

基于 AI 大语言模型的叙事资产定价研究

樊嘉诚 林建浩 李宗余*

摘要: 本文利用 BERT 大语言模型提取新闻文本信息,并通过 Sparse IPCA 方法构建叙事资产定价模型。实证发现状态变量显著反映宏观预期、产出及股市收益,揭示了新闻叙事的经济信息价值。相比传统的股票特征因子,基于文本数据的叙事因子能更好地解释资产异象,并具有更强的样本外策略表现。研究进一步证明了大语言模型的定价效果优于传统的 LDA 主题模型,并探索其在可解释性、高频因子挖掘及与传统因子联合应用方面的潜力。

关键词: 横截面资产定价;大语言模型;文本分析

DOI: 10.13821/j.cnki.ceq.2026.03.07

一、引言

理解资产回报的横截面变动规律始终是资产定价研究的核心议题。Merton(1973)的跨期资本资产定价模型(Intertemporal Capital Asset Pricing Model, ICAPM)提供了经典理论框架,认为资产价格受某些潜在“状态变量”的影响,对这些变量的冲击构成了投资者面临的基本面风险,而资产对风险的暴露程度决定了收益率的横截面差异。在实证层面,如何有效识别状态变量和风险因素,存在着不同的观点和技术路线。早期研究以宏观经济变量作为 ICAPM 状态变量的代理(Cochrane, 1996; Rossi and Timmermann, 2015);随后,基于股票特征构建因子模拟组合成为主流(Fama and French, 1996、2016; Hou et al., 2015),并衍生出大量定价因子,形成了“因子动物园”(factor zoo)(Feng et al., 2020)。

近年来,Shiller(2017)提出的“叙事经济学”(Narrative Economics)为理解资产价格变动提供了新视角。他强调,经济叙事传播迅速,影响微观主体决策乃至改变经济和社会的走向。这一思想与基于风险(risk)的解释和基于错误定

* 樊嘉诚、林建浩,中山大学岭南学院;李宗余,广发证券股份有限公司。通信作者及地址:林建浩,广东省广州市海珠区新港西路 135 号中山大学岭南学院,510275;电话:020-84112104;E-mail:linjh3@mail.sysu.edu.cn。本研究得到国家自然科学基金面上项目(72273156、72073148),以及国家自然科学基金青年学生基础研究项目(723B2024)的资助。感谢《经济学》(季刊)编辑部和三位匿名审稿人,以及陈海强、姜富伟、杨学伟、唐国豪等专家学者对本文的宝贵意见。当然,文责自负。

价(mispricing)的解释两大资产定价理论支柱相呼应(Fama and French, 1996; Shleifer and Vishny, 1997)。一方面,新闻报道、社交媒体等文本作为叙事的主要载体,具有较强的时效性和前瞻性,能够及时反映经济状态或市场预期,从而影响投资者的风险感知(Larsen and Thorsrud, 2019; Bybee et al., 2024)。另一方面,叙事内容本身也包含与情绪波动和市场行为相关的大量信息(Gentzkow et al., 2019; 洪永森等, 2023),其传播可能引发投资者的过度反应或情绪化决策,使资产价格偏离基本面。尽管经济叙事与资产定价的理论联系较为清晰,但文本数据具有高维、非结构化和低信噪比等特征,在实践中如何高效提取叙事信息并将其合理映射到定价因子仍面临挑战。同时,在中国金融市场的特殊情境下^①,相关的资产定价实证研究依然十分有限。

本文利用 AI 大语言模型(Large Language Models, LLMs)从新闻媒体文本中提取叙事信息,在 ICAPM 框架下结合股票收益率构造叙事因子定价模型,以考察新闻叙事对我国 A 股市场的横截面定价效果。具体而言,本文收集 2007—2021 年七家专业财经媒体的新闻报道数据,通过 BERT(Bidirectional Encoder Representations from Transformers)预训练模型将非结构化文本表示为 768 维的嵌入向量,从而实现文本叙事的信息提取。然后,使用 Bybee et al. (2023)提出的 Sparse IPCA (Instrumented Principal Component Analysis)方法^②,将高维文本叙事主题作为条件信息嵌入 IPCA 因子模型(Kelly et al., 2019),并通过分组 LASSO 实现变量选择,以提取少数具有资产定价解释力的叙事风险因子,以及与经济基本面相关的状态变量。研究发现:第一,基于新闻文本估计的状态变量与多种宏观经济及金融市场指标相关,表明其可作为 ICAPM 理论中状态变量的有效代理。第二,基于 LLMs 的叙事因子模型的资产定价效果优于传统 Fama-French-Carhart 六因子模型(Fama and French, 1996; Carhart, 1997)^③;不仅对资产异象的解释力更强,其样本外交易策略也具有更高的年化夏普比率。第三,与传统词袋思想^④的 LDA(Latent Dirichlet Allocation)主题模型相比,LLMs 仅需更少因子即可获得良好的定价效果。

进一步地,本文拓展了 Bybee et al. (2023)的研究结果:一是将文本情绪纳

① 一方面,A股市场以散户投资者为主体,其专业知识和分析能力有限,往往依赖市场中传播的各类信息进行交易决策。因此,新闻报道作为信息传递的主要载体,发挥着重要作用。另一方面,中文作为一种高语境语言,其表达具有语义含蓄、一词多义等独特性,使得信息的有效提取更加困难。

② Sparse IPCA 在 IPCA 框架下引入稀疏化变量选择,利用可观测特征信息刻画资产的时变因子载荷,并通过分组 LASSO 在高维特征中进行稀疏化筛选与降维。

③ 该模型包括市场、规模、价值、盈利、投资和动量六大因子,是资产定价领域进行风险调整和异常收益检验的基础模型。

④ 词袋模型(Bag-of-Words, BOW)将文本表示为一个包含了所有词语的集合(即词袋),其中每个词语是独立的。但是,这种方法忽略了词语在文本中的顺序关系和词语本身的含义。此外,在大规模语料库中,词袋模型通常会生成一个非常稀疏的数值向量,将大大增加后续模型的复杂度和计算成本。

入基于 LDA 主题关注度的叙事因子模型,二是使用周度数据构造高频的叙事因子模型。上述分析使得模型能够更精准地捕捉到市场情绪波动和短期变化对资产价格的影响,从而提升定价表现。为了打开 LLMs 定价结果的“黑箱”,本文首先根据 LDA 模型识别出对我国股票市场具有重要影响的叙事主题,发现其呈现政策导向特征;然后,在由此类主题主导的子样本上训练基于 LLMs 的叙事因子模型,发现其能更好地解释资产异象并获得可观的夏普比率;最后,本文探讨了叙事因子与传统股票特征因子在联合应用方面的潜力,并进行稳健性分析。

本文的主要贡献有以下三个方面。第一,本文率先将 AI 大语言模型应用于横截面资产定价的实证研究,为利用文本数据乃至其他另类数据进行资产配置、投资策略构建等金融学研究议题拓展了思路。第二,本文考察新闻叙事对 A 股市场横截面收益率的影响,识别出关键的叙事主题,对理解我国股市随机折现因子(Stochastic Discount Factor, SDF)的构成具有参考价值。第三,本文从三个维度提供了 LLMs 优于传统文本分析方法的经验证据,包括捕捉上下文语义、理解复杂语言关系以及处理短文本三个方面。同时,本文结合 LLMs 与 LDA 两类文本分析方法,在一定程度上缓解了 LLMs 作为“黑箱”模型可解释性不足的问题,为后续研究提供了方法借鉴。

二、文献述评

(一) 文本分析与资产定价实证

现有研究普遍证实,文本数据(特别是新闻媒体报道)是影响资产价格的重要因素。相关实证工作主要分为两类:一类侧重于构造文本指标(情绪或关注度),并分析其对收益率的时序影响。^①例如,Tetlock et al.(2008)发现媒体情绪可预测股价波动,消极报道较多的公司股价和盈利更低。Ke et al.(2019)结合新闻情绪分析与有监督学习构造了一种交易策略,林建浩等(2022)进一步改进该方法并在沪深 300 指数成分股中验证其有效性。游家兴和吴静(2012)指出,新闻报道中的情绪会导致股票价格偏离基本价值;而 Dim et al.(2023)则发现,股票对媒体关注度变化的暴露程度越高,异质性风险越大、股价信息效率越低。另一类则聚焦于因子挖掘,利用文本信息解释或预测资产回报的横截面差异。代表性文献如 Bybee et al.(2023),他们使用 LDA 主题模型和 Sparse IPCA 方法,构建了基于叙事关注度的因子模型。Aleti and Bollerslev(2025)则

^① 相关文献综述可参考 Gentzkow et al.(2019)、沈艳等(2019)以及唐国豪等(2016)。

利用高频数据估计 SDF, 并通过词频计数法识别与其变动高度相关的新闻主题。除此之外, Engle et al.(2020)和 Faccini et al.(2023)基于文本构建气候变化指数, 研究气候风险对股票定价的影响。Liu and Matthies(2022)通过新闻报道衡量投资者对经济增长的关注, 发现其不仅能预测消费增长, 所衍生的单因子模型也优于 Fama-French 三因子模型。

然而, 现有将文本分析应用于资产定价实证的研究仍存在两方面局限。一是这些研究仍依赖传统的词袋模型进行文本数据建模, 未充分探讨大语言模型在测度文本叙事和构建资产定价模型中的应用潜力。二是这些研究主要聚焦于美国市场, 对中国市场的关注相对不足, 尤其缺乏将文本数据与横截面因子定价相结合的系统研究。本文旨在弥补上述不足, 探索大语言模型在资产定价中的应用, 并检验其在中国市场的适用性。

(二) 大语言模型的优势与应用

传统文本分析方法通常依赖分词技术将文本转化为数值型词频矩阵, 但存在三方面局限: 一是仅保留词频信息, 忽略语序与语义结构, 对同义词替换稳健性差; 二是词频矩阵高维稀疏, 统计效率低; 三是跨领域泛化能力弱, 难以适应文本更新与新词出现。相比之下, 大语言模型(LLMs)采用更细粒度的子词单元, 缓解了数据稀疏问题并提升了通用性; 同时可引入特定领域的文本数据, 通过“进一步预训练”增强专业文本分析能力。大量研究验证了 LLMS 的优势: Jin et al.(2020)通过对抗攻击实验表明, LLMS 在保持语义、语法和分类准确性方面具有更强稳健性; Leippold(2023)在金融文本分析场景下佐证了这一结论。Garrido-Merchan et al.(2023)则在多项文档分类任务中显示, BERT 模型优于传统词频-逆文档频率方法。

得益于其强大的文本分析能力, LLMS 正被日益广泛应用于经济金融研究中。例如, 刘青和肖柏高(2023)通过 BERT 模型识别专利文本中的劳动节约型技术创新, 金星晔等(2024)也采用类似方法测度企业数字化转型。在金融学领域, 多项研究借助 LLMS 提取文本信息以预测股票收益。Chen et al.(2022)和 Tan et al.(2023)利用 LLMS 提取新闻文本的数值表征, 并应用于股票回报预测。实证结果表明, 在不同市场和多种语言环境中, 新闻信息的价格吸收存在一定滞后, 而这一特征可被利用于构建实时的交易策略。姜富伟等(2024)训练了一个中文金融大语言模型, 并开展金融市场情绪测度和资产价格风险预测。

以上研究虽已将 LLMS 应用于金融市场分析, 但也存在一定局限性: 一是多聚焦于文本情绪测度和时间序列预测, 尚未挖掘 LLMS 在横截面资产定价中

的作用。二是尚未系统比较 LLMs 相较于传统文本分析方法的优势来源,也未充分揭示哪些文本信息在资产定价中发挥关键作用及其经济机制。本文补充了上述研究,探讨 LLMs 在横截面因子定价中的应用,并评估其相较于传统文本分析方法的增量价值。

三、模型构建

(一) 文本模型

1. 大语言模型

本文使用在 BERT 中文模型的基础上通过 CoSENT(Cosine Sentence)方法进一步微调得到的预训练模型,参数量约 4 亿,输出 768 维实数向量。^① 借鉴 Peters et al.(2018)的做法,本文直接利用该模型生成文档层面的数值表征(即嵌入向量),而不再对其做任何改动(包括参数、输入输出设定等)。由于 BERT 类模型对输入文本的长度有限制,为了充分利用每篇新闻的全部信息,本文先对文本进行“段落”分割处理,然后通过标题主导各个段落嵌入向量的语义聚合,从而得到整篇文档的数值表征。

2. LDA 模型

为与 LLMs 对比,本文采用 Blei et al.(2003)提出的 LDA 主题模型。该模型假设文档由若干主题混合构成,每个主题对应一个词汇概率分布。在模型训练前,首先对新闻文本进行预处理:使用 Python 软件包 jieba 进行中文分词,过滤数字、英文、标点及停用词,保留长度不少于两字且至少在 50 篇文章中出现过的词汇,进而构建文档-词频矩阵。本文采用变分期望最大化(Variational Expectation-Maximization, VEM)算法估计包含 30 个主题的 LDA 模型,以文档主题分布作为叙事信息,关键词分布描述主题内容。^②

3. 叙事新息

令 $D_{j,m,\tau}$, $j=1, \dots, N_{m,\tau}$ 表示媒体 m 在第 τ 天发布的第 j 篇新闻文本,其中 $N_{m,\tau}$ 为该媒体在当天的发文总量。通过 BERT 或 LDA 模型生成该文档的数值表征,得到 L 维^③向量 $\theta_{j,m,\tau}$ 。在构造日度叙事信息 θ_τ 时,考虑到不同媒体

^① BERT 模型最早由 Devlin et al.(2019)提出,谷歌公司(Google)在其算法基础上推出了中文版本的预训练模型,哈工大讯飞联合实验室(HFL)对其进行微调训练。本文使用的 BERT-CoSENT 模型可以在 Hugging Face 社区官网公开获取(<https://huggingface.co/shibing624/text2vec-base-chinese>)。本文采用该模型生成文档级向量表示,附录 I 对其适用性和应用细节进行了补充说明。限于篇幅,附录未在正文列示,感兴趣的读者可在《经济学》(季刊)官网(<https://ceq.ccer.pku.edu.cn>)下载。

^② 有关 LDA 模型的数学描述与估计细节,Larsen and Thorsrud(2019)和 Bybee et al.(2024)等文献已作详细介绍,本文不再赘述。附录 I 报告了 LDA 模型主题数量的选择过程和估计结果。

^③ 在 BERT 模型中, $L = 768$;在 LDA 模型中, $L = 30$ 。

的发量差异,本文先在媒体层面对当天的所有新闻取均值: $\theta_{m,\tau} = N_{m,\tau}^{-1} \times \sum_{j \in m} \theta_{j,m,\tau}$; 然后,以当日发量在总文章数的占比为权重,对媒体层面的叙事信息进行加权平均: $\theta_\tau = (\sum_{m \in \tau} N_{m,\tau})^{-1} \sum_{m \in \tau} N_{m,\tau} \theta_{m,\tau}$ 。最后,沿用文献的常见做法,对 θ_τ 减去过去 5 天的移动平均得到叙事新息(Narrative Innovations) z_τ :

$$z_\tau \triangleq \theta_\tau - \frac{1}{5} \sum_{l=1}^5 \theta_{\tau-l}. \quad (1)$$

(二) 收益率模型

本文使用 Bybee et al.(2023)的方法,利用新闻文本和股票收益率构造状态变量和定价因子。具体而言,假设状态变量 x_τ 为 $K \times 1$ 向量($K \ll L$), z_τ 通过一个 $L \times K$ 矩阵 $\tilde{\mathbf{A}}$ 与 x_τ 相关联:

$$z_\tau = \tilde{\mathbf{A}} x_\tau + \eta_\tau, \quad (2)$$

其中, η_τ 表示叙事新息中与状态变量无关的部分。对式(2)的一种解释是投资者会基于新闻文本形成对当前状态的评估,进而指导他们对未来经济状况的预期和消费投资决策。

针对股票超额收益的横截面结构,假设其遵循条件因子定价模型:

$$r_{i,t+1} = \beta_{i,t} f_{t+1} + \epsilon_{i,t+1}, \quad (3)$$

其中, $r_{i,t}$ 为股票超额收益, $K \times 1$ 向量 f_t 为因子模拟组合的收益率,即本文所指的叙事因子(Narrative Factor, NF); $\beta_{i,t}$ 为时变的因子暴露, $\epsilon_{i,t}$ 为误差项; 下标 $i=1,2,\dots,N$ 表示股票, $t=1,2,\dots,T$ 表示时间^①。ICAPM 理论表明,上述横截面方程的截距为零(即不存在定价误差),误差 $\epsilon_{i,t}$ 的均值为零且与 f_t 正交。

状态变量 x_t 决定了因子暴露 $\beta_{i,t}$ 。文献中估计可观测 x_t 的因子暴露通常使用 Fama-MacBeth 回归,但本文的 x_t 不可观测。因此,参考 Bybee et al.(2023)的做法,将状态变量 x_t 在资产超额收益空间 f_t 上做投影,投影残差 $\nu_t \triangleq x_t - f_t$ 与所有超额收益正交。通过式(2)得到:

$$z_t = \mathbf{A} f_t + g_t, \quad (4)$$

其中, g_t 为由 ν_t 和 η_t 生成的复合残差,同样正交于超额收益;系数矩阵 \mathbf{A} 刻画了叙事因子 f_t 与叙事新息 z_t 之间的关系。由于 f_t 不可观测,将式(4)代入 $\text{Cov}(r_{i,t+1}, z_{t+1})$, 定义 $\sum_{ff} \triangleq \text{Cov}_t(f_{t+1})$, $\text{cov}_{i,t} \triangleq \text{Cov}_t(r_{i,t+1}, z_{t+1})$, $\tilde{\mathbf{F}} \triangleq \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \sum_{ff}^{-1}$, 便得到 IPCA 估计方程:

^① 下标 t 和 τ 均表示时间,其中 t 表示月度而 τ 表示日度。由于因子模型的数据频率一般为月度,因此在收益率模型部分,使用月度下标 t 以便与实际数据相匹配。

$$r_{i,t+1} = \text{cov}_{i,t} \tilde{\Gamma} f_{t+1} + \varepsilon_{i,t+1}, \quad (5)$$

其中, $r_{i,t}$ 可观测, 协方差 $\text{cov}_{i,t}$ 可用样本数据估计, 因而可以用 IPCA 方法估计 $L \times K$ 的系数矩阵 $\tilde{\Gamma}$ 和因子收益率 f_t 。^①

进一步地, 因子暴露 $\beta_{i,t}$ 的估计方程为:

$$\beta_{i,t} \triangleq \text{cov}_{i,t} \tilde{\Gamma}. \quad (6)$$

联立式(2)和式(5)可以得到状态变量的估计:

$$x_t = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T z_t = \sum_{ff} \tilde{\Gamma}^T z_t. \quad (7)$$

值得注意的是, IPCA 方法识别的叙事因子和状态变量存在旋转不确定性 (Rotational Indeterminacy)。但经过均值方差有效 (Mean-Variance Efficient, MVE) 组合后, 因子和状态变量唯一确定且经济含义明确, 即定价核 (Pricing Kernel) 或随机折现因子 (SDF)。为兼顾计量可识别性和经济可解释性, 后文分析中的“叙事因子”和“状态变量”均转换为基于原始序列的 MVE 组合。具体地, 对于一组叙事因子 f_t 和状态变量 x_t 的估计值, 其 MVE 组合的权重向量为

$$w_{MVE} = \sum_{ff}^{-1} \hat{\mu}_f, \quad (8)$$

其中, $\hat{\mu}_f$ 和 \sum_{ff} 为 f_t 估计值的样本均值和样本协方差矩阵, 对应的定价核分别为 $f_t^{MVE} = w_{MVE}^T f_t$ 和 $x_t^{MVE} = w_{MVE}^T x_t$ 。

四、实证分析

(一) 数据说明

在新闻文本数据方面, 本文参照张一帆等 (2023) 的标准, 选择目前在我国具有广泛影响力且知名度和权威性较高的七家专业财经报纸作为媒体池^②。我们通过慧科新闻搜索引擎^③, 全文检索含“经济”关键词的报道并下载原始文本。基于 ICAPM 理论框架, 研究旨在从文本中识别宏观状态变量与风险因素, 因此根据“审议”“公示”“公告”“披露”等关键词, 剔除了与个股高度相关的新闻。在股票价格和因子数据方面, 本文选取沪深 A 股主板和创业板的股票作为研究样本, 并按中国证监会《上市公司行业分类指引 (2012 年修订)》剔除金融业股票和

^① 限于篇幅, 模型估计的细节见附录 II。由于叙事新息具有高维性, 但并非所有维度都与资产收益率相关。因此, 我们通过分组 LASSO 惩罚项引入稀疏性, 而超参数根据最大化因子模型的夏普比率调优。

^② 具体包括《证券日报》《证券时报》《金融时报》《第一财经日报》《21 世纪经济报道》《中国经营报》和《每日经济新闻》, 涵盖了官方媒体和市场化媒体两类。

^③ 慧科新闻搜索引擎 (WiseSearch) 是慧科讯业 (Wisers) 提供的媒体资讯检索平台, 支持基于关键词检索相关媒体报道。

所有 ST 股票。参照 Liu et al.(2019)的做法,进一步剔除新股上市前 6 个月、上一年交易少于 120 天和上个月交易少于 15 天的样本。股票市场数据来自 Wind,将股价转换成对数收益率。对于传统的基准因子模型,本文考虑了 CAPM、Fama-French 三因子模型(FF3)、五因子模型(FF5)及包含 Carhart 动量因子的六因子模型(FFC6)。相关数据来源于 CSMAR。本文以月度为周期进行主要的定价检验,在拓展分析中进一步考察高频(周度)模型的定价效果。最终,样本包括 2007 年 1 月至 2021 年 12 月的新闻文本共 360 488 篇,涵盖 3 426 家上市公司。

(二) 状态变量检验

本文基于全样本数据估计叙事因子模型。具体而言,以 1 年为预热期(burn-in)初始化参数,然后利用剩余样本进行模型估计。与 Bybee et al.(2023)保持一致,设定叙事因子维度 $K = 1, 2, \dots, 6$ 。

首先,我们对叙事因子模型识别的状态变量进行经济属性分析。根据 ICAPM 理论,状态变量包含投资者未来投资机会集变化的相关信息,因而与经济基本面有着密切联系。表 1 报告了状态变量的定价核(x^{MVE})与宏观经济和金融市场指标的时序相关性^①。结果显示,基于 LLMs 的叙事因子模型(NF-LLM)所提取的状态变量定价核 x_{LLM}^{MVE} 与消费者信心指数和股市收益率具有较强的正相关关系,与产出、通胀等指标也均具有一定的正向关联。初步表明状态变量具有顺周期特征: x_{LLM}^{MVE} 的增加伴随着信心提升、经济扩张和金融市场繁荣,这符合 ICAPM 的理论预测。相比而言,尽管 NF-LDA 模型的状态变量 x_{LDA}^{MVE} 也呈现一定的顺周期性,但其与基本面变量的相关性略低于 x_{LLM}^{MVE} ,说明 LLMs 在提取与经济相关的叙事信息方面更具优势。

表 1 状态变量定价核与宏观、金融指标的时序相关系数

经济指标(Var)	Corr(x_{LLM}^{MVE}, Var)		Corr(x_{LDA}^{MVE}, Var)	
	疫情前	全样本	疫情前	全样本
消费者信心指数	0.30	0.41	0.13	0.28
规模以上工业增加值同比	0.10	0.07	0.01	0.03
社会消费品零售总额同比	0.08	0.03	0.02	0.05
房地产开发投资同比	0.09	0.01	0.01	-0.02
消费者价格指数同比	-0.04	-0.07	-0.01	-0.10
生产者价格指数同比	0.17	0.23	0.13	0.18

^① 由于中国的产出、消费和投资等宏观指标具有明显的趋势特征,我们参考张一帆等(2023)的做法对其进行去趋势处理。附录 III 展示了状态变量的时间序列和其他检验结果。

(续表)

经济指标(Var)	Corr(x_{LLM}^{MVE}, Var)		Corr(x_{LDA}^{MVE}, Var)	
	疫情前	全样本	疫情前	全样本
上证综指收益率	0.18	0.16	0.03	0.01
深证成指收益率	0.14	0.14	0.01	0.01
10年期国债收益率变化	0.09	0.06	0.21	0.15

注:为减少新冠疫情冲击导致的数据结构性变化对实证结果的影响,本文以2008年1月至2021年12月为全样本期,并进一步选取2008年1月至2019年12月作为疫情前样本进行稳健性检验。由于疫情期间消费、投资、工业增加值等宏观变量的同比增速出现突发性下滑,本文采用斯皮尔曼秩相关系数计算变量相关性,以在一定程度上降低离群值的影响。

(三) 资产异象检验

本文采用李斌等(2019)构造的96个A股异象^①(包括交易摩擦、动量、价值和财务流动性等),考察模型对这些资产异象的解释能力。表2报告了叙事因子模型(NF-LLM和NF-LDA)、基准因子模型以及统计因子模型(记作PC)^②的全样本检验结果。表2 Panel A结果显示,NF-LLM模型的定价表现随因子数 K 增加而持续改善(如平均定价误差由0.15%降至0.02%,GRS统计量由6.91降至2.04),表明更多因子有助于充分利用LLMs提取的叙事信息。相比NF-LDA模型(Panel B),在 K 较低时两者对异象的解释力相近,但当 $K \geq 5$ 后,NF-LLM模型显示出明显优势。上述结果说明,LLMs嵌入向量相较于LDA主题分布的信息含量更加丰富,有助于提升叙事因子定价模型的定价效果。

与基准因子模型(Panel C)相比,叙事因子模型表现更优:NF-LLM(NF-LDA)六因子模型的异象显著比例较FFC6模型低约30%(12%),GRS统计量远低于FFC6的6.14。与统计因子模型(Panel D)对比发现,NF-LDA与PC模型解释资产异象方面的表现整体相当,而NF-LLM模型则展现出额外的增量信息。综合以上分析,基于LLMs的叙事因子相较于Fama-French因子具有更好的异象解释能力,并且相比传统文本分析方法(LDA)构造的因子,也表现出一定的边际提升。更重要的是,叙事因子模型对资产定价能力的改进,不仅在于统计方法的优化,更在于文本信息本身所蕴含的经济价值。

^① 由于数据可获得性等客观限制,资产异象检验部分采用的数据集仅覆盖至2018年11月,因此表2的样本期与其他部分存在差异。异象数据可以在《中国工业经济》网站获取附件。

^② 本文参考Kozak et al.(2018)的做法,采用主成分分析(Principal Component Analysis, PCA)方法从股票收益率中提取共同因子,并构造统计因子模型进行对比分析。感谢匿名审稿专家提出对比统计因子模型的宝贵建议。受篇幅限制,PC模型的构造方法和定价检验详见附录IV。

表2 因子模型对96个异象的全样本检验结果

Panel A: NF-LLM				
K	平均 α (%)	平均 $ t(\alpha) $	显著比例 (%)	GRS
1	0.15	1.58	22.92	6.91
2	0.13	1.29	17.71	5.11
3	0.08	1.24	19.79	3.63
4	0.07	1.23	18.75	2.62
5	0.01	0.88	10.42	2.07
6	0.02	0.86	9.38	2.04
Panel B: NF-LDA				
K	平均 α (%)	平均 $ t(\alpha) $	显著比例 (%)	GRS
1	0.15	1.58	22.92	6.91
2	0.13	1.21	17.71	4.57
3	0.14	1.16	15.62	4.25
4	0.14	1.17	15.62	4.06
5	0.15	1.20	16.67	4.11
6	0.15	1.16	17.71	3.87
Panel C: 基准因子模型				
因子模型	平均 α (%)	平均 $ t(\alpha) $	显著比例 (%)	GRS
CAPM	0.15	1.52	21.88	6.55
FF3	0.06	1.86	31.25	6.26
FF5	0.08	2.10	38.54	6.28
FFC6	0.08	2.27	39.58	6.14
Panel D: 统计因子模型				
因子模型	平均 α (%)	平均 $ t(\alpha) $	显著比例 (%)	GRS
PC1	0.24	2.02	38.54	5.41
PC2	0.99	3.60	57.29	10.27
PC3	1.08	2.04	39.58	4.01
PC4	0.32	1.29	22.92	2.86
PC5	0.14	1.12	15.62	2.30

注:PC1 仅包含第一个主成分因子,PC2 包含前两个主成分因子,以此类推。第一列报告所有异象资产的平均定价误差 α ,第二列为其 t 值绝对值的平均值;第三列统计定价误差在 95%水平下显著的异象比例;最后一列为 GRS 统计量,用于检验 96 个异象的定价误差是否联合为 0。样本期为 2008 年 1 月至 2018 年 11 月。

(四) 交易策略表现

进一步地,本文从样本外交易策略的角度探究叙事因子模型是否能获得显著的超额回报。为此,我们通过滚动窗口,用样本内训练得到的因子暴露 $\hat{\beta}_{i,t}$ 来构建可交易的因子模拟组合(代表样本外的因子收益率),并考察这些投资组合的夏普比率。^①具体地,第 $t+1$ 月样本外因子收益率表示为:

$$f_{t+1}^{OOS} = \left[\left(\sum_i \hat{\beta}_{i,t} \hat{\beta}_{i,t}^T \right) + 2 \mathbb{I}_K \right]^{-1} \sum_i \hat{\beta}_{i,t} r_{i,t+1}, \quad (9)$$

其中, $\hat{\beta}_{i,t} = c_{it} \hat{\Gamma}$, $\hat{\Gamma}$ 由训练集样本估计。令 $\hat{\mu}_f$ 和 $\hat{\Sigma}_{ff}$ 为样本内因子收益率的均值和协方差矩阵,则样本外因子 MVE 组合的收益率为:

$$f_{t+1}^{OOS-MVE} = \hat{\mu}_f^T \hat{\Sigma}_{ff}^{-1} f_{t+1}^{OOS}. \quad (10)$$

表 3 汇报了各因子模型的样本外年化夏普比率。整体而言,NF-LLM 在多数设定下均优于 NF-LDA 模型。具体来看,最优 NF-LLM 模型的夏普比率达 0.95,显著高于最优 NF-LDA 模型(0.88);且 NF-LLM 在因子数较少时($K=2,3$)已表现优异,说明其信息提取效率更高。与传统基准模型相比,叙事因子模型在实时策略中优势明显:最优 NF-LLM 的夏普比率(0.95)是最优 FFC 模型(0.27)的 3.5 倍。即使与统计因子模型相比,叙事因子仍具竞争力:NF-LLM 双因子模型已优于最优 PC 模型,而 NF-LDA 的夏普比率最高是其两倍。这一优势可能源于新闻文本能及时且广泛地捕捉市场情绪、政策与宏观动态,从而提升样本外策略收益。

表 3 因子模型的样本外年化夏普比率

Panel A: NF-LLM						
K	1	2	3	4	5	6
NF-LLM	0.06	0.68	0.86	0.90	0.86	0.95
Panel B: NF-LDA						
K	1	2	3	4	5	6
NF-LDA	0.06	0.27	0.34	0.88	0.81	0.74
Panel C: 基准因子模型						
CAPM	FF3	FF5	FF6			
0.18	0.04	0.24	0.27			

^① 初始训练样本为 2008 年 1 月至 2010 年 12 月。为降低计算负荷,模型于每年末重新训练,并构建未来一年的样本外因子收益率。本文对基准因子和统计因子也使用相同方法构建样本外因子收益率及 MVE 组合收益率,以确保可比性。

(续表)

Panel D: 统计因子模型				
PC1	PC2	PC3	PC4	PC5
0.43	0.20	-0.01	0.28	0.40

注：样本期为 2008 年 1 月至 2021 年 12 月。

(五) 稳健性检验

本文对上述资产定价检验结果进行稳健性分析(相关结果见附录 III)。第一,我们使用其他因子模型(如中国版四因子^①)为基准对比叙事因子模型的定价效果,发现 NF-LLM 模型仍能获得显著的超额收益和更高的夏普比率,而 NF-LDA 模型的增量信息有限。第二,为控制新闻文档写作风格、报道偏好等非经济因素对 LLMs 嵌入向量的潜在影响,本文借鉴 Shapiro et al.(2022)的思路,提取嵌入向量的时间固定效应作为叙事信息度量。结果显示,NF-LLM 模型仍保持稳健定价能力,部分情形下甚至优于基准。第三,在不同 LDA 主题数设定下,本文主要结论均未发生实质性变化。

五、进一步分析

(一) 大语言模型的优势

1. 统计性证据

如前所述,LLMs 相较传统文本分析方法的优势在于其强大的语义理解能力和通用性。为验证这一点,本文从统计角度进行了三方面检验。首先,我们检验上下文语义关系对定价效果的影响。借鉴词袋模型思路,将新闻文本切分为无顺序、无语义关联的关键词组合,再对每个词进行嵌入表示,并在文档层面取平均以构建“NF-LLM(词嵌入平均)”模型。图 1(a)显示,在绝大多数因子数设定下,基准 NF-LLM 模型的样本外年化夏普比率均高于 NF-LDA 和词嵌入平均模型,说明保留上下文语义关系的 LLMs 能有效提升定价表现。即使不考虑语义结构,词嵌入平均模型因信息维度更高,在低因子数下仍优于 NF-LDA。

其次,本文考察 LLMs 在处理复杂语言关系和不同长度文本的能力。^②

^① Liu et al.(2019)针对中国 A 股市场构造了中国版四因子模型,包括市场、规模、价值和换手率因子。其中,规模因子在构造时剔除了最小 30% 股票以降低壳价值干扰,价值因子基于盈利市值比构造,换手率因子则用于刻画中国市场中由异常换手率反映的投资者情绪或交易特征。

^② 以高否定性语境文本的识别为例:基于中国知网构建的否定词词典统计每篇新闻的否定词数量,选取日度样本中否定词数高于中位数的文本作为高否定性语境子样本。高模糊性语境文本采用 You et al.(2018)提供的包含模糊性词汇的词典识别,短篇幅文本则以文本预处理后的总词数为判断标准。然后,在子样本中分别训练 NF-LLM 和 NF-LDA 模型并构造样本外交易策略。

图1(b)显示,在全部文本中,NF-LLM的夏普比率较NF-LDA高约7.95% $(= (0.95/0.88-1) \times 100\%$,数值对应于表3)。在高否定性语境的文本中,其优势进一步扩大至18.39%。在高模糊性语境下,NF-LDA表现有所下降,而NF-LLM仍保持稳健,较之提升10.13%。在篇幅较短的新闻中,NF-LDA的定价能力显著减弱,而NF-LLM仍保持较高夏普比率,性能提升达20.27%。上述结果表明,LLMs不仅能有效理解复杂语义关系,并且在信噪比更低的短文本中仍能有效提取定价信息。

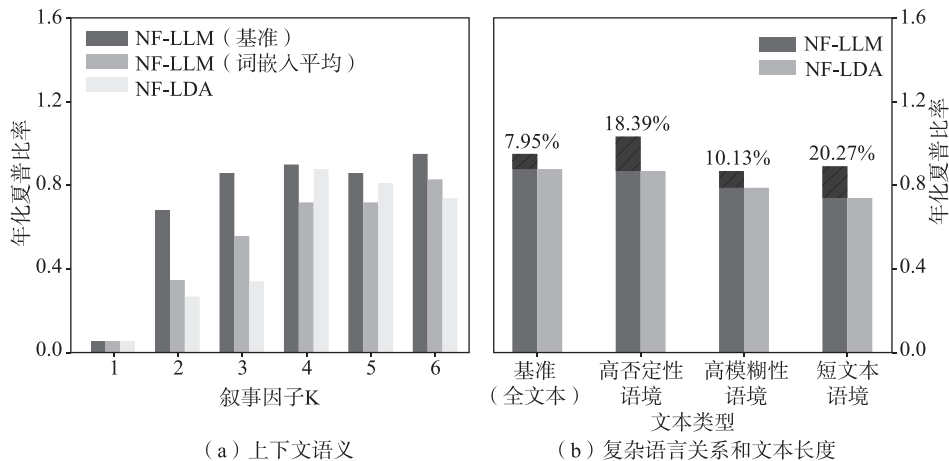


图1 NF-LLM 优于 NF-LDA 的统计性证据

注:图1(b)中柱状图顶部的黑色阴影区域表示NF-LLM相较于NF-LDA的年化夏普比率提升,柱上百分比为对应的提升幅度。

2. 情绪的作用

Shiller(2017)认为经济叙事尤其与情绪相关,而文本等非结构化大数据中包含了大量与经济主体预期和情绪相关的信息。大量金融学研究表明,情绪信息对资产价格具有重要影响,故而有必要将文本情绪纳入叙事因子模型。借鉴Larsen and Thorsrud(2019)的思路,本文在文档层面对主题概率分布进行情绪调整,以融合情绪信息与主题信息^①。将调整后的主题序列重新训练NF-LDA模型,并对比引入情绪前后的定价表现。

表4结果显示,在全样本异象检验中,情绪调整的NF-LDA模型在因子数K较低时解释力有限;但随着K的增加,其表现有所改善。例如,五因子模型

^① 尽管理想状态下应直接测度每个主题的情绪倾向,但在技术实践中,由于同一文档中不同主题可能交织混合情绪(如某主题同时包含乐观和悲观词汇),直接分离主题层面的情绪可能存在较大噪声。在实现过程中,本文采用姜富伟等(2021)的中文金融情绪词典,利用词频法统计文本中正面词汇和负面词汇出现的频次,进而计算每篇文档的情绪得分 s_j ;然后,使用文档情绪得分调整该文档的LDA主题分布 $\theta_{j,t} : \hat{\theta}_{j,t} = s_j \times \theta_{j,t}$ 。若文档中未出现任何情感词,则将文档情绪得分设为1($s_j = 1$),从而保留文档原始的主题分布。

可解释近 90% 的资产异象, 平均 $|t(\alpha)|$ 也低于基准的 NF-LDA 模型。在样本外策略方面, 引入情绪信息同样具有价值。特别是, 当 K 较小时, 情绪调整的模型显著提升了夏普比率。与表 2 横向比较发现, NF-LLM 在解释异象和交易策略方面均优于情绪调整的 NF-LDA, 说明引入情绪信息虽能提升 NF-LDA, 但其效果仍不及 LLMs 嵌入向量所捕捉的丰富文本信息。

表 4 情绪调整后 NF-LDA 模型的资产定价效果

K	NF-LDA(情绪调整)			NF-LDA(基准)		
	异象检验		交易策略	异象检验		交易策略
	平均 $ t(\alpha) $	显著比例	年化夏普比率	平均 $ t(\alpha) $	显著比例	年化夏普比率
1	1.58	22.92	0.17	1.58	22.92	0.06
2	1.60	30.21	0.38	1.21	17.71	0.27
3	1.32	22.92	0.75	1.16	15.62	0.34
4	1.33	21.88	0.75	1.17	15.62	0.88
5	1.12	12.50	0.84	1.20	16.67	0.81
6	1.11	12.50	0.91	1.16	17.71	0.74

(二) 叙事因子模型的可解释性

叙事因子模型通过将可读的新闻文本与风险因素相结合, 为理解不可观测的 SDF 提供了一种新工具。在可解释性方面, NF-LDA 模型能够揭示影响资产定价的核心叙事主题, 而 LLMs 嵌入向量虽可捕捉复杂语义关系, 但其经济学含义却相对模糊^①。为提升 NF-LLM 模型的可解释性, 本文结合两类文本分析方法, 通过评估 NF-LDA 识别的关键主题在 LLMs 中的重要性, 在一定程度上打开“黑箱”。

1. 关键叙事主题

首先, 本文依据 Bybee et al.(2023)的做法, 对 NF-LDA 定价模型进行经济层面的解释, 识别出对 SDF(即状态变量的定价核)具有重要影响的关键叙事主题。^② 图 2 展示了影响系数排名前十的关键叙事主题: 与定价核正相关的主题(如“风险管理”“经济发展”“经济改革”)多与经济基本面改善同步; 而负相关主

^① 在经济学研究中, 对文本分析的可解释性通常要求识别出哪些具体词语或主题类别发挥关键作用, 而大语言模型在这一方面存在天然劣势(Cao et al., 2024)。其一, LLMs 主要捕捉文本的高阶语义特征(如语境、语序及词语间的深层关联), 而非依赖单一词语的独立作用。其二, LLMs 并不具备明确的数据生成过程, 嵌入向量与词语之间缺乏直接对应关系, 难以追踪特定词语对模型输出的影响。

^② 具体地, 使用 Bybee et al.(2023)文中的公式(11)计算定价核 x_t^{MVE} 与 LDA 主题关注度 z_t 之间的映射关系 $I_{z \rightarrow MVE}$ 。与其一致, 我们在全样本上训练 NF-LDA 三因子模型, 计算系数向量 $I_{z \rightarrow MVE}$, 并按绝对值排序。

题(如“美国经济”“资本运作”“房地产市场”)则常预示基本面弱化或波动。不难发现,A股市场的核心叙事偏向政策主导,关键主题往往围绕宏观经济与政府政策展开。^①这一方面源于国内财经媒体对宏观新闻的重点报道;另一方面,在我国以散户投资者为主的市场环境下,投资者普遍存在较强的“政策依赖心理”和“过度自信”倾向(Liu et al., 2022; Jones et al., 2024),使其投资决策易受政策信号影响。

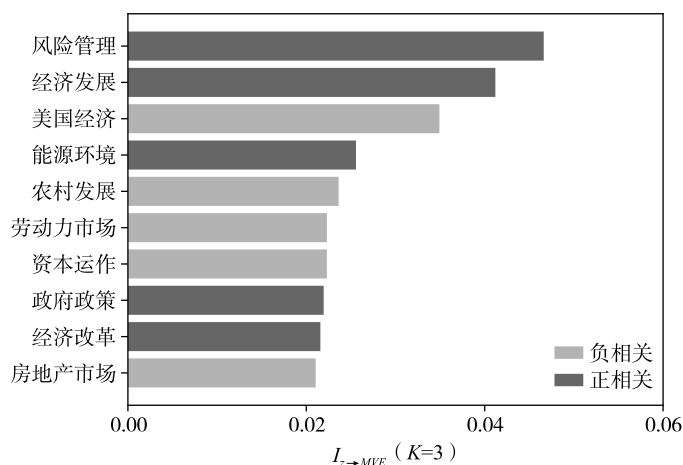


图2 NF-LDA模型的可解释性分析

2. 结合两类叙事因子模型

为了将LDA与LLMs构建的两类叙事因子模型进行结合,本文基于上述识别的关键主题,将新闻样本分为关键叙事主题主导的子样本与非关键叙事主题主导的子样本两类,然后分别利用两组子样本的LLMs嵌入向量训练NF-LLM模型,并进行与全样本模型相同的资产定价测试。^②这种方法实现了两类文本分析方法的结合,有助于揭示哪些叙事在大语言模型驱动资产定价中发挥关键作用。

表5报告了两组NF-LLM模型的性能,包括对96个资产异象的定价检验和叙事因子的样本外策略表现。整体而言,在关键叙事主题主导子样本上训练的NF-LLM模型不仅可以更好地解释资产异象,还能获得更高的年化夏普比率。例如,仅有7.29%的异象在由关键叙事主题构造的NF-LLM六因子模型中显著,远低于同规格的非关键叙事主题模型;年化夏普比率最高可达0.59,约

^① 本文结合Campbell收益率分解,实证发现LDA识别的叙事主题与基本面相关的现金流信息具有更强的关联性,进一步佐证上述结果。相关内容见附录III。感谢审稿专家提出的宝贵建议。

^② 样本分组标准如下:第一,使用LDA模型输出每篇新闻文档的主题分布概率矩阵,提取概率值排名前五的主题;第二,若其中包含三个及以上关键叙事主题,则将该新闻归入关键叙事主题主导子样本,反之则划入非关键主题主导子样本。经统计分析,最终获得的关键叙事主题主导子样本占全样本的比重约为23.28%。

为非关键叙事主题模型最高值的 1.5 倍。这一结果表明,LLMs 模型确实能够有效捕捉关键主题的叙事信息,而正是这些叙事在 A 股市场的资产定价中发挥着关键作用。

表 5 子样本 NF-LLM 模型的资产定价效果

K	关键叙事主题			非关键叙事主题		
	异象检验		交易策略	异象检验		交易策略
	平均 $ t(\alpha) $	显著比例	年化夏普比率	平均 $ t(\alpha) $	显著比例	年化夏普比率
1	1.58	22.92	0.04	1.55	22.92	0.07
2	0.86	10.42	0.27	1.62	23.96	0.25
3	0.88	8.33	0.29	1.61	25.00	0.26
4	0.88	7.29	0.35	1.59	22.92	0.34
5	0.87	8.33	0.45	1.57	23.96	0.38
6	0.90	7.29	0.59	1.55	21.88	0.40

(三) 高频因子模型和联合因子模型

新闻叙事既可能包含投资者的理性预期,也可能反映短期市场情绪(姜富伟等,2021;Dim et al., 2023)。在 A 股市场,较高的换手率和投资者对短周期的关注,使得高频叙事信息可能在资产定价中发挥更重要的作用。为此,本文构建基于周度数据的高频叙事因子模型,并进一步考察其定价效果。^① 表 6 的 Panel A 展示了周度 NF-LDA 模型的资产定价结果。整体上,周度模型较月度模型展现出更强的异象定价能力。例如,周度四因子模型可解释 90% 以上的异象,GRS 统计量仅为 1.65,均显著优于同规格月度模型。随着因子维度 K 增加,模型解释力逐步提升,尤其在 $K > 3$ 后定价效果显著增强,说明高频数据的大样本优势有助于更充分利用文本信息。即便与月度 NF-LLM 模型(表 2)相比,周度 NF-LDA 模型仍表现优异,在因子数较少时更具优势。第三列进一步显示,周度叙事因子模型的交易策略夏普比率最高达 1.06,较月度最优值(0.88)提升 20%。

最后,本文在传统 FFC 六因子模型的基础上,引入 NF-LLM 叙事因子构建联合定价模型,以探究文本信息与股票特征之间究竟是互补还是替代? 表 6 的 Panel B 报告了基于联合模型构建交易策略的夏普比率。^② 分析表明,叙事因子与传统定价因子具有互补性。纵向来看,在基准模型中加入叙事因子可显著提升样本外夏普比率,且效果随因子数量增加而单调上升。例如,“NF+FFC6”模

^① 由于 LLMs 模型的文档嵌入维度极高,估计周度 NF-LLM 模型涉及庞大的计算量且非常耗时,导致作者未能完成训练,故周度模型仅报告 NF-LDA 模型的结果。

^② NF-LDA 与 FFC 联合模型的交易策略结果,以及联合模型的资产异象检验结果报告在附录 III 中。

型在 $K=3$ 时夏普比率达 1.07, 远高于单一 FFC6 模型的 0.27。横向来看, 在叙事因子模型中加入传统因子也能改善收益能力, 尤其在 $K \geq 3$ 时效果显著。

表 6 高频因子模型和联合因子模型的资产定价结果

Panel A: 高频因子模型						
K	NF-LDA(周度)			NF-LDA(月度)		
	异常检验	交易策略		异常检验	交易策略	
	平均 $ t(\alpha) $	显著比例	年化夏普比率	平均 $ t(\alpha) $	显著比例	年化夏普比率
1	1.54	20.83	0.53	1.58	22.92	0.06
2	1.22	19.79	0.49	1.21	17.71	0.27
3	0.99	14.58	0.48	1.16	15.62	0.34
4	0.97	8.33	0.74	1.17	15.62	0.88
5	0.77	7.29	1.06	1.20	16.67	0.81
6	0.77	7.29	0.88	1.16	17.71	0.74

Panel B: 联合因子模型(交易策略)							
K	0	1	2	3	4	5	6
仅 NF-LLM		0.06	0.68	0.86	0.90	0.86	0.95
NF+CAPM	0.18	0.29	0.81	0.86	0.88	0.83	0.88
NF+FF3	0.04	0.39	0.86	0.95	0.97	0.80	0.86
NF+FF5	0.24	0.40	1.17	1.06	1.14	0.86	1.10
NF+FFC6	0.27	0.42	1.20	1.07	1.16	0.89	1.10

六、政策含义和研究展望

本文结论具有如下政策启示。第一, 加大“人工智能+金融”的政策支持力度, 推动其在市场分析、风险管理和交易优化等应用场景的广泛落地。为充分释放这一潜力, 政策层面可积极引导金融机构与高校科研院所加大研发投入, 深化产学研协同创新, 推进经济大模型的研发与应用。同时, 应注重提升人工智能模型的可解释性与透明度, 确保其在金融实践中的安全性与合规性。第二, 在 AI 时代, 信息传播的速度和广度显著提升, 金融市场的叙事驱动效应愈发突出。我国资本市场以散户投资者为主, 对不同经济叙事的敏感度较高, 容易因市场情绪波动而产生非理性交易行为, 从而加剧系统性金融风险。监管机构应积极利用人工智能技术, 提升监管的精准度和前瞻性, 防范因信息失真导致的市场失序。

在本文分析的基础上, 仍有两个关键问题需要在后续的研究与实践中共予

以进一步关注和拓展。一方面,尽管本文已尝试关注大语言模型生成叙事因子的“黑箱”问题,但对其经济含义的解读仍较有限。未来可结合主题建模(如BERTopic模型)、注意力权重可视化等可解释性机器学习方法,更准确地识别哪些叙事信息对资产定价具有驱动效果。另一方面,本文通过滚动窗口和考虑新闻发布时间等措施,在一定程度上缓解了“前瞻性偏误”(look-ahead-bias)问题,但也存在不足:如LDA主题序列基于全样本一次性估计而非实时更新,以及生成周度因子时需要在月度层面估计协方差。未来研究可进一步采用更严谨、完善的计量方法,增强结论的稳健性与可推广性。

参考文献

- [1] Aletti, S., and T. Bollerslev, “News and Asset Pricing: A High-Frequency Anatomy of the SDF”, *The Review of Financial Studies*, 2025, 38(3), 712-759.
- [2] Blei, D. M., A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, 2003, 3, 993-1022.
- [3] Bybee, L., B. Kelly, A. Manela, and D. Xiu, “Business News and Business Cycles”, *The Journal of Finance*, 2024, 79(5), 3105-3147.
- [4] Bybee, L., B. Kelly, and Y. Su, “Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text”, *The Review of Financial Studies*, 2023, 36(12), 4759-4787.
- [5] Cao, S., B. Yang, and A. L. Zhang, “Beyond the Lines: Uncovering Private Information from Fund Managers’ Disclosures”, *SSRN Electronic Journal*, 2024.
- [6] Carhart, M. M., “On Persistence in Mutual Fund Performance”, *The Journal of Finance*, 1997, 52(1), 57-82.
- [7] Chen, Y., B. T. Kelly, and D. Xiu, *Expected Returns and Large Language Models*. Rochester, NY, 2022.
- [8] Cochrane, J. H., “A Cross-Sectional Test of an Investment-Based Asset Pricing Model”, *Journal of Political Economy*, 1996, 104(3), 572-621.
- [9] Devlin, J., M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding”, arXiv, 2019.
- [10] Dim, C., F. Sangiorgi, and G. Vilkov, “Media Narratives and Price Informativeness”, *SSRN Electronic Journal*, 2023.
- [11] Engle, R. F., S. Giglio, B. Kelly, H. Lee, and J. Stroebel, “Hedging Climate Change News”, *The Review of Financial Studies*, 2020, 33(3), 1184-1216.
- [12] Faccini, R., R. Matin, and G. Skiadopoulos, “Dissecting Climate Risks: Are They Reflected in Stock Prices?”, *Journal of Banking & Finance*, 2023, 155, 106948.
- [13] Fama, E. F., and K. R. French, “Multifactor Explanations of Asset Pricing Anomalies”, *The Journal of Finance*, 1996, 51(1), 55-84.
- [14] Fama, E. F., and K. R. French, “Dissecting Anomalies with a Five-Factor Model”, *The Review of Financial Studies*, 2016, 29(1), 69-103.

- [15] Feng, G., S. Giglio, and D. Xiu, "Taming the Factor Zoo: A Test of New Factors", *The Journal of Finance*, 2020, 75(3), 1327-1370.
- [16] Garrido-Merchan, E. C., R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT Against Traditional Machine Learning Models in Text Classification", *Journal of Computational and Cognitive Engineering*, 2023, 2(4), 352-356.
- [17] Gentzkow, M., B. Kelly, and M. Taddy, "Text as Data", *Journal of Economic Literature*, 2019, 57(3), 535-574.
- [18] 洪永淼、刘倬奇、薛润坡, "政府与市场心理因素的经济影响及其测度", 《管理世界》, 2023年第3期, 第30—51页。
- [19] Hou, K., C. Xue, and L. Zhang, "Digesting Anomalies: An Investment Approach", *The Review of Financial Studies*, 2015, 28(3), 650-705.
- [20] 姜富伟、刘雨旻、孟令超, "大语言模型、文本情绪与金融市场", 《管理世界》, 2024年第8期, 第42—64页。
- [21] 姜富伟、孟令超、唐国豪, "媒体文本情绪与股票回报预测", 《经济学》(季刊), 2021年第4期, 第1323—1344页。
- [22] Jin, D., Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment", *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(5), 8018-8025.
- [23] 金星晔、左从江、方明月、李涛、聂辉华, "企业数字化转型的测度难题: 基于大语言模型的新方法与新发现", 《经济研究》, 2024年第3期, 第34—53页。
- [24] Jones, C. M., D. Shi, X. Zhang, and X. Zhang, "Retail Trading and Return Predictability in China", *Journal of Financial and Quantitative Analysis*, 2024, 1-37.
- [25] Ke, Z. T., B. T. Kelly, and D. Xiu, *Predicting Returns with Text Data*. National Bureau of Economic Research, 2019.
- [26] Kelly, B. T., S. Pruitt, and Y. Su, "Characteristics Are Covariances: A Unified Model of Risk and Return", *Journal of Financial Economics*, 2019, 134(3), 501-524.
- [27] Kozak, S., S. Nagel, and S. Santosh, "Interpreting Factor Models", *The Journal of Finance*, 2018, 73(3), 1183-1223.
- [28] Larsen, V. H., and L. A. Thorsrud, "The Value of News for Economic Developments", *Journal of Econometrics*, 2019, 210(1), 203-218.
- [29] Leippold, M., "Sentiment Spin: Attacking Financial Sentiment with GPT-3", *Finance Research Letters*, 2023, 55, 103957.
- [30] 李斌、邵新月、李玥阳, "机器学习驱动的基本面量化投资研究", 《中国工业经济》, 2019年第8期, 第61—79页。
- [31] 林建浩、张一帆、陈良源、邓益萌, "基于新闻情绪的机器学习交易策略", 《计量经济学报》, 2022年第4期, 第881—908页。
- [32] Liu, H., C. Peng, W. A. Xiong, and W. Xiong, "Taming the Bias Zoo", *Journal of Financial Economics*, 2022, 143(2), 716-741.
- [33] Liu, J., R. F. Stambaugh, and Y. Yuan, "Size and Value in China", *Journal of Financial Economics*, 2019, 134(1), 48-69.
- [34] 刘青、肖柏高, "劳动力成本与劳动节约型技术创新——来自 AI 语言模型和专利文本的证据", 《经

- 济研究》,2023年第2期,第74—90页。
- [35] Liu, Y., and B. Matthies, “Long-Run Risk: Is It There?”, *The Journal of Finance*, 2022, 77(3), 1587-1633.
- [36] Merton, R. C., “An Intertemporal Capital Asset Pricing Model”, *Econometrica*, 1973, 41(5), 867-887.
- [37] Peters, M. E., M. Neumann, L. Zettlemoyer, and W. Yih, “Dissecting Contextual Word Embeddings: Architecture and Representation”, arXiv, 2018.
- [38] Rossi, A. G., and A. Timmermann, “Modeling Covariance Risk in Merton’s ICAPM”, *The Review of Financial Studies*, 2015, 28(5), 1428-1461.
- [39] Shapiro, A. H., M. Sudhof, and D. J. Wilson, “Measuring News Sentiment”, *Journal of Econometrics*, 2022, 228(2), 221-243.
- [40] 沈艳、陈赞、黄卓,“文本大数据分析在经济学和金融学中的应用:一个文献综述”,《经济学》(季刊), 2019年第4期,第1153—1186页。
- [41] Shiller, R. J., “Narrative Economics”, *American Economic Review*, 2017, 107(4), 967-1004.
- [42] Shleifer, A., and R. W. Vishny, “The Limits of Arbitrage”, *The Journal of Finance*, 1997, 52(1), 35-55.
- [43] Tan, L., H. Wu, and X. Zhang, *Large Language Models and Return Prediction in China*. Rochester, NY, 2023.
- [44] 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年第11期,第137—147页。
- [45] Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy, “More than Words: Quantifying Language to Measure Firms’ Fundamentals”, *The Journal of Finance*, 2008, 63(3), 1437-1467.
- [46] You, J., B. Zhang, and L. Zhang, “Who Captures the Power of the Pen?”, *The Review of Financial Studies*, 2018, 31(1), 43-96.
- [47] 游家兴、吴静,“沉默的螺旋:媒体情绪与资产误定价”,《经济研究》,2012年第7期,第141—152页。
- [48] 张一帆、林建浩、樊嘉诚,“新闻文本大数据与消费增速实时预测——基于叙事经济学的视角”,《金融研究》,2023年第5期,第152—169页。

Narrative Asset Pricing with Large Language Models

FAN Jiacheng LIN Jianhao*

(Sun Yat-sen University)

LI Zongyu

(GF Securities Co., Ltd.)

Abstract: We use large language model (BERT) to extract narrative information from news texts, and construct a narrative asset pricing model via sparse instrumented principal component analysis (IPCA). We find that state variables reflect macroeconomic expectations, output, and stock returns. Compared with traditional factors based on stock characteristics, text-based narrative factors better explain asset anomalies and exhibit stronger out-of-sample performance. We provide evidence that BERT outperforms topic models such as Latent Dirichlet Allocation (LDA), and examine its potential for interpretability, high-frequency factor extraction, and integration with traditional factors in asset pricing.

Keywords: cross-sectional asset pricing; large language models; textual analysis

JEL Classification: C38, G11, G12

* Corresponding Author: LIN Jianhao, Lingnan College, Sun Yat-sen University, No. 135 Xingang West Road, Haizhu District, Guangzhou, Guangdong 510275, China; Tel: 86-20-84112104; E-mail: linjh3@mail.sysu.edu.cn.