政策与舆论信息能否提升行业 ETF 资金流的短期预测能力？——基于多模型比较的实证分析

蒋文浩 2025104245

2025 年 12 月 27 日

目录

1 引言

1.1 研究背景与目标

2025 年，中国资本市场正式步入“后复苏”时代的深水区。在这一年中，全球地缘政治格局持续剧烈演变，尤其是 4 月爆发的新一轮中美贸易冲突，不仅重新点燃了全球投资者对供应链扰动与经济衰退风险的担忧，也导致全球主要股票指数在短期内出现同步且非理性的剧烈下挫。

在此背景下，为对冲外部不确定性并提振国内需求，各部委及地方政府密集推出一系列消费刺激和稳预期政策。然而，随着政策密度与信息复杂度的急剧提升，市场对政策事件的反应机制正出现范式性转变。一方面，不同行业主题 ETF 的资金流入流出（记为 Y_t）呈现显著异质性；另一方面，传统依赖基本面与市场走势的解释框架已难以充分刻画投资资金的即时反应机制。政策冲击（记为 D_t）改变公众与市场主体对行业前景、产业逻辑或政策倾斜的预期，影响相关议题的搜索强度、讨论密度（表现为 OI_t ），与投资者情绪倾向（记为 ET_t ），两者作为舆论热度一体两面的另一表征，深刻影响着市场叙事。基于此，本文以“公开可获取的数据”为基础，构建一个日度行业面板数据集，将政策文本信息（记为 $D_{i,t}$）、网络舆论指标（$OI_{i,t}$ 与 $ET_{i,t}$）与市场控制变量（$X_{i,t}$）在时间维度上对齐，并围绕以下两个任务展开：

- **探索性分析（EDA）：**描述资金流、政策与舆论指标的基本分布、时间走势与相关关系，展示清洗与预处理过程。
- **预测建模：**将资金流预测视为监督学习问题，比较不同模型在样本外预测中的表现，并分析模型差异的原因。

1.2 研究目标、分析思路与研究意义

本文围绕**“信息变量是否能够提升行业 ETF 资金流的可预测性”**这一核心问题，提出两条具有可操作性的研究问题：

第一，政策强度指数与舆论指标在数据层面是否与行业 ETF 资金流存在稳定的统计关联结构？该关联是否具有跨行业的一致性或显著异质性？

第二，在相同的训练/测试划分下，时间序列模型、线性回归模型与机器学习模型的预测效果有何差异？进一步地，引入政策与舆论特征是否能够降低样本外误差、改善方向判断能力？

围绕上述问题，本文选取六个行业主题 ETF 作为研究对象，构建涵盖政策信息、舆论信息与市场环境的多源数据集。采用如上介绍的变量体系，进行建模分析。在机制设想层面，本文认为政策冲击可能通过改变市场对行业前景的预期结构，经由舆论关注度与情绪倾向的变化被投资者“感知—解读—放大”，并最终转化为资金流入与流出的市场行为。其基本逻辑如图1所示：

在分析路径上，本文遵循**“由数据到模型、由预测到机制”**的研究思路，整体分为三个层次：

（一）探索性数据分析（EDA）。在正式建模之前，本文对核心变量开展系统性 EDA，重点考察资金流、政策强度与舆论指标的时间演化特征、变量间相关结构与滞后关系、以及跨行业异质性。EDA 旨在识别信息变量在数据层面的潜在预测价值，并为后续模型设定提供经验依据。

（二）多模型预测性能比较（核心实证）。本文将样本按时间顺序划分为训练集与测试集，并在完全一致的数据划分下比较不同建模范式的样本外预测表现。时间序列基准模型包括 ARIMA（仅利用历史 Y_t）与 ARIMAX（在 ARIMA 框架中引入 D_t, OI_t, ET_t 及 X_t 等外生变量）。在此

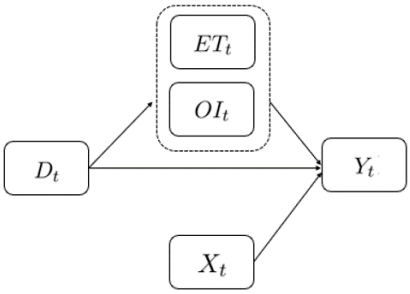


图 1: 政策冲击—舆论变化—ETF 资金流变化的机制设想

基础上，采用两类机器学习模型：梯度提升树（Boosting）与随机森林（Random Forest），以捕捉非线性与高阶交互结构。模型表现以样本外误差指标（如 RMSE、MAE）与方向预测准确率等进行评估，从而检验信息变量是否具有可稳定复现的预测增量。

（三）DML 机制分析（辅助分析）。在完成预测性能比较后，本文进一步采用 Double Machine Learning（DML）框架对政策与舆论变量的作用机制进行分析。DML 的目的并非追求最优预测精度，而是在高维控制变量情形下对政策冲击及舆论指标对资金流的边际影响进行相对稳健的估计，从而为“预测改进是否对应稳定的边际效应/机制通道”提供解释支撑。

总体而言，本文的研究意义体现在以下三个方面：（1）从预测视角出发，系统评估政策与舆论信息在行业 ETF 资金流预测中的增量价值；（2）在统一数据与样本划分下比较时间序列模型、线性模型与机器学习模型的适用性与局限；（3）通过 DML 补充机制解释，在预测表现与经济含义之间建立联系。

全文结构安排如下：第二部分介绍数据来源、变量构造与探索性分析；第三部分展示多模型预测框架与样本外结果比较；第四部分基于 DML 框架分析政策与舆论变量的作用机制；第五部分给出结论与讨论。

2 数据处理与变量构造

2.1 原始数据来源与预处理

政策信息的主要来自我国各级政府官方网站、北大法宝政策数据库以及新华网等权威媒体报道。中央层面的政策主要来源于相关部委官网及国务院统一发布平台，其中国家发展改革委官网（规范性文件库）；文化和旅游部官网（含政策法规库、政府信息公开目录）；国家新闻出版署官网（游戏审批信息公示栏目）是与本文研究高度相关的核心发布渠道。地方层面的政策数据主要来自各省市政务公开平台，如省（区、市）文化和旅游行政主管部门官网，以及地方政府官方网站的政务公开或政策文件栏目等。政策文件统一提取文本内容与发布时间用于构造政策冲击变量 D_t 。

舆论数据来自金融平台与大众平台两类渠道。其中金融平台选取东方财富股吧（https://guba.eastmoney.com/）作为主要数据源。覆盖行业相关 ETF/基金吧、指数/板块吧及龙头个股吧，不同行业的标的选择依据行业覆盖度与市场关注度确定，其完整对应关系见附录表。按日期爬取其上部分相关帖文与评论文本，整理以构造舆论情感倾向变量 ET_t 大众舆论数据来自百度指数，通过选取每个行业的三个代表性关键词，获取其 PC 端与移动端搜索量，以衡量市

场舆论关注度的日度波动。所得搜索指数用于构造舆论关注度变量 OI_t。

市场数据主要来源于 WIND 与 Tushare 接口，通过对应 ETF 换手率、成交量、成交额等指标构建控制变量集合 X_t。

2.2 变量构造

2.2.1 政策冲击指数 D 的构造

本文依据国家与地方两级、与行业（按 GB/T 4754）匹配的全部政策文本，构造行业 i 的日度政策暴露指数。对每条政策 j，为量化政策强度，本文采用 Qwen3 1.7B 对政策文本进行结构化标注，提取五类特征：层级、新颖度、财政性、约束力度与覆盖范围，并按照规则赋予权重。具体的，L_j：国家级 $=1$、省级 $=0.6$、市级 $=0.3$；依照是否明确财税条款 $B_j \in \{1.3, 1\}$；依照实施细则/规章到位程度 $R_j \in \{1.3, 1, 0.8\}$；依照覆盖范围为全行业、部分产业、细分产业赋 $C_j \in \{1.2, 1, 0.8\}$；依照政策是首发、调整、跟发赋 $N_j \in \{1.2, 1, 0.8\}$。

$$w_j = \underbrace{L_j}_{\text{层级}} \times \underbrace{B_j}_{\text{财税/软硬性}} \times \underbrace{R_j}_{\text{约束/可执行}} \times \underbrace{C_j}_{\text{覆盖范围}} \times \underbrace{N_j}_{\text{新颖度}}. \tag{2.1}$$

政策暴露量采用“有限记忆”的指数衰减机制。设行业 i 当日政策到达量为 $a_{i,t}$，记忆长度为 M（本文取 $M=14$），则政策库存按 $D_{i,t} = \delta D_{i,t-1} + a_{i,t} - \delta^M a_{i,t-M}$ 更新，其中日衰减因子 $\delta = \exp(-\lambda)$，并由在 M 日后衰减至比例 ρ 得 $\delta^M = \rho$、$\lambda = -\log(\rho)/M$ 反解。为减弱政策密集期的累积效应，对 $D_{i,t}$ 施加滚动标准化，定义最近 W（本文取 $W=30$）日到达量 $\mathrm{Den}_{i,t} = \sum_{k=t-W+1}^{t} a_{i,k}$，并令标准化指数 $D_{i,t}^{\mathrm{norm}} = D_{i,t}/(\mathrm{Den}_{i,t} + \varepsilon)$，其中 ε 为防止分母为零的常数。进一步地，分别对国家与地方政策构造 $D_{i,t}^{\mathrm{nat}}$ 与 $D_{i,t}^{\mathrm{loc}}$，并取总暴露量 $D_{i,t}^{\mathrm{total}} = D_{i,t}^{\mathrm{nat}} + D_{i,t}^{\mathrm{loc}}$ 作为解释 Y_t、OI_t 与 ET_t 变化的日度政策冲击指数。

2.2.2 变量 M1：行业舆论热度 OI 的构造

行业舆论热度来自关键词篮子的百度指数。对单指标先做双侧 winsorize（5–99%）及节假日异常清洗，再取 $\log(1+\mathrm{IDX}_{k,t})$ 进单个搜索词指数的标准化 $z_{k,t} = \big(\log(1+\mathrm{IDX}_{k,t}) - \mu_k\big)/\sigma_k$。将同一行业日度舆论热度取等权聚合：$\mathrm{OI}_{i,t} = (1/K)\sum_{k=1}^{K} z_{k,t}$，并作为中介 OI。

2.2.3 变量 M2：行业情感倾向 ET 的构造

为刻画日度舆论情绪，本文采用 Qwen3 1.7B 对所爬取的股吧每日随机抽取三条文本内容做情感识别，得到离散情感分数 $\mathrm{score}_p \in [1, 5]$。行业 i 在日 t 的情绪均值记为 $\mathrm{ET}_{i,t}^{\mathrm{mean}} = (1/P_{i,t})\sum_{p=1}^{P_{i,t}} \mathrm{score}_p$（不足三条样本仍取等权）。随后在行业内做 z-score 标准化，并以指数移动平均平滑：$\mathrm{ET}_{i,t} = \alpha\,\mathrm{ET}_{i,t}^{\mathrm{mean}} + (1-\alpha)\mathrm{ET}_{i,t-1}$，其中 $\alpha = 1 - 2^{-1/h}$、$h \approx 7$ 为半衰期日。若对小样本日敏感，可加入总体均值的贝叶斯收缩作为稳健性检验。

2.2.4 被解释变量 Y 的处理

行业间 ETF 流量量纲差异显著且符号跳点频繁，本文采用符号对数（signed-log）形式统一尺度：$Y_{i,t}^{\mathrm{slog}} = \mathrm{sign}(\mathrm{Flow}_{i,t}) \log\big(1 + |\mathrm{Flow}_{i,t}|/s_i\big)$，其中 s_i 为行业稳健尺度（历史 P50 或 IQR）。

该变换兼具横向可比性与尾部稳定性。

2.2.5 控制变量 X 处理

控制向量包括：（i）固定效应：行业固定效应吸收长期均值差异，日或周固定效应吸收统一行情与节假日冲击；（ii）市场与宏观变量：无风险利率、货币利率（如 DR007）、价格/景气指标、市场收益与波动（大盘与行业收益、换手率、波动度）、汇率（USD/CNY 变化）、海外基准（如 S&P500）等，必要时加入滞后项控制动态自相关。将所有连续控制变量在样本内 z-score 标准化，以便线性与非参数方法学习条件结构。

2.3 补充说明

鉴于政策冲击在构造上设定为具有约 14 天的“有限寿命”，本文所用样本区间为 5 月 1 日至 10 月 31 日。其中，因果识别部分的分析窗口取为 5 月 15 日至 10 月 31 日，以保证每一期观测都充分包含此前政策的剩余效应。预测能力比较中，则将 9 月 30 日及以前的数据作为训练集，将 10 月份样本作为测试集，对不同模型的样本外表现进行评估。

在控制变量处理中，部分宏观与市场指标存在个别日期缺失。对于呈现平滑时间趋势的连续变量（如利率、货币供应量、指数收益与波动等），本文先在行业层面按日期对齐，再采用按时间顺序的一维插值（如线性插值）补全内部缺失值；对于仍残留的零星缺失，则使用样本期内的截面中位数进行填补。该策略在尽量保留时间序列结构的同时，避免因大规模删除观测或人为引入强烈跳点而干扰因果估计与预测结果的稳健性。

3 探索性数据分析（EDA）

在正式建模之前，本文首先对样本的数据完整性进行检查。总体来看，各行业面板数据在研究区间内具有较好的时间连续性，核心变量（资金流、政策强度、舆论指标）无数据缺失。个别宏观或市场控制变量存在缺失，后续建模中采用稳健的时间序列插值或中位数填补方法进行处理。

3.1 核心变量的分布特征与相关性结构

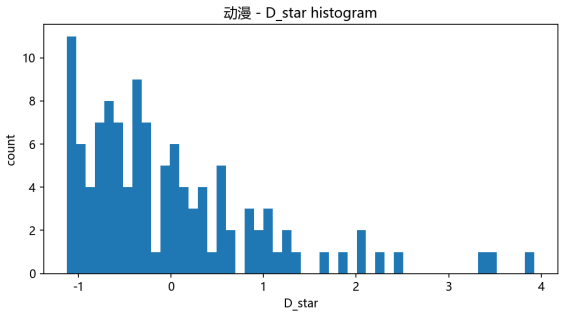
图2展示了以动漫行业为例的政策强度指数（D_t）、舆论情绪指标（ET_t）、舆论关注度指标（OI_t）以及 ETF 资金流变量（Y_t）的分布特征，所呈现的分布形态在其余行业中具有一定的代表性。

从政策强度指数 D_t 的分布可以看出，其普遍呈现出明显的右偏与厚尾特征：大多数观测值集中于较低区间，而在少数政策密集或强干预时期出现显著的高值。这表明政策冲击在时间维度上具有“低频—高强度”的结构性特征，其影响更可能通过阶段性集中释放而非线性平滑累积的方式体现。

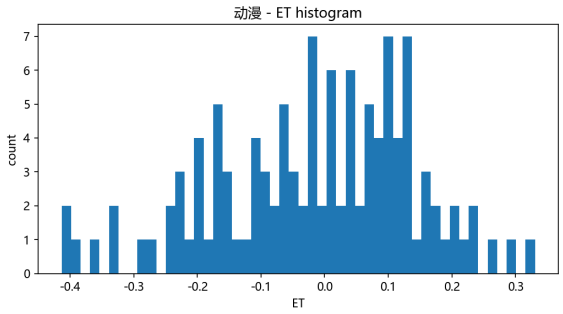
舆论情绪指标 ET_t 的分布相对集中，整体围绕零附近波动，但仍存在一定的偏态与尾部延伸。这一特征反映出，在多数交易日中市场情绪保持相对中性，而在特定事件或信息冲击下，情绪变量可能出现短期偏离，从而对市场行为产生阶段性影响。

舆论关注度指标 OI_t 的分布范围相对更广，左右尾部均较为明显，显示投资者关注度在时间维度上具有更强的波动性。相比情绪指标，关注度更容易受到突发事件或热点议题驱动，其变化幅度和持续性也更为显著，因而在预测框架中可能蕴含更丰富的动态信息。

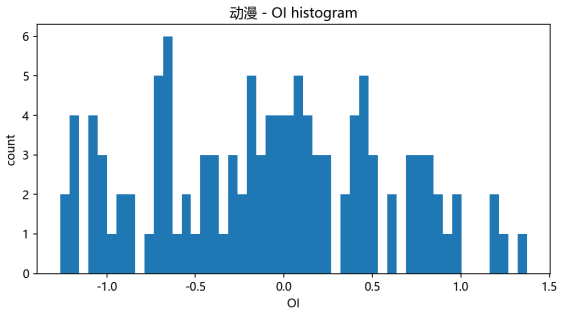
资金流变量 Y_t 整体呈现出明显的非对称与多峰分布特征，且尾部较厚，表明行业 ETF 资金流在样本期内存在频繁的正负切换以及少数幅度较大的极端流入或流出情形。这一分布特征意味着基于正态性或线性结构的模型可能难以充分刻画资金流行为，为后续引入能够处理非线性与异质性的预测方法（如树模型）提供了数据层面的动机。



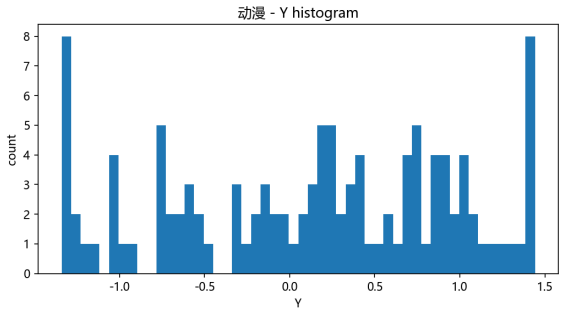
(a) 政策强度指数 D_t



(b) 舆论情绪指标 ET_t



(c) 舆论关注度指标 OI_t



(d) ETF 资金流 Y_t

图 2: 动漫行业核心变量的分布特征

图 **??** 汇总展示了六个行业中核心变量之间的 Pearson 相关系数热力图，涵盖 ETF 资金流（Y_t）、政策强度指数（D_t）、舆论关注度（OI_t）、舆论情绪指标（ET_t）以及主要市场控制变量。该图的目的并非用于因果推断，而在于刻画变量之间的线性相关结构，为后续模型设定与特征选择提供依据。

首先，从整体上看，六个行业中 Y_t 与政策变量 D_t、舆论变量 OI_t 和 ET_t 的同期相关系数整体处于中低水平，且符号在不同行业间存在差异。这一特征表明，政策与舆论因素对资金流的影响并非通过简单的静态线性关系体现，而更可能以滞后效应、非线性响应或与其他市场变量交互的方式发挥作用。这一观察为后文引入滞后项（如 ARIMAX）以及非线性机器学习模型提供了数据层面的动机。

其次，在舆论变量内部，OI_t 与 ET_t 在多个行业中呈现出中等程度的正相关，但该相关性并不稳定。这说明投资者关注度的上升并不必然伴随情绪的单向变化，二者在机制上具有区分度，也支持在建模时将关注度与情绪作为不同的信息维度分别纳入，而非简单合并处理。

再次，资金流 Y_t 与市场控制变量（如溢价因子与短期动量因子）在多数行业中表现出相对更为稳定的相关关系，这反映出资金流对传统风险因子与市场状态变量的敏感性。这一结果验

证了在后续预测与因果分析中引入控制变量集合 X_t 的必要性，以避免将宏观或市场共同因素误识别为政策或舆论效应。

总体而言，相关系数热力图揭示了以下几个特征：其一，政策与舆论变量与资金流之间的关系具有明显的行业异质性与非强线性特征；其二，舆论关注度与情绪指标在信息维度上具有互补性；其三，市场因子在资金流解释中仍占据重要地位。这些描述性发现共同构成了后续模型设计的基础，即在时间序列模型中引入外生信息，在机器学习模型中允许非线性映射，并在因果框架中通过控制变量与中介结构进行进一步辨析。

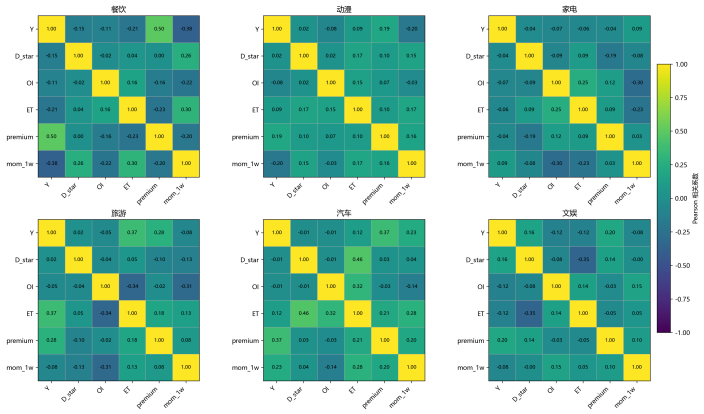


图 3: 动漫行业资金流与政策强度的先行—滞后相关关系

3.2 解释变量与资金流的时间关系

为进一步考察政策信息在时间维度上对资金流的作用方式，本文计算了动漫行业中 ETF 资金流 Y_t 与政策强度指数 D_{t-k} 之间的先行—滞后相关系数，即 $\mathrm{corr}(Y_t, D_{t-k})$，其中 k 表示政策变量相对于资金流的滞后期数。相关结果如图4所示。

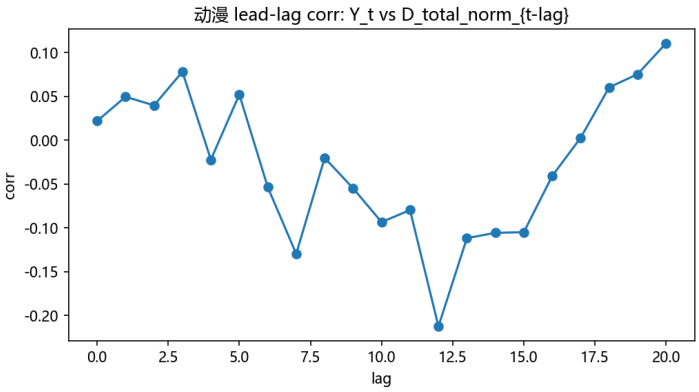


图 4: 动漫行业资金流与政策强度的先行—滞后相关关系

从结果可以看出，政策强度与资金流之间的相关性在当期（$k = 0$）并非达到最大值，而是在若干滞后期内出现更为显著的相关结构。具体而言，在中短期滞后区间内，相关系数的绝对值有所放大，表明政策冲击对资金流的影响并非即时完成，而可能通过信息消化、预期调整与交易决策等过程逐步体现。

这一现象说明，政策变量在预测资金流时具有潜在的动态信息价值，仅依赖当期政策强度

可能低估其实际影响。该发现为后续预测模型中引入滞后项结构，以及在 ARIMAX 与机器学习模型中系统性使用政策与舆论的历史信息提供了经验依据。

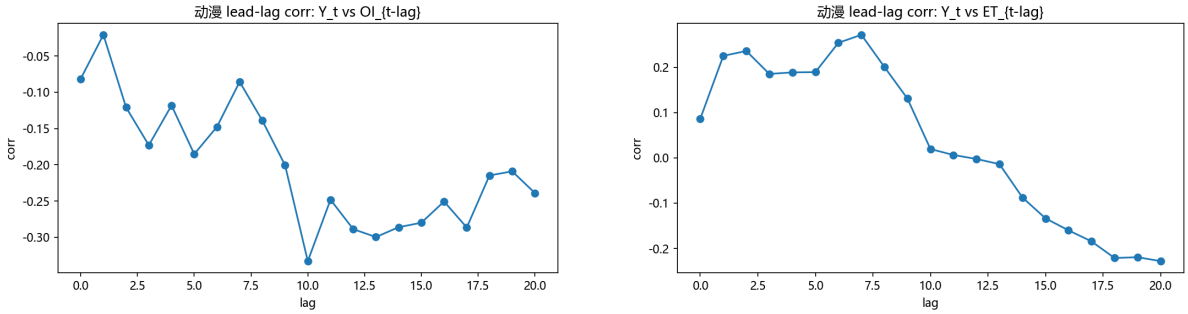
3.2.1 舆论变量与资金流的先行—滞后关系

除政策强度变量外，本文进一步考察了舆论关注度指标（OI_t）与舆论情绪指标（ET_t）相对于资金流 Y_t 的时间先后关系。具体而言，计算了 $\mathrm{corr}(Y_t, OI_{t-k})$ 与 $\mathrm{corr}(Y_t, ET_{t-k})$ 在不同滞后期 k 下的变化情况。相关结果的示例如图5所示。

从整体模式来看，舆论变量与资金流之间普遍呈现出明显的时间结构特征。舆论关注度 OI_t 与资金流的相关性在当期及短期滞后内并不稳定，而在若干滞后期内其相关性幅度有所放大，表明投资者关注度的变化并非即时转化为资金流入或流出，而可能通过信息扩散与交易决策过程逐步体现。

相比之下，舆论情绪指标 ET_t 在较短滞后区间内与资金流表现出更为清晰的相关结构，随后相关性随滞后期增加逐渐减弱甚至发生方向变化。这一现象暗示，情绪变量对资金流的影响更偏向于短期反应，其作用窗口相对有限。

上述结果并不意味着舆论变量与资金流之间存在稳定的线性因果关系，而是表明二者在时间维度上具有可识别的动态关联结构。该发现对于预测建模具有重要启示意义：一方面，仅使用当期舆论指标可能低估其信息含量；另一方面，在模型中系统性引入舆论变量的滞后项，有助于捕捉其对资金流的潜在预测价值。



(a) 资金流与舆论关注度的先行—滞后相关关系　　(b) 资金流与舆论情绪的先行—滞后相关关系

图 5: 舆论变量与资金流之间的先行—滞后相关结构示例

4 预测评估与模型比较

4.1 总体分析思路

在本节中，本文对 ARIMA、ARIMAX、Boosting 与 Random Forest 模型在相同训练/测试划分下的预测结果进行统一比较，主要关注以下指标：均方根误差（RMSE）以及预测方向一致性。通过这一比较，本文旨在回答两个问题：一是引入政策与舆论特征是否在不同建模范式下稳定改善预测表现；二是非线性机器学习模型相较于传统时间序列模型在短期资金流预测中是否具有系统性优势。

4.2 建模流程

4.2.1 ARIMA 与 ARIMAX 模型

对每个行业的资金流序列 Y_t，先做平稳性检验（ADF）并按需要差分，得到近似平稳的 $\Delta^d Y_t$。结合 ACF/PACF 与信息准则（AIC、BIC）选取 ARIMA(p,d,q)：

$$\Phi(L)\Delta^d Y_t = \Theta(L)\varepsilon_t,$$

其中 $\Phi(L)$、$\Theta(L)$ 为自回归和滑动平均多项式，ε_t 为白噪声。考虑外生变量时，引入政策与舆论相关指标构建 ARIMAX：

$$\Phi(L)\Delta^d Y_t = \Theta(L)\varepsilon_t + \beta_1 D_t + \beta_2 OI_t + \beta_3 ET_t + \beta_4' Z_t.$$

预测时采用滚动一步法：对测试期的每一日，以截至 $t-1$ 的历史数据估计模型并生成 $\hat{Y}_{t|t-1}$；外生变量在预测点采用其实际观测值，从而得到“条件于给定政策与舆论路径”的预测。

4.2.2 机器学习预测模型：Boosting 与 Random Forest

在时间序列基准模型（ARIMA / ARIMAX）之外，本文进一步引入两类树模型为代表的机器学习方法：梯度提升树（Boosting）与随机森林（Random Forest），以评估在允许非线性与高阶交互的情形下，政策与舆论信息是否能够提升资金流的样本外预测能力。

与 ARIMAX 模型不同，机器学习方法不依赖明确的动态结构设定，而是将预测问题表述为一个监督学习任务：在给定历史信息集 $\mathcal{F}_{t-1}$ 的条件下，直接学习

$$Y_t = f(D_{t-1}, OI_{t-1}, ET_{t-1}, Z_{t-1}) + \varepsilon_t,$$

其中 $f(\cdot)$ 为未知的非线性函数，由数据驱动进行近似。

特征构造与时间序列约束 为避免信息泄露（look-ahead bias），所有机器学习模型均仅使用 $t-1$ 及以前时点的解释变量作为输入特征，包括政策强度指标、舆论关注度、情绪指标及宏观控制变量。模型训练与预测严格遵循与 ARIMA / ARIMAX 相同的时间切分方案，在训练集上拟合模型参数，并在测试集上进行滚动预测评估。

梯度提升树（Boosting） Boosting 方法通过逐步拟合残差的方式，将多个弱学习器（通常为浅层回归树）加权组合为一个强预测模型。其基本思想是在第 m 轮迭代中，针对前 $m-1$ 轮模型的预测误差进行修正，从而不断降低整体损失函数：

$$\hat{f}_M(x) = \sum_{m=1}^{M} \nu\, h_m(x),$$

其中 $h_m(x)$ 为第 m 个基学习器，ν 为学习率。

本文在训练集中采用交叉验证，对以下关键参数进行网格搜索与比较：树的数量（number of estimators）、单棵树的最大深度（max depth）、学习率（learning rate）以及子样本比例

（subsample）。通过在候选参数组合下比较交叉验证误差，选取在训练集上表现最优的一组参数，并将其固定用于测试期预测。

表 1: Boosting 模型超参数选择结果汇总（六行业）

行业	CV-RMSE	min_leaf	max_leaf_nodes	max_iter	max_depth	learning_rate	L2
动漫	0.556368	10	31	400	5	0.03	1
旅游	0.650360	10	31	400	5	0.03	1
文娱	0.682683	10	63	400	2	0.03	1
家电	0.848112	10	31	400	5	0.03	1
汽车	0.910794	20	31	400	5	0.01	5
餐饮	0.942309	10	63	400	2	0.03	1

随机森林（Random Forest） 随机森林通过对样本与特征进行随机抽样，构建多棵相互独立的回归树，并对其预测结果取平均：

$$\hat{Y}_t = \frac{1}{B}\sum_{b=1}^{B} T_b(x_t),$$

其中 $T_b(\cdot)$ 表示第 b 棵回归树。该方法能够有效降低单棵决策树的方差，对异常值与噪声具有较强的稳健性。

随机森林的调参重点主要集中在控制模型复杂度与随机性水平。本文在训练集中对森林规模（树的数量）、单棵树的最大深度、最小叶节点样本数以及每次分裂可选特征数等参数进行系统比较，并通过交叉验证误差最小化的原则确定最终模型配置。

表 2: Random Forest 模型超参数选择结果汇总（六行业）

行业	CV-RMSE	n_estimators	min_split	min_leaf	max_features	max_depth	bootstrap
汽车	0.376886	200	5	1	0.8	12	FALSE
旅游	0.395299	200	5	2	0.5	12	FALSE
动漫	0.448006	800	10	1	1.0	8	TRUE
文娱	0.583038	200	5	2	0.5	12	FALSE
家电	0.631217	200	5	1	0.8	12	FALSE
餐饮	0.858341	800	10	5	0.8	12	TRUE

4.3 预测性能比较与结果分析

基于前述统一的数据构造与滚动预测设定，本文进一步比较了传统时间序列模型（ARIMA、ARIMAX）与机器学习模型（Boosting、Random Forest）在六个行业 ETF 资金流预测中的样本外表现。评价指标主要包括均方根误差（RMSE）与方向预测准确率（Direction Accuracy），分别刻画预测幅度误差与涨跌方向判断能力。

图 6 汇总展示了不同模型在六个行业测试集上的 RMSE 表现。从整体上看，ARIMA 模型在多数行业中表现出较高的预测误差，其预测结果往往呈现明显的“平滑化”特征，难以刻画资金流的短期波动。在引入政策与舆论等外生变量后，ARIMAX 模型的预测误差在部分行业中

有所下降，说明外生信息在一定程度上提升了线性时间序列模型对资金流变动的解释能力，但其改进幅度整体有限。

相比之下，Boosting 与 Random Forest 等机器学习模型在大多数行业中取得了更低的 RMSE，尤其在波动幅度较大、非线性特征较为明显的行业中优势更为突出。这一结果表明，资金流序列中可能同时存在非线性关系与高阶交互结构，传统线性模型难以充分捕捉，而基于树结构的集成学习方法在此类情形下具有更强的拟合与泛化能力。

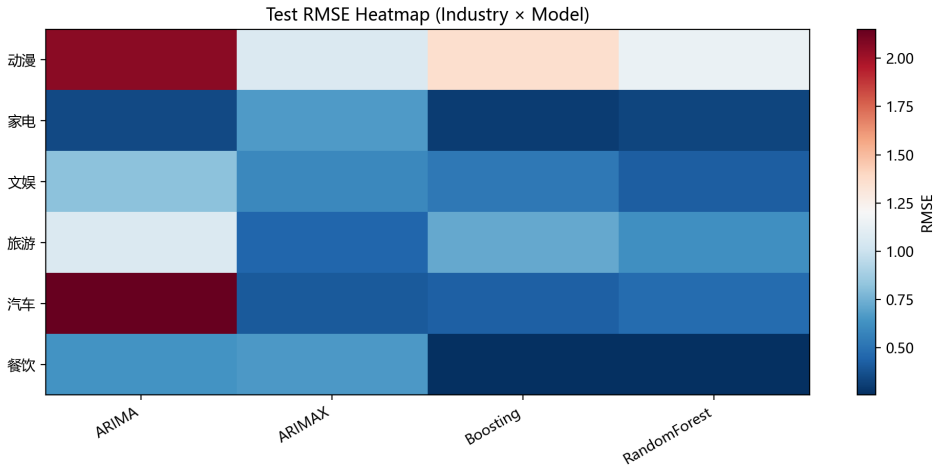


图 6: 不同模型在六个行业上的测试集 RMSE 热力图

图 7 则从方向预测准确率的角度对模型进行比较。可以观察到，ARIMA 模型在方向判断上的表现相对不稳定，而 ARIMAX 模型在若干行业中能够明显提高方向预测准确率，反映出政策与舆论变量在捕捉资金流“方向性变化”方面具有一定信息含量。机器学习模型在方向预测上整体表现较优，尤其是 Boosting 模型，在多数行业中取得了较高且较为稳定的方向准确率。

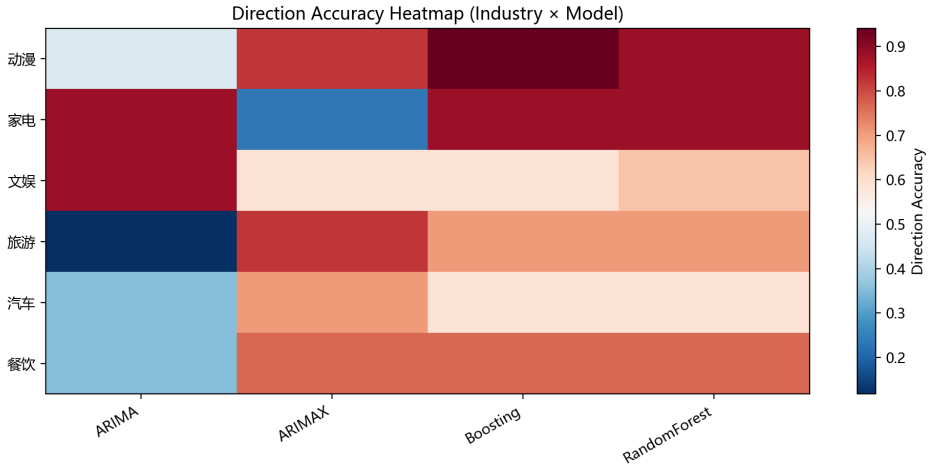
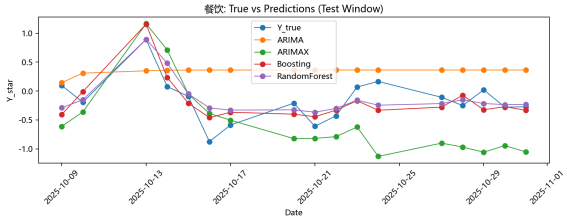


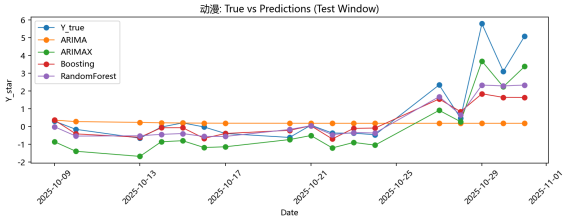
图 7: 不同模型在六个行业上的方向预测准确率热力图

为进一步直观比较不同模型在时间维度上的预测特征，图 8 给出了六个行业测试窗口内的真实资金流与各模型预测值的对比折线图。可以看到，ARIMA 模型的预测路径通常较为平滑，对局部极值与突发波动的响应不足；ARIMAX 模型在部分关键节点上能够更好地贴合真实走势，但仍存在系统性偏差。相比之下，Boosting 与 Random Forest 模型在多数行业中能够更准

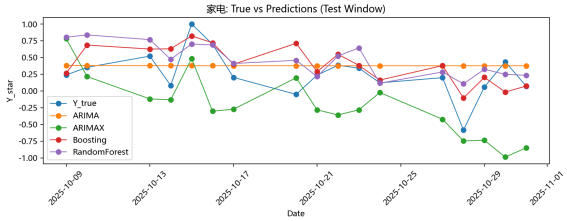
确地跟踪资金流的短期波动，并在拐点附近表现出更强的适应性。



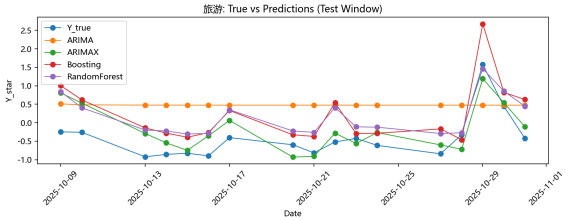
(a) 餐饮行业



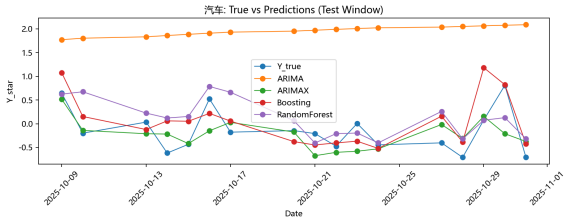
(b) 动漫行业



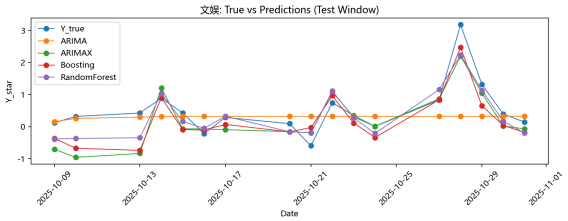
(c) 家电行业



(d) 旅游行业



(e) 汽车行业



(f) 文娱行业

图 8: 六个行业测试窗口内真实资金流与不同模型预测结果对比

总体而言，预测结果表明：在仅依赖历史信息的情形下，传统时间序列模型对资金流的刻画能力有限；引入政策与舆论等外生变量后，线性模型的预测性能有所改善，但提升幅度相对有限；基于树模型的机器学习方法在处理非线性关系和复杂特征交互方面具有明显优势，能够在多数行业中实现更低的预测误差和更高的方向判断准确率。这一发现为后续结合机器学习方法进行因果机制分析与稳健性检验提供了经验基础。

5 机制分析：基于 DML 的补充证据

5.1 总体分析思路

前文已围绕“信息变量能否提升行业 ETF 资金流预测表现”这一核心问题，系统比较了时间序列模型、线性模型与机器学习模型在样本外预测中的差异。在确认政策强度与舆论相关变量在预测层面具有一定信息增量后，本文进一步希望回答一个自然的延伸问题：这些信息变量是通过何种机制影响资金流动的？

基于前文提出的机制逻辑，为识别因果效应需要满足若干关键假设。首先，政策冲击的发生时间及其强度被视为相对外生，其在短周期内不受 ETF 资金流量反向影响，从而满足时间先后性假设。其次，在给定控制变量集合 X_t 后，政策冲击与误差项之间不再存在系统性相关性，从而满足条件独立性假设。再次，舆论关注度与情绪在时间维度上发生于政策冲击之后、资金流反应之前，其时间顺序支持中介路径的识别。此外，X_t 中包含市场风险因子、行业交易特征

及宏观变量，有助于降低遗漏变量对因果估计的影响。

在方法选择上，本文采用 Double Machine Learning（DML）以实现对因果效应的稳健评估。DML 能够利用机器学习模型灵活拟合干扰项，在高维情境下获得理论上无偏的一致估计，。

5.2 DML 的设定与结果分析

以行业 i 为单位，在给定控制变量向量 X_t（包括行业固定效应、市场与宏观控制等）下，结构关系写作 $M_{1,t} = h_1(D_t, X_t) + u_{1,t}$、$M_{2,t} = h_2(D_t, M_{1,t}, X_t) + u_{2,t}$、$Y_t = h_3(D_t, M_{1,t}, M_{2,t}, X_t) + \varepsilon_t$，并据此定义总效应 TE、自然直接效应 NDE 与沿“$D \to M_1 \to M_2 \to Y$”链条的路径依赖效应 $\mathrm{PIE}_{M_1 \to M_2}$。

DML 采用机器学习（本文采用 LightGBM）分别估计各个条件期望（例如 $E[Y_t \mid D_t, M_{1,t}, M_{2,t}, X_t]$、$E[M_{2,t} \mid D_t, M_{1,t}, X_t]$、$E[M_{1,t} \mid D_t, X_t]$ 等），并通过交叉拟合得到残差化变量（如 $\tilde{Y}_t = Y_t - \hat{E}[Y_t \mid \cdot]$、$\tilde{D}_t = D_t - \hat{E}[D_t \mid X_t]$ 等），在“正交化”后的低维残差空间中再用简单线性回归构造各类效应的估计量 $\widehat{\mathrm{TE}}$、$\widehat{\mathrm{NDE}}$、$\widehat{\mathrm{PIE}}_{M_1 \to M_2}$。为获得区间不确定性，本文在行业内对样本进行重抽样，利用 bootstrap 计算三类效应的 95% 置信区间。该 DML-中介链组合既保留了机器学习对复杂扰动项的拟合能力，又通过 Neyman 正交和交叉拟合保证在高维场景下的因果效应可解释性。

在此基础上，利用 DML-中介链框架得到政策冲击的总效应 TE、自然直接效应 NDE 与沿“$D \to M_1 \to M_2 \to Y$”路径的路径依赖效应 PIE，并通过 bootstrap 计算 95% 置信区间。表 3 汇总了六个行业的 DML 结果。

表 3: DML 的政策总效应、直接效应与中介效应比较

：DML 估计（点估计与 95% CI）			
industry	TE	NDE	PIE
餐饮	1.543 [0.356, 1.941]	0.644 [0.126, 0.967]	0.899 [0.200, 1.145]
动漫	0.989 [0.472, 1.979]	0.468 [0.190, 1.597]	0.521 [0.188, 0.938]
汽车	0.301 [0.158, 1.307]	0.037 [-0.059, 0.741]	0.263 [0.106, 0.730]
文娱	0.908 [0.474, 1.911]	0.337 [0.200, 1.037]	0.571 [0.211, 1.104]
旅游	0.956 [0.351, 1.518]	0.636 [0.280, 1.094]	0.320 [0.050, 0.552]
家电	0.641 [0.316, 1.845]	0.274 [0.143, 1.143]	0.367 [0.107, 0.943]

从表 3 可以看到，在 DML 框架下，政策冲击对行业 ETF 资金流的影响在不同产业中呈现出显著的异质性，但整体传导方向具有一致性。首先，从中介路径效应 PIE 来看，六个行业的估计值均为正，且对应的 95% 置信区间均不包含零，表明在控制高维混杂因素并允许非线性关系存在的条件下，政策冲击通过“舆论关注度—情绪反应”这一链式通道，对资金流变化具有稳定且显著的放大作用。这一结果支持了本文的核心机制假设，即政策信息并非仅通过基本面预期直接影响资金配置，而是显著依赖投资者关注度和情绪的中介传导。

其次，从直接效应 NDE 的估计结果看，其在多数行业中显著小于对应的总效应 TE，且在汽车等行业中，其置信区间覆盖零，显示政策冲击在剔除舆论与情绪通道后的“净效应”较弱。这意味着，对于部分周期性或高度市场化的行业而言，政策信号本身并不足以直接驱动资金流向，其作用更多体现在激发市场情绪、改变投资者预期结构之后，才逐步反映到资金流动中。

进一步比较 TE 与 PIE 的相对大小可以发现，在餐饮、动漫和文娱等行业中，中介路径效应占据了总效应的主要部分，说明这些行业对舆论信息和情绪波动更为敏感；而在旅游和家电等行业中，虽然直接效应和中介效应均为正，但中介通道的重要性相对略弱，反映出不同行业在信息传导机制上的结构性差异。

总体而言，DML 的估计结果不仅在方向上与理论预期保持一致，而且在允许复杂非线性和高维控制变量存在的情况下，提供了更为稳健的因果效应刻画。该结果表明，将政策冲击纳入“舆论—情绪—资金流”的链式中介框架，有助于更全面地理解政策信息在金融市场中的实际传导方式，也为后续将机制分析与预测模型相结合提供了经验依据。

6 总结与讨论

本文围绕“政策与舆论信息是否有助于预测行业 ETF 资金流”这一问题，基于六个行业的日度数据，系统比较了时间序列模型、引入外生信息的扩展模型以及机器学习模型在短期预测中的表现，并在预测分析的基础上，进一步借助 DML 框架对潜在的信息传导机制进行了补充刻画。

首先，从预测结果看，仅依赖历史信息的 ARIMA 模型在多数行业中难以有效刻画资金流的短期波动，其预测路径往往过于平滑，对突发变化的响应不足。在引入政策强度与舆论变量后，ARIMAX 的预测误差在多个行业中明显下降，表明政策与舆论信息在统计意义上为资金流预测提供了重要的增量信息。这一结果说明，行业资金流并非完全由自身历史动态决定，外部信息在短期内具有不可忽视的预测价值。

其次，机器学习模型在多数行业中展现出更强的拟合与泛化能力。无论是 Boosting 还是 Random Forest，其在测试窗口内的 RMSE 与方向预测准确率整体优于传统时间序列模型，尤其在资金流波动幅度较大、非线性特征更明显的行业中优势更为突出。这表明，当政策与舆论变量以非线性方式与资金流发生作用时，基于树模型的机器学习方法能够更充分地利用多维信息，捕捉复杂的交互结构。同时，不同行业中最优模型并不完全一致，反映出资金流生成机制在行业层面具有显著异质性，也提示预测模型的选择需要结合具体应用场景。

在此基础上，本文进一步引入 DML 框架，对政策强度、舆论关注度与情绪变量在统计意义上的作用路径进行了机制层面的补充分析。相关结果显示，在控制市场与宏观信息后，政策变量与资金流之间的关联在很大程度上伴随着舆论关注度与情绪指标的变化，提示信息扩散与投资者情绪可能是政策影响资金流的重要中介通道。需要强调的是，该分析并不以严格的因果识别为目标，而是用于解释为何包含政策与舆论变量的模型在预测中表现更优，从而为预测结果提供机制层面的合理性说明。

总体而言，本文的研究表明，政策与舆论信息不仅在经济含义上具有解释力，也在预测层面展现出切实可用的价值。通过将时间序列模型、机器学习方法与机制分析相结合，本文展示了不同建模思路在金融预测任务中的互补性。未来研究可以在此基础上进一步拓展至更高频数据、跨行业信息溢出以及结构性变化情形，以更全面地评估信息变量在资产流动预测中的作用。

附录

A 东方财富股吧中各行业舆论数据对应标的

表 4: 东方财富股吧中各行业舆论数据对应标的

行业	代表 ETF/基金	指数/板块	具有影响力个股
汽车	of004854，of009067，of012543，of015527	is931008，is931230，if930721，zssz399432	比亚迪（002594.SZ）等
家电	of012461，of013053，of018646，sh561120	is931241，bk0456	美的集团（000333.SZ）等
食品饮料	sh516900，sh515170，sz159843，sz159736	zssh000807，is000807，bk0438，bk0896	贵州茅台（600519.SH）等
旅游	of159766，of562510，sh562510，sz159766	is930633，bk0485，bk0692	中国中免（601888.SH）等
动漫游戏	of012728，of012729，of012768，of012769	is930901	哔哩哔哩（09626.HK）等
文娱传媒	of516190，sh516190	ish30365	泡泡玛特（09992.HK）等

B 变量构造与关键词示例

表 5: 各行业在百度指数中选取的舆论关键词示例

行业名称	关键词 1	关键词 2	关键词 3
新能源汽车	增程式与插电式混动区别	固态电池	续航
家电	智能家居	能效等级	洗衣机
食品饮料	预制菜	轻食	奶茶
旅游	自驾游	民宿	露营
动漫游戏	电竞比赛	游戏攻略	二次元
文娱传媒	潮玩宇宙	首映	手办

C 政策文本权重赋值示例

表 6: 政策文本权重赋值示例

标题	制定机关	公布日期	层级	新颖度	财政性	约束力度	覆盖范围	分析推理
《道路机动车辆生产企业及产品》(第 394 批)，《享受车船税减免优惠的节约能源使用新能源汽车车型目录》(第七十三批),《减免车辆购置税的新能源汽车车型目录》(第十七批)	工业和信息化部	20250521	1	1.3	1.3	1	1.2	三项目录同步发布，涉及新能源汽车税收优惠与生产管理。
关于开展 2025 年新能源汽车下乡活动的通知	工业和信息化部	20250530	1	1.3	1.3	1.3	1	工信部推动新能源汽车下乡，含财政补贴，约束力度较强，覆盖全行业。
关于维护公平竞争秩序,促进行业健康发展的倡议	中国汽车工业协会	20250531	0.6	1.3	1	0.8	1.2	行业协会倡议规范竞争，无财政安排，但覆盖全行业。

D 舆论情感标注示例

表 7: 舆论情感倾向标注示例

作者	评论内容	更新时间	情感评分	分析说明
剑剑牛	你也是小票，今天为何不怎么涨	10 月 31 日	3	语气带质疑但未明确表达强烈不满，整体偏中性或轻微悲观。
波波大兄弟	百将牌不氪金悠闲好玩	10 月 31 日	5	强调游戏“免费、悠闲、好玩”，情绪明显积极。
兴趣使然的韭菜	假突破？	10 月 31 日	1	使用“假突破”暗示行情虚弱，对后市判断明显悲观。