

专家审稿单

	投资者情绪与股票收益：网络社交媒体文本分析
外 审 意 见	1. 创新性的特点不够突出：个股的投资者情绪量化指标构建方法，并通过投资策略回测，通过分析互联网社交媒体信息，研究市场参与者的行为。这些已经在许多文献中论证过了，这里并没有显示出与众不同的特色。 2. 投资者关注度构建指标没有说明M_{it}^M是什么含义？ 3. 投资者意见趋同指数构建的依据是什么？ 4. 本文的主要研究对象为投资者情绪，将投资者关注与投资者意见趋同指数放进模型的意图是什么？ 5. 用投资者情绪预测股票收益的理论基础没有进一步说明； 6. 第六部分的量化投资策略的实证，似乎显得并不必要。 2024 年 1 月 18 日
作 者 回 复	首先诚挚感谢审稿老师与编辑部老师对我们论文评审的辛苦付出！衷心感谢审稿老师富有建设性的修改意见，这些建议极大提高了本文理论逻辑和实证研究的严谨性。我们根据审稿意见，逐条对论文进行了仔细修改与意见回复，具体说明如下，烦劳编辑和审稿人再次审阅。 1. 创新性的特点不够突出：个股的投资者情绪量化指标构建方法，并通过投资策略回测，通过分析互联网社交媒体信息，研究市场参与者的行为。这些已经在许多文献中论证过了，这里并没有显示出与众不同的特色。 在情绪的量化指标理论方法上，我们并没有创新。但我们认为，本文在情绪指标的构建和应用方面，仍然有一定的特色和创新之处。 第一，情绪指标的构建方面，本文使用作者自己标注的股吧评论数据作为训练集，且在对比后使用了更适用于金融领域评论情感分类任务的"Chinese-Word-Vectors"和 Bi-LSTM 模型进行情感分类，情感分类的准确率得到了提升。 第二，投建投资者情绪时，我们使用了股吧中 700 多万条帖子信息的数据。相比以往文献一般用 1-2 年样本期或者总体市场指数情况，我们的数据量更大，样本期更长。 2. 投资者关注度构建指标没有说明M_{it}^M是什么含义？ 我们在文中给予了说明，M_{it}^M表示股票 i 论坛中第 t 周的中立帖子数。 3. 投资者意见趋同指数构建的依据是什么？ 与构建情感指标一样，我们这里使用的是 Antweiler 和 Frank

	(2004)的方法构建投资者意见趋同指数，对此正文（p.10）进行了详细说明，“其中，$M_{i,t}^B$，$M_{i,t}^M$和$M_{i,t}^S$分别代表股票 i 论坛中第 t 周的看涨帖子数、中立帖子数和看跌帖子数。” 4. 本文的主要研究对象为投资者情绪，将投资者关注与投资者意见趋同指数放进模型的意图是什么？ 除了投资者看涨指数之外，我们还计算了投资者关注与投资者意见趋同指标，是因为投资者对股票的关注度和趋同度（即分歧度）在不同角度上也可以衡量投资者情绪。而且，以往文献表明，除了投资者看涨/看跌情绪之外，投资者对股票的关注度和分歧度也可能会影响到股票收益。例如，Kumar 和 Lee（2006）发现，市场关注度高的股票，投资者情绪和股票收益之间的关系更强。所以，本文研究的投资者情绪指标，包含三个维度：情感指标（看涨指数）、关注度指标和意见趋同指标（情感指标反映投资者看跌、看涨或者中立）。 5. 用投资者情绪预测股票收益的理论基础没有进一步说明。 在本文第一段，我们梳理了投资者情绪预测股票收益相关文献以及争议。我们写道：“投资者情绪如何影响股票价格？这是经济学家和金融业界长期争论的问题。至少自凯恩斯的‘动物精神（animal spirits）’概念提出开始，不少研究者便认识到……”由于理论上的争议，投资者情绪预测股票收益问题具有实证检验的价值。 6. 第六部分的量化投资策略的实证，似乎显得并不必要。 基于这条审稿意见，我们删除了关基于注度指标和意见趋同指标的投资策略业绩回测表现。不过，我们保留了基于投资者情感因子的投资策略业绩回测表现，只是将篇幅大幅度压缩，由原稿中的第六部分压缩为第五部分的一个小节。 这部分内容是为了论证我们所构建的情绪指标对股票收益率的预测能力，这可以展示该情绪指数在交易策略中的实际应用价值。基于投资者情感因子的投资策略业绩回测表现最佳，所以我们保留了这部分内容。
--	--

外审评议和作者回复（若有）是文章的过程性材料，隐名后的有价值讨论可公开。如研究中使用该内容，辛苦您在研究成果上注明引文和出处。

专家审稿单

	投资者情绪与股票收益：网络社交媒体文本分析
外 审 意 见	文章选题并不新，这是一个比较老的话题，不管是投资者情绪与股票收益，还是说基于网络社交媒体的文本分析。换言之，就是贡献稍显不足。文章仅选择了上证 50 成分股作为主要研究对象，这可能削弱了作者想突出表达个股情绪方面的结论。原因在于上证 50 通常是蓝筹股，大盘龙头股，占据核心地位，这也意味着它们的个股情绪和大盘整体情绪相关性很强。同时，这也意味着结果的背后具有较强的特质性，而未必具有普遍性。但一个可取之处是作者确实做了大量的工作，在方法上也有一定的改进。除此之外，还有以下建议： 1. 第四章中方程（5）写着的是RF_t，但对应的文中写的是R_{ft}，对不上号。此外，规模因子和价值因子使用的是全部 A 股的股票计算的，但是市场指数收益率选取的却是沪深 300 指数，因此计算的市场因子与其他两个因子的内涵不同，却放在同一个方程中，这不合理。并且无风险收益率采用的是 3 月期的 SHIBOR，依据是？ 2. 作者使用的是面板固定效应回归，但是从表 5 来看，仅固定了公司层面的，却忽略了年份的固定效应，这可能会导致结果有偏差。同时，根据当前做法的惯例，对应的固定效应最好是使用对应的聚类标准误，这在文中似乎没有得到体现。文中也未表明是一般标准误，还是稳健标准误，这也使得结果的有效性易被质疑。建议考虑不同的固定效应类型的组合，增强可信度。此外，多因子模型更通常的做法是使用 Fama-MacBeth 回归，然后使用特征变量来做面板回归进行稳健性的对照检验。文中仅使用面板回归分析，在做法上不符合资产定价文献的常规研究范式。 3. 作者使用了 PVAR 模型，但是未报告单位根检验、协整检验、滞后阶数的具体结果、稳定性检验、还有方差分解分析，以及对应的方程，仅报告了脉冲相应函数图，这使得该结果缺乏可信度。同时，作者在使用 PVAR 模型之前，阐述了“情绪因子和股票收益率可能同时受其他因素的影响——比如宏观经济指标、行业信息或公司个体信息，还可能是股价的变化引起了投资者情绪的变化。”然后接着写道为了预测性使用 PVAR 模型，这前后逻辑不通。并且，建议报告格兰杰因果检验的结果，而不是仅仅文字阐述。 4. 缺乏相应的稳健性检验，例如作者可以使用单变量组合分析法，这是传统的资产定价实证方法；以及上述说到的 Fama-MacBeth 回归进行相互印证；此外，还可以考虑增加更多的因子，例如五因子、四因子、反转因子等来进行稳健性检验。
作 者 回 复	首先诚挚感谢审稿老师与编辑部老师对我们论文评审的辛苦付出！衷心感谢审稿报告中富有建设性的修改意见，这些建议极大提高了本文理论逻辑和实证研究的严谨性。我们根据审稿意见，逐条对论文进行了仔细修改，具体说明如下，烦劳编辑和审稿人再次审阅。

	1. 第四章中方程（5）写着的是RF_t，但对应的文中写的是$R_{f,t}$，对不上号。此外，规模因子和价值因子使用的是全部 A 股的股票计算的，但是市场指数收益率选取的却是沪深 300 指数，因此计算的市场因子与其他两个因子的内涵不同，却放在同一个方程中，这不合理。并且无风险收益率采用的是 3 月期的 SHIBOR，依据是？ 衷心感谢评审老师指出这种表述的不一致性，我们在方程（5）的表述中已经改正。为了保证三个因子计算方法的一致性，市场指数因子也采用全部 A 股股票按照流通市值加权。本文使用的 Fama-French 三因子数据的详细说明见锐思数据库： https://db.resset.com/db/download/dataDictionary.jsp?tableName=THRFACDAT_WEEKLY。 关于无风险收益率，我们采用的是 3 月期的 SHIBOR，因为 SHIBOR 作为中国银行间拆借利率，在研究文献中一般被用作短期无风险利率。相比短期银行存款利率，SHIBOR 更加市场化，反映了金融机构自主定价能力。相比债券回购利率，SHIBOR 相对平稳，适合作为指导货币市场产品定价的基准。 2. 作者使用的是面板固定效应回归，但是从表 5 来看，仅固定了公司层面的，却忽略了年份的固定效应，这可能会导致结果有偏差。同时，根据当前做法的惯例，对应的固定效应最好是使用对应的聚类标准误，这在文中似乎没有得到体现。文中也未表明是一般标准误，还是稳健标准误，这也使得结果的有效性易被质疑。建议考虑不同的固定效应类型的组合，增强可信度。此外，多因子模型更通常的做法是使用 Fama-MacBeth 回归，然后使用特征变量来做面板回归进行稳健性的对照检验。文中仅使用面板回归分析，在做法上不符合资产定价文献的常规研究范式。 感谢评审老师指出这些问题，针对这些问题，我们进行了如下修改。 第一，改为使用双向固定效应回归模型（见表 5）。 第二，采用 Prais-Winsten 方法计算了面板矫正的稳健标准误（原稿中是一般标准误），见表 5。由于本文的样本为长面板数据，故而双向固定效应回归模型使用面板矫正的标准误(PCSE)似乎更为合适。 第三，使用 Fama-MacBeth 回归方法进行了稳健性检验（结果见表 5 第 4 列），同时在正文中给予了解释。 3. 作者使用了 PVAR 模型，但是未报告单位根检验、协整检验、滞后阶数的具体结果、稳定性检验、还有方差分解分析，以及对应的方程，仅报告
--	--

	了脉冲相应函数图，这使得该结果缺乏可信度。同时，作者在使用 PVAR 模型之前，阐述了“情绪因子和股票收益率可能同时受其他因素的影响——比如宏观经济指标、行业信息或公司个体信息，还可能是股价的变化引起了投资者情绪的变化。”然后接着写道为了预测性使用 PVAR 模型，这前后逻辑不通。并且，建议报告格兰杰因果检验的结果，而不是仅仅文字阐述。 根据这些意见，我们在正文中新增了格兰杰因果检验的结果（表 6）。我们尽管进行了单位根检验、单位根检验、方差分解等工作，但在正文中仅对基本结果进行了文字说明，并未报告具体数值结果。这主要是考虑到论文的篇幅限制，此外，经验而言，股市收益率系列一般不存在单位根，因此我们没有重点介绍协整检验和稳定性检验的结果。我们将 PVAR 模型中，变量单位根检验、协整检验、方差分解等具体结果列在本报告文后，谨供审阅（由于这些变量都不存在单位根，故不必进行协整检验、模型稳定性检验）。 4.缺乏相应的稳健性检验，例如作者可以使用单变量组合分析法，这是传统的资产定价实证方法；以及上述说到的 Fama-MacBeth 回归进行相互印证；此外，还可以考虑增加更多的因子，例如五因子、四因子、反转因子等来进行稳健性检验。 我们补充了 Fama-MacBeth 回归模型的结果作为稳健性检验（表 5）。不过，我们没有在模型中加入更多的因子，因为我们考虑到本文的研究目的不在于进行股票收益率的分解，而是在于研究投资者情绪因子的构建和预测价值。根据正文中已有的结果，在控制常用的 FF 三因子情况下，便可得到情绪因子的显著性结果以及显著的预测价值。考虑到集中研究主题不能太过发散的问题，我们没有在回归模型中再加入其它因子。
--	--

外审评议和作者回复（若有）是文章的过程性材料，隐名后的有价值讨论可公开。如研究中使用该内容，辛苦您在研究成果上注明引文和出处。

附录：

PVAR 模型中，变量单位根检验、模型滞后阶数选择和方差分解结果

为了防止出现伪回归的情况，本文采用了 Fisher-ADF 和 Fisher-PP 方法对本文的解释变量与被解释变量开展平稳性检验，结果见附表 1。由于 P 值均小于 1%，说明可以拒绝假定有单位根的原假设。因此，本文的变量均不存在单位根，即序列平稳。

附表 1 平稳性检验表（解释变量与被解释变量）

变量	Fisher-ADF		Fisher-PP		是否平稳
	Chi2 统计量	P 值	Chi2 统计量	P 值	
$R_{i,t}$	3160.1068	<0.001	3171.8415	<0.001	平稳
$BI_{i,t}$	2250.5959	<0.001	3159.2069	<0.001	平稳
$ATTI_{i,t}$	887.3741	<0.001	1440.3871	<0.001	平稳
$AI_{i,t}$	2726.1720	<0.001	3171.8415	<0.001	平稳

附表 2 是 PVAR 模型的最优滞后阶数选择结果。根据贝叶斯信息准则（BIC），我们设定 PVAR 模型的滞后阶数为 3。

附表 2 PVAR 模型最优滞后阶数选择

	AIC	BIC	HQIC
1	2.02748	2.15867	2.07175
2	1.92639	2.07326	1.97604
3	1.84874*	2.0124*	1.90418*
4	2.03092	2.2126	2.09259
5	1.96567	2.16593	2.03378

PVAR 模型的方差分解结果见附表 3。从中我们可以看到，股票收益率（ $R_{i,t}$ ）对自身的贡献率最大，从第 8 期开始稳定在了 81.5%。其次为投资者关注度因子（ $ATTI_{i,t}$ ），对其贡献率达到了 14.5%，代表投资者关注度变化在长期内对股票收益率变化有较大影响。

附表 3 方差分解结果表

	期数	$R_{i,t}$	$BI_{i,t}$	$ATTI_{i,t}$	$AI_{i,t}$
$R_{i,t}$	1	1	0	0	0
$R_{i,t}$	2	0.953	0.003	0.031	0.012
$R_{i,t}$	3	0.902	0.008	0.069	0.02
$R_{i,t}$	4	0.835	0.015	0.127	0.024
$R_{i,t}$	5	0.823	0.015	0.138	0.023
$R_{i,t}$	6	0.819	0.015	0.142	0.024
$R_{i,t}$	7	0.817	0.015	0.144	0.024

$R_{i,t}$	8	0.816	0.015	0.144	0.024
$R_{i,t}$	9	0.816	0.015	0.145	0.024
$R_{i,t}$	10	0.815	0.015	0.145	0.024

本文得到东湖高新区国家智能社会治理实验基地项目“金融资产定价与金融资产智能配置研究”的资助。感谢匿名审稿专家的宝贵意见，文责自负。

我们使用如下公式计算投资者对股票 i 在时期 t 内的看涨指数 (Bullishness Index):

$$BI_{i,t} = \ln[(1 + M_{i,t}^B)/(1 + M_{i,t}^S)] \quad (2)$$

其中, $M_{i,t}^B$ 和 $M_{i,t}^S$ 分别代表股票 i 股吧中第 t 期内的看涨帖子数和看跌帖子数。 $BI_{i,t}$ 反映了看涨贴和看跌贴的比例, 其值越大表明投资者看涨该股票的情绪越强烈。当看涨与看跌帖子数相等时, $BI_{i,t}$ 取值为 0。

投资者关注度指数 (Attention Index) 用于反映一段时间内投资者对某股票的关注程度。我们使用论坛中的发帖数量衡量投资者在第 t 周对股票 i 的关注度, 即:

$$ATTI_{i,t} = \ln(M_{i,t}^B + M_{i,t}^M + M_{i,t}^S) \quad (3)$$

其中, $M_{i,t}^B$, $M_{i,t}^M$ 和 $M_{i,t}^S$ 分别代表股票 i 论坛中第 t 周的看涨帖子数、中立帖子数和看跌帖子数。

投资者意见趋同指数 (Agreement Index) 用于反映对股票 i 第 t 周的投资者的意见一致程度:

$$AI_{i,t} = 1 - \sqrt{1 - \left(\frac{M_{i,t}^B - M_{i,t}^S}{M_{i,t}^B + M_{i,t}^S}\right)^2} \quad (4)$$

该指数反映了看涨和看跌帖子数量的差异程度, 取值范围为[0, 1]。当所有帖子均看涨, 或者所有帖子均看跌时, 均有 $AI_{i,t}=1$, 代表投资者的意见极端一致。而当一半帖子看涨而另一半帖子看跌时, $AI_{i,t}=0$, 此时投资者的意见分歧最大。

我们首先使用多因子模型, 考察投资者情绪与股票收益的关系。我们估计如下方程:

$$R_{i,t} = c + \beta(R_{m,t} - R_{f,t}) + sSMB_t + hHML_t + b_1BI_{i,t} + b_2ATTI_{i,t} + b_3AI_{i,t} + \varepsilon_{i,t} \quad (5)$$

其中, $R_{m,t} - R_{f,t}$, SMB_t 和 HML_t 分别为 Fama-French 三因子模型中的市场因子、规模因子和价值因子。 $R_{m,t} - R_{f,t}$ 表示市场指数 (全部 A 股股票按照流通市值加权计算) 收益率与无风险利率 (3 月期 SHIBOR) 之差, SMB_t 和 HML_t 的构建方法参考 Fama & French (1993)。¹

方程 (5) 是在三因子模型基础上新增了三个表示投资者有关个股情绪的自变量, 即上节中计算的投资者对股票 i 在时期 t 内的看涨指数 ($BI_{i,t}$)、关注度指数 ($ATTI_{i,t}$) 和意见趋

¹ 每年底将全部 A 股股票 (剔除 ST 股) 按照总市值排序, 市值较小的一半个股记为 S 组, 市值较大的一半个股记为 B 组; 将股票按账面市值比 (B/M) 排序, B/M 值最高的 30% 个股记为 H 组 (价值股), 最低的 30% 个股记为 L 组 (成长股), 中间 40% 记为 M 组 (平衡组)。以流通市值为权重, 构成六个组合: 小盘成长股 (收益率表示为 SL_t)、小盘价值股 (SH_t)、小盘平衡组 (SM_t)、大盘成长股 (收益率表示为 BL_t)、大盘价值股 (BH_t) 和大盘平衡组 (BM_t)。市值因子即小公司股票与大公司股票的收益率之差: $SMB_t = [(SL_t + SM_t + SH_t) - (BL_t + BM_t + BH_t)]/3$; 价值因子即价值股与成长股的收益率之差: $HML_t = [(SH_t + BH_t)2 - (SL_t + BL_t)]/2$ 。

同指数 ($AI_{i,t}$)。Fama-French 三因子均为市场层面变量, 与个股特征无关。根据有效市场假说, 资产的预期收益由系统性风险所决定, 而与异质性特征无关。如果有效市场假说成立, 则投资者有关个股情绪变量的系数应接近于零; 反之, 如果方程 (5) 中个股情绪变量的系数显著异于零, 意味着股票收益而不完全由系统性风险因子所决定, 而是与投资者情绪相关。

表 4 样本描述性统计

变量	观测数	均值	中位数	标准差	最小值	最大值
$R_{i,t}$ (%)	11086	0.084	0.008	1.148	-10.035	10.008
$BI_{i,t}$	11086	-0.2	-0.223	0.451	-1.981	1.897
$ATTI_{i,t}$	11086	5.988	5.966	0.973	2.303	9.924
$AI_{i,t}$	11086	0.03	0.016	0.037	0	0.413
$TO_{i,t}$ (%)	11086	7.258	5.129	6.943	0	71.395
$R_{m,t} - R_{f,t}$ (%)	252	0.107	0.155	2.337	-5.96	8.64
SMB_t (%)	252	0.132	0.2	1.693	-8.32	6.05
HML_t (%)	252	-0.023	-0.205	1.636	-4.29	6.27

注：本表数据为 2019—2023 年间周度数据，共 252 周。2023 年底的上证 50 指数成分股中，剔除 2019 年之后新上市的股票，共 44 只个股票。 $R_{i,t}$ 为股票 i 在第 t 周的收益率， $R_{m,t} - R_{f,t}$ 、 SMB_t 和 HML_t 为 Fama-French 三因子， $BI_{i,t}$ 、 $ATTI_{i,t}$ 和 $AI_{i,t}$ 分别表示投资者对股票 i 在第 t 周的看涨指数、关注度指数和意见趋同指数。 $TO_{i,t}$ 为股票 i 在第 t 周的换手率。

数据来源：论坛帖子来源为东方财富网股吧，股价和交易量数据来自 Wind 数据库，Fama-French 三因子数据来自锐思数据库。

表 5 多因子模型的双向固定效应回归结果

因变量： $R_{i,t}$	(1)	(2)	(3)	(4)
$BI_{i,t}$	1.618*** (0.032)		1.430*** (0.026)	1.303*** (0.051)
$ATTI_{i,t}$	0.016 (0.025)		0.012 (0.015)	0.046** (0.014)
$AI_{i,t}$	-0.307 (0.303)		-0.392 (0.260)	-0.791* (0.415)
SMB_t		0.213*** (0.005)	0.163*** (0.005)	
HML_t		-0.078*** (0.007)	-0.056*** (0.007)	
$R_{m,t} - R_{f,t}$		0.039*** (0.008)	0.017** (0.008)	
c	0.277 (1.130)	1.012* (0.538)	-1.504*** (0.537)	0.048 (0.076)
Company	Yes	Yes	Yes	No
Week	Yes	Yes	Yes	No

R^2	0.345	0.199	0.453	0.377
N	11086	11086	11086	11086

注：小括号内为 Prais-Winsten 方法计算的面板矫正稳健标准误，列(4)是使用 Fama-MacBeth 方法的估计结果。***、**、*分别代表在 1%、5%、10%水平下显著。其它说明见表 4。

表 6 格兰杰因果检验结果表

原假设	CHI2 统计量	P 值	结果
$BI_{i,t}$ 不是 $R_{i,t}$ 的格兰杰原因	9.9509	0.019	拒绝
$ATTI_{i,t}$ 不是 $R_{i,t}$ 的格兰杰原因	15.185	0.002	拒绝
$AI_{i,t}$ 不是 $R_{i,t}$ 的格兰杰原因	14.303	0.003	拒绝
$R_{i,t}$ 不是 $BI_{i,t}$ 的格兰杰原因	29.978	0.000	拒绝
$R_{i,t}$ 不是 $ATTI_{i,t}$ 的格兰杰原因	1.3461	0.718	接受
$R_{i,t}$ 不是 $AI_{i,t}$ 的格兰杰原因	24.893	0.000	拒绝

表 7 策略回测结果

		情感因子排序		关注度因子排序		意见趋同因子排序	
指标	基准组合	高	低	高	低	高	低
周度							
年化收益率 (%)	11.68	20.86	0.66	7.51	14.99	9.71	2.38
年化波动率 (%)	16.1	18.98	21	21.43	17.29	18.89	18.37
最大回撤 (%)	21.09	17.08	38.05	41.45	19.79	23.86	32.6
夏普率	0.65	1	0.04	0.35	0.79	0.48	0.11
月度							
年化收益率 (%)	11.22	17.79	7.24	5.21	17.66	11.39	2.05
年化波动率 (%)	16.04	18.86	24.46	22.77	17.01	22.78	16.79
最大回撤 (%)	21.12	15.99	33.5	41.07	15.57	22.99	29.27
夏普率	0.6	0.85	0.31	0.24	0.91	0.48	0.07

注：最大回撤率表示投资组合 p 的最大损失率，计算公式为： $\max [(D_i - D_j)/D_i]$ ， $\forall i < j$ ，其中， D_i 为投资组合在第 i 期净值。夏普率即组合超额收益率与波动率的比值，计算公式为： $[\overline{R_p} - R_f]/\sigma_p$ ，其中， $\overline{R_p}$ 为组合收益率均值， σ_p 为标准差， R_f 为无风险利率，以一个月期 Shibor 表示。