	39.0	33.1	—
全面注册制	实施前	32.6	41.0	26.4	+4.2/-5.8/-0.4
全面注册制	实施后	36.8	35.2	26.0	—

4.2 投资者关注的差异化特征

不同投资者类型的关注主题分布见表4。机构投资者价值主题占比为45.2%，比普通投资者高16.5个百分点；普通投资者投机主题占比为43.5%，比机构投资者高12.8个百分点。政策主题差异较小，说明机构和普通投资者对政策信息均保持关注，但二者在价值与投机之间的配置差异明显。

表4 不同投资者类型的关注主题分布(单位：%)

投资者类型	价值主题	投机主题	政策主题	样本(条)	差异说明
机构投资者	45.2	30.7	30.1	18.7万	价值关注较高
普通投资者	28.7	43.5	27.8	109.9万	投机关注较高
差异值	+16.5	-12.8	+2.3	—	差异明显

不同市场板块的关注结构见表5。主板价值主题占比最高，为35.6%；创业板投机主题占比最高，为41.8%；科创板政策主题占比最高，为33.2%。卡方检验结果为128.3，且在0.1%水平显著，说明板块属性会显著影响投资者关注结构。

表5 不同市场板块的关注主题分布(单位：%)

市场板块	价值主题	投机主题	政策主题	上市公司数(家)	卡方值
主板	35.6	37.1	27.3	1528	—
创业板	30.2	41.8	28.0	1215	—
科创板	29.7	37.1	33.2	542	—
合计	—	—	—	—	128.3***

企业特征也会影响关注主题。表6显示，市值规模越高，价值主题关注占比越高；市值规模越低，投机主题关注占比越高。价值主题和投机主题的F值分别为42.8和51.3，均在0.1%水平显著。表7显示，账面市值比越高，价值关注越强；账面市值比越低，投机关注越强，说明投资者会根据企业估值属性形成差异化关注。

表6 不同市值规模企业的关注主题分布(单位：%)

市值规模	价值主题	投机主题	政策主题	样本占比	F 值(价值)	F 值(投机)
高市值	38.9	32.2	28.9	1/3	—	—
中市值	32.1	38.5	29.4	1/3	—	—
低市值	27.3	45.7	27.0	1/3	—	—
统计检验	—	—	—	—	42.8***	51.3***

表7 不同账面市值比企业的关注主题分布(单位：%)

账面市值比	价值主题	投机主题	政策主题	样本占比	F 值(价值)	F 值(投机)
高(价值型)	36.4	35.2	28.4	1/3	—	—
中	32.5	38.6	28.9	1/3	—	—
低(成长型)	29.8	42.3	27.9	1/3	—	—
统计检验	—	—	—	—	38.6***	45.2***

4.3 三大关注主题的互动特征(基于VAR模型)

建模前，本文对价值、投机和政策主题序列进行ADF单位根检验。表8显示，三类主题原序列P值均大于0.05，不能拒绝单位根假设；一阶差分后P值均小于0.01，满足平稳性要求。

表8 VAR模型平稳性检验结果(ADF检验)

变量	检验类型	ADF 统计量	P 值	结论
V(价值主题)	(c, t, 2)	-1.23	0.67	不平稳
S(投机主题)	(c, t, 2)	-1.08	0.73	不平稳
P(政策主题)	(c, t, 2)	-0.95	0.78	不平稳
D(V)	(c, 0, 1)	-4.89	0.001	平稳
D(S)	(c, 0, 1)	-5.12	0.000	平稳
D(P)	(c, 0, 1)	-4.67	0.002	平稳

最优滞后阶数选择结果见表9。滞后2期时AIC、SC和HQ准则均达到最小，LR统计量也在0.1%水平显著，因此本文将VAR模型设定为VAR(2)。AR根检验表明所有根均位于单位圆内，模型满足稳定性要求。

表9 最优滞后阶数选择结果

滞后阶数	AIC	SC	HQ	LR 统计量
0	-2.15	-2.08	-2.12	—
1	-2.98	-2.81	-2.91	98.3***
2	-3.25	-3.01	-3.16	128.6***
3	-3.18	-2.87	-3.05	85.2***
4	-3.09	-2.71	-2.93	62.8**

VAR(2)模型核心估计结果见表10。三类主题均存在显著自滞后效应,其中政策主题滞后1期系数为0.51,持续性最强。价值与投机主题之间存在显著双向负向影响,说明价值关注提升会抑制投机关注,投机关注升温也会挤出价值关注。政策主题对价值主题为正向影响,对投机主题为负向影响,表明政策预期能够发挥关注引导作用。

表10 VAR(2)模型估计系数表(核心结果)

被解释变量	解释变量	滞后1期系数	滞后2期系数	t 值	P 值
价值主题(V)	V 滞后	0.32***	0.18**	4.56	0.000
价值主题(V)	S 滞后	-0.27***	-0.15*	-3.89	0.001
价值主题(V)	P 滞后	0.16**	0.09	2.34	0.021
投机主题(S)	V 滞后	-0.29***	-0.17**	-4.12	0.000
投机主题(S)	S 滞后	0.45***	0.22***	5.31	0.000
投机主题(S)	P 滞后	-0.14*	-0.08	-2.11	0.038
政策主题(P)	V 滞后	0.08	0.05	1.42	0.158
政策主题(P)	S 滞后	-0.09	-0.06	-1.57	0.121
政策主题(P)	P 滞后	0.51***	0.25***	6.02	0.000

脉冲响应结果见表11。价值主题冲击对投机主题产生短期负向影响,峰值为-0.23,持续8个月;政策主题冲击对价值主题产生长期正向引导作用,峰值为0.18,持续超过10个月;政策主题冲击对投机主题产生短期负向抑制作用,峰值为-0.15,持续7个月。

表11 脉冲响应核心结果汇总

冲击变量	被冲击变量	方向	峰值	峰值期数(月)	持续期(月)	经济意义
价值主题	投机主题	负向	-0.23	2	8	价值关注抑制投机
政策主题	价值主题	正向	0.18	3	>10	政策引导价值关注
政策主题	投机主题	负向	-0.15	2	7	政策抑制投机行为
价值主题	自身	正向	0.32	1	>10	价值关注持续
投机主题	自身	正向	0.45	1	>10	投机关注持续
政策主题	自身	正向	0.51	1	>10	政策关注持续性最强

表12 VAR模型方差分解结果(第10期)

被解释变量	自身贡献度	价值主题贡献度	投机主题贡献度	政策主题贡献度
价值主题(V)	62.3	—	21.5	16.2
投机主题(S)	58.7	23.8	—	17.5
政策主题(P)	89.6	4.2	6.2	—

方差分解结果见表12。政策主题波动89.6%由自身解释,说明政策关注具有较强独立性;价值主题波动由自身、投机主题和政策主题共同解释;投机主题波动也同时受价值主题和政策主题影响。这表明价值与投机之间存在相互约束,政策主题则通过引导预期影响市场关注风格。

4. 4稳健性检验结果

表13显示,将LDA主题概率占比替换为关键词词频占比后,核心系数符号和显著性未发生实质变化。价值与投机主题依然表现为双向负向关系,政策主题对价值主题保持正向影响,对投机主题保持负向影响。

表13 稳健性检验(替换核心解释变量): VAR模型核心系数

被解释变量	解释变量	滞后1期系数	P 值	滞后2期系数	P 值
价值主题(V)	S 滞后	-0.25***	0.000	-0.13*	0.082
投机主题(S)	V 滞后	-0.27***	0.000	-0.16**	0.035
价值主题(V)	P 滞后	0.14**	0.041	0.08	0.163
投机主题(S)	P 滞后	-0.12**	0.047	-0.07	0.189

表14显示,剔除2020年1—6月疫情初期异常数据后,政策主题自滞后效应仍显著,价值与投机主题的双向负向影响仍在0.1%水平显著,说明结论未受疫情初期特殊样本影响。

表14 稳健性检验(调整样本区间): VAR模型核心系数

被解释变量	解释变量	滞后1期系数	P 值	滞后2期系数	P 值
政策主题(P)	P 滞后	0.49***	0.000	0.23***	0.001
价值主题(V)	S 滞后	-0.26***	0.000	-0.14*	0.076
投机主题(S)	V 滞后	-0.28***	0.000	-0.15**	0.042

表15显示,采用VECM模型后,V与S的长期均衡系数为-0.24,V与P的长期均衡系数为0.15,S与P的长期均衡系数为-0.13,均与VAR模型结论一致。由此说明三类关注主题的互动机制不仅存在于短期动态中,也存在于长期均衡关系中。

表15 稳健性检验(改变模型形式): VECM模型核心系数

变量关系	长期均衡系数	P 值	短期调整系数	P 值
V 与 S 的协整关系	-0.24***	0.000	-0.18**	0.029
V 与 P 的协整关系	0.15**	0.032	0.11*	0.087
S 与 P 的协整关系	-0.13**	0.045	-0.09*	0.093

5 结论与建议

5. 1研究结论

本文基于2020—2024年沪深交易所互动平台文本,运用LDA模型识别投资者关注主题,并结合统计分析 with VAR模型考察主题互动机制。研究发现:A股投资者关注呈现“投机短期主导、价值稳步上升、政策事件驱动”的特征,全面注册制改革推动关注由投机向价值转移;不同投资者、板块和企业特征下关注存在显著差异,机构、高市值和高账面市值比企业更偏价值,普通投

资者及低市值企业更偏投机;政策主题具有引导作用,对价值关注产生长期正向影响,对投机关注产生短期抑制,价值与投机之间存在双向替代关系。

5.2政策建议

第一,强化政策引导稳定性。监管部门应提升政策沟通的连续性与可解释性,围绕注册制、信息披露、分红回报和公司治理等议题开展预期管理,引导投资者形成长期价值判断。第二,实施差异化监管。对创业板、低市值企业及短期题材集中领域加强异常关注监测,并对投机关注过热板块及时风险提示,同时发挥机构投资者的价值引领作用。第三,构建动态监测体系。将主题关注占比、主题替代关系和政策引导效应纳入风险监测,当投机关注挤出价值关注或政策引导弱化时,及时预警并开展投资者教育。

5.3研究局限与未来方向

本文仍存在一定局限:一是数据主要来自交易所互动平台,尚未纳入股吧、新闻评论、社交媒体等多源数据;二是研究重点在关注结构与主题互动,尚未进一步检验关注主题对收益率、波动率和流动性的直接影响。未来可结合多平台文本、深度学习主题模型和高频市场数据,进一步揭示投资者关注、市场行为与监管政策之间的传导路径。

[基金项目]

江西省高校人文社会科学研究项目阶段性成果(项目编号:JC24109)。

[参考文献]

[1]Simon H A.A behavioral model of rational choice[J].The Quarterly Journal of Economics,1955,69(1):99-118.

[2]Peng L,Xiong W.Investor attention,overconfidence and category learning[J].Journal of Financial Economics,2006,80(3):563-602.

[3]Da Z, Engelberg J,Gao P.In search of attention[J].The Journal of Finance,2011,66(5):1461-1499.

[4]Bouchaud J P,Krüger P,Landier A,et al.Sticky expectations and the profitability anomaly[J].The Journal of Finance,2019,74(2):639-674.

[5]朱红兵,王婉菁,张兵.投资者关注、信息不对称与名人持股效应[J].管理科学,2022,35(1):152-164.

[6]张学勇,唐国梅.行业关注度与股票横截面收益率——基于百度行业搜索指数的研究[J].经济学(季刊),2022,22(3):773-794.

[7]李雪婧,郑璐瑶,张永冀.投资者关注对信息披露质量的影响及机制——基于董秘履职视角[J].北京理工大学学报(社会科学版),2023,25(6):162-183.

[8]白兰,魏宇.投资者公共卫生事件关注度与我国行业股票市场信息溢出效应研究——来自TVP-VAR模型的经验证据[J].中国管理科学,2024,32(1):54-64.

作者简介:

吴华风(1979--),女,汉族,江西丰城人,本科,景德镇陶瓷大学实验员,主要从事管理学研究。