

社会情绪与经济波动的联动机制研究方案

231098150 金浩然

November 2025

Contents

1	社会情绪、宏观经济与沪深 300 波动的联动机制	2
1.1	研究背景和目的	2
1.2	变量说明	2
1.3	总体技术路线	2
1.3.1	数据对齐	2
1.3.2	衍生变量构建（特征工程）	3
1.3.3	宏观剥离与情绪修正（月度）	4
1.3.4	实证模型构建	5

1 社会情绪、宏观经济与沪深 300 波动的联动机制

1.1 研究背景和目的

弹幕、微博、新闻等文本实时承载社会情绪，其量化指标可能对资产价格产生领先或同步影响。我们希望以“沪深 300”为研究对象，同时引入 PMI、PPI、CICSI 等宏观变量及恶性事件频次，回答两个问题：

- (1) 社会情绪能否解释或预测沪深 300 的短期（周度）与中期（月度）波动？
- (2) 情绪与宏观冲击孰轻孰重？是否需要“剔除宏观”后的情绪净指标？

1.2 变量说明

变量符号	变量含义
HS300	沪深 300 日收盘指数
PPI	工业生产者出厂价格指数（上年同月 =100）
PMI	制造业采购经理指数，含累乘
CICSI	投资者情绪指数（标准化-剔除宏观经济因素）
Score	弹幕情绪评分

1.3 总体技术路线

1.3.1 数据对齐

基于前面的环节我们将 B 站弹幕成功的转换成了情绪数据，依照评分模型的差异主要有下述三种情绪数据：

- (1) 基于情感词典的弹幕情绪评分；
- (2) 基于传统机器学习的弹幕情绪评分；
- (3) 基于深度学习轻量级 BERT 的弹幕情绪评分；

这些指标最初是以周度频率计算的。而所能获取到的沪深 300 指数则是传统的日度数据，所以需要对于这些数据进行时间维度上的对齐，我们将其按自然周和自然月进行均值聚合：

周度数据聚合公式：

$$I_w = \frac{1}{7} \sum_{d \in \text{week}(w)} I_t$$

月度数据聚合公式：

$$S_{0,m} = \frac{1}{N_m} \sum_{d \in \text{month}(m)} S_{0,d}$$

对于对齐后的数据若出现数据值的缺失，则考虑采用线性插值的方式填补缺失值：

$$S_{i,j} = \frac{S_{i,j} + S_{i,j}}{2}$$

1.3.2 衍生变量构建（特征工程）

在《微博情感分析综述》等相关研究文献中，研究人员在相关实证研究中，都提出将原始周度情绪指标直接纳入模型往往难以获得稳健且具有经济解释力的结果，其原因可能在于“原始情绪序列”与“资产收益率序列”在统计属性与信息结构层面存在系统性错配：首先，从微观结构视角看，投资者对文本情绪的反应并非机械地依赖于绝对水平，而是对“边际变化”“趋势强度”以及“相对预期偏差”更为敏感；原始周度分值仅提供当期水平的噪声估计，无法捕捉这些跨期比较信息；其次，原始情绪序列通常呈现高波动性、弱平稳性与低信噪比特征，ADF 检验常无法拒绝单位根假设，直接建模易陷入伪回归陷阱，且 Granger 因果检验的功效显著下降；第三，情绪与收益率的领先-滞后关系通常具有时变非线性特征，例如情绪需持续升温数周方能转化为资金流入，而单日极端情绪可能因“过度反应”而被市场快速修正；若不通过滚动平均、差分或变化率等变换加以显式刻画，此类动态结构将被淹没于白噪声中。

因此在上一步将情绪数据和相关指标进行时间维度上进行对齐后，我们借鉴了不同文献中的不同的三种方式构建衍生变量，以捕捉情绪与市场的趋势、波动与反应强度：

（一）滚动平均（4 期/8 期）

原始情绪序列中潜在的白噪声成分高，与沪深 300 的简单相关系数可能会被

被“毛刺”淹没。使用滚动平均的方式，用于平滑短期波动，提取中期趋势：

$$\bar{X}_w = \frac{1}{4} \sum_{k=0}^3 X_{w-k}$$

（二）一阶差分（变化量）

真正影响市场的是“情绪变化”，而非情绪绝对水平，所以我们同样也引入了对于数据的一阶差分处理（计算变化量），用于衡量变量在相邻周期间的绝对变化：

$$\Delta X_w = X_w - X_{w-1}$$

（三）百分比变化率

考虑到沪深 300 的股市收益率本身就是百分比形式，“情绪冲击”与之对话时必须量纲一致，所以核算百分比变化率，用于衡量变量变化的相对幅度，便于与收益率等指标对比：

$$\% \Delta X_w = \frac{X_w - X_{w-1}}{X_{w-1}} \times 100\%$$

1.3.3 宏观剥离与情绪修正（月度）

2021-07 之前的情感得分存在误差：B 站历史弹幕最早只能获取到 2021 年 7 月，因此这之前的弹幕是从 2021 年 7 月开始爬取的，因此这些弹幕不是对应时间内发布的，导致计算的情绪得分存在偏误，所以我们使用国泰安数据库中的投资者情绪指数 CICI（消费者情绪指数，标准化-剔除宏观经济因素）作为“纯净情绪”基准来修正 2021 年 7 月之前的情绪评分，具体做法是将投资者情绪指数的线性趋势叠加到原来的情绪得分上：

$$\text{Score}_t^* = b + (\text{Score}_t - \mu_{\text{pre}})$$

其中：

$$b = \text{CICI}_t \cdot \frac{\text{Score}_\tau - \text{Score}_{\min}}{\text{CICI}_\tau - \text{CICI}_{\min}}$$

注： $\tau = 2021 - 07$ 为断点， μ_{pre} 为 2021 年 7 月前的均值；CICSI: Consumer Internet Consumer Sentiment Index, 网络消费者情绪指数。

1.3.4 实证模型构建

我们计划采用以下 4 种计量模型对情绪与股市之间的关系进行系统检验：

(1) 交叉相关分析（周度）

我们希望通过利用交叉相关分析，来回答“情绪是否存在提前的现象”，也为后续 VAR(p) 模型中的初值选择提供依据。

$$\rho_k = \text{Corr}(\Delta S_{w-k}, \% \Delta I_w), \quad k = 0, 1, 2, 3, 4$$

相关指标：最大交叉相关系数 ρ_{max} 及其滞后阶数 k^*

(2) OLS 回归模型（月度）

而采用 OLS 回归模型则是希望解决“情绪本身对于沪深市场行为是否存在解释力”，来体现情绪对于其的独立贡献是多少。

$$\% \Delta I_m = \alpha + \beta_1 \Delta S_m^* + \beta_2 \Delta \text{PMI}_m + \beta_3 \Delta \text{PPI}_m + \gamma \text{VEI}_m + \varepsilon_m$$

相关检验：t 检验（ β 显著性）、F 检验（整体显著性）、调整 R^2 （相关性）

(3) VAR(p) 模型（月度）

使用 VAR(p) 模型则是希望通过观察得到的 Granger 因果检验和脉冲响应函数（IRF），观察情绪冲击对股市的动态影响路径。

$$\begin{bmatrix} S_m^* \\ I_m \\ \text{PMI}_m \end{bmatrix} = A_1 \begin{bmatrix} S_{m-1}^* \\ I_{m-1} \\ \text{PMI}_{m-1} \end{bmatrix} + \cdots + A_p \begin{bmatrix} S_{m-p}^* \\ I_{m-p} \\ \text{PMI}_{m-p} \end{bmatrix} + u_m$$

采用 Granger 因果检验（“有 X 的过去比没 X 的过去，更能预测 Y 的现在”）：

1. 受限模型（Y 仅用自身滞后预测）

$$Y_t = \alpha_0 + \sum_{i=1}^k \alpha_i Y_{t-i} + \varepsilon_{0t} \quad (1)$$

2. 非受限模型（加入 X 的滞后项）

$$Y_t = \alpha_0 + \sum_{i=1}^k \alpha_i Y_{t-i} + \sum_{j=1}^k \beta_j X_{t-j} + \varepsilon_{1t} \quad (2)$$

3. F 统计量（Wald 形式）

$$F = \frac{(\text{RSS}_R - \text{RSS}_U)/k}{\text{RSS}_U/(T - 2k - 1)} \quad (3)$$

其中：

- RSS_R ：受限模型残差平方和
- RSS_U ：非受限模型残差平方和
- k ：滞后阶数（约束个数）
- T ：样本容量

$$F > F_\alpha(k, T - 2k - 1) \quad \text{或} \quad p\text{-value} < \alpha \Rightarrow \text{拒绝 } H_0, X \text{ Granger-导致 } Y \quad (4)$$

(4) 方差分解（Variance Decomposition）

最后一步的方差分解则是希望在前三步的基础上，得到情绪对股指波动的贡献率曲线，更直观的看到其对于股指波动影响的重要程度。

$$\Sigma_k = LL^\top, \quad \text{其中 } L \text{ 为下三角矩阵}$$