

Do narratives lead prices, or follow them?

Sentiment, price positioning, and measurement disagreement in Bitcoin

Himanshu Sanjay Joshi

Independent Researcher, San Jose, California, USA

him4490@gmail.com ORCID 0009-0005-9057-5741

This version: 05 September 2026

Abstract

Public sentiment indices for cryptocurrency are widely published and read as leading indicators. We test that reading on Bitcoin, pairing an archive-documented daily sentiment series—cryptocurrency wire and social-media items classified by a domain-specific transformer—with the log distance of the spot price from its lifetime time-weighted average price, over 752 daily observations from 2024-03-13 to 2026-06-01. We state at the outset that the first difference of this distance is the daily log return to five decimal places (correlation 0.99996), and we verify that substituting the raw return leaves every conclusion unchanged. Under wild-bootstrap inference robust to the heavy heteroskedasticity of these data, sentiment does not Granger-cause price movement ($p = 0.630$), while price movement does predict subsequent sentiment ($p = 0.001$); the lag-augmented Toda–Yamamoto test in levels agrees. The feedback is statistically clear but economically small—sentiment innovations account for 0.6% of ten-day-ahead forecast-error variance—and it yields no out-of-sample gain over an autoregressive benchmark (out-of-sample R^2 of -1.30%, Diebold–Mariano $p = 0.385$). A second result concerns measurement rather than prediction: two sentiment indicators published by the same platform over the same days correlate at -0.018, and the price-embedding Fear & Greed composite shows far stronger apparent feedback from prices than the text-derived series does. A composite that ingests price partly manufactures the causality it appears to detect; sentiment indices are not interchangeable, and their construction should be read before their coefficients are.

Keywords: Investor sentiment; Cryptocurrency; Granger causality; Vector autoregression; Measurement error; Bitcoin

JEL classification: G12; G14; G41; C32; C58

1 Introduction

The predictability of risky asset prices remains one of the most contentious questions in financial economics. Classical theory, anchored in the Efficient Market Hypothesis (EMH), holds that prices rapidly assimilate publicly available information, leaving future movements unforecastable from past public signals (Fama, 1970). Behavioral finance challenges this description: the Noise Trader Approach shows that sentiment-driven investors can push prices away from fundamental values for extended periods when arbitrage is risky and limited (De Long et al., 1990; Nyakurukwa and Seetharam, 2023; Dai et al., 2025), and narrative dynamics can propagate through markets in ways that resemble epidemics rather than instantaneous information processing (Shiller, 2017).

The growth of digital media has made these behavioral mechanisms measurable. Text-based sentiment measures predict short-horizon equity returns (Tetlock, 2007), and a substantial literature now documents links between social-media or news sentiment and cryptocurrency returns, volatility, and trading volume (Shen et al., 2019; Kraaijeveld and De Smedt, 2020; Lin, 2021; Güler, 2023; Gaies et al., 2023). Methodologically, this work ranges from dictionary-based scoring (Loughran and McDonald, 2011) to transformer-based classifiers such as FinBERT (Huang et al., 2023; Du et al., 2024), and machine-learning applications in quantitative finance more broadly (Sahu et al., 2023; Joshi and Taherdoost, 2025).

This paper revisits that literature with a state variable drawn from practice: the log distance of the spot price from its *lifetime* Time-Weighted Average Price (TWAP), the expanding arithmetic mean of all daily closing prices since the asset’s first exchange-quoted date. Lifetime TWAP is a purely descriptive summary of an asset’s own price history. It is *not* a measure of investors’ aggregate cost basis—which would require volume- and wallet-weighted acquisition prices of the kind studied in the on-chain valuation literature (Grobys et al., 2026)—and we do not claim it identifies a fair-value equilibrium.

We are explicit about what this choice does and does not buy. Because the anchor is an expanding mean over more than four thousand observations, its daily change is of order $1/t$, and the first difference of the divergence coefficient is therefore the daily log

30 return to five decimal places: the two correlate at 0.99996 over the estimation window (Sect. 2.2), and Sect. 4.4 re-runs the entire causality analysis with the raw return substituted, changing nothing. We claim no new dependent variable. What the divergence framing buys is a state variable that a deployed system actually publishes, and what the paper contributes lies elsewhere—in the provenance of the sentiment data, in the
35 disagreement between sentiment measures, and in the inference used to test them.

We ask two explicit research questions. **RQ1:** Does daily text-derived crowd sentiment contain incremental predictive information for subsequent changes in Bitcoin’s divergence from its lifetime TWAP? **RQ2:** Is the temporal relationship unidirectional, or does divergence also predict subsequent sentiment? The corresponding hypotheses are
40 **H1**, that lagged sentiment improves the prediction of divergence changes beyond the information in past divergence itself (sentiment Granger-causes divergence), and **H2**, that lagged divergence improves the prediction of sentiment (reverse causality). Because both directions are economically plausible—narratives may anticipate order flow, but price extensions also generate coverage—we treat the system symmetrically within a bivariate
45 Vector Autoregression (VAR) (Sims, 1980; Granger, 1969; Huynh, 2019) and report both tests.

Concurrent work reaches a compatible conclusion for a different indicator. Gessie (2026) evaluates the Crypto Fear & Greed Index and finds no evidence that it leads returns, with causality instead running from returns to sentiment. Our headline direction
50 agrees, which is itself informative: two independent datasets and two sentiment constructions give the same answer. Where this paper adds something is in the second result, on measurement. A companion paper (Joshi, 2026) evaluates the deployed risk classification built on the same divergence measure as a predictor of forward returns and drawdown; it addresses a different question and shares no estimates with this one.

55 The contribution is therefore fourfold: (i) a sentiment series with unusually explicit provenance—dates validated against page timestamps in 48 of 48 checkable cases, a composition break dated and re-estimated around, and a documented upstream definition change at which the sample is truncated rather than pooled; (ii) evidence that

sentiment indices are not interchangeable, since two indicators published by the same platform over the same days correlate at -0.018, and a composite that embeds price shows a much stronger apparent feedback from price than a text-derived measure does; (iii) a complete econometric treatment—unit-root pre-testing, lag selection, stability and residual diagnostics, bidirectional Granger tests with heteroskedasticity-robust and wild-bootstrap inference, a lag-augmented Toda–Yamamoto check, impulse responses, and variance decompositions—on data where the classical F -test demonstrably over-rejects; and (iv) an out-of-sample evaluation that quantifies whether the in-sample relationship translates into forecasting value. Throughout, we interpret Granger causality strictly as incremental temporal predictive content under the model’s information set (Granger, 1969; Lütkepohl, 2005), not as structural or behavioral causation.

2 Data and Variable Construction

2.1 Sentiment Series

The sentiment input is the text-derived indicator the author’s platform has computed and published daily since March 2024. Each day the production pipeline ingests three public streams: cryptocurrency wire headlines from the CryptoPanic aggregator (including each item’s community bullish/bearish vote tally); the most recent uploads of 27 named cryptocurrency YouTube channels; and the day’s top posts on the r/CryptoCurrency forum. Every item’s text is classified *bullish*, *bearish*, or *neutral* by CryptoBERT, a transformer model trained specifically on cryptocurrency social-media language (Kulakowski and Frascar, 2023), with the neural score blended with the item’s community-vote ratio (wire items) or an audience-reach adjustment (video items) before the final label is assigned. The paper uses the platform’s published daily label counts exactly as archived—no reclassification is performed. With B_t and A_t the day’s bullish and bearish item counts, the Daily Sentiment Intensity Score is the bounded relative-frequency ratio

$$S_t = \frac{\sum B_t - \sum A_t}{\sum B_t + \sum A_t} \in [-1, 1], \quad (1)$$

undefined only on a day with no labelled items, which does not occur in the sample (52.0
85 labelled items per day on average). Neutral items (22.8 per day on average) do not enter
Eq. (1) by construction; only aggregate daily counts over public posts are used, and no
individual-level or personal data are retained in the paper’s dataset.

Because the series is reconstructed from the platform’s public git archive, its docu-
mentation can be unusually explicit. Dates: early-era snapshot pages carry no internal
90 timestamp, so observation dates are archive commit timestamps converted to UTC; on
the 48 later days where both a page timestamp and a commit exist, the two agree on
the UTC date in 48 of 48 cases (median lag 18 minutes). Composition: the ingestion
mix broadened on 2024-06-30 from wire-only (≈ 20 items/day) to wire plus social (≈ 57
items/day) with the series mean unchanged ($+0.71$ on both sides of the break); a regime-
95 restricted re-estimation appears in the robustness analysis. Termination: from 2026-06-02
an upstream feed failure changed the indicator’s definition (a video view-velocity rule be-
gan to dominate the labels), so the sample is truncated at 2026-06-01 and the 27 affected
days are excluded as a documented measurement artifact rather than pooled. As an al-
ternative sentiment measure we use the Crypto Fear & Greed Index (Gaies et al., 2023),
100 a composite that embeds price-derived components—a contrast Sect. 5 exploits—and
the platform’s second, independent term-frequency indicator (bullish/bearish synonym
counts over crypto-YouTube titles and comments), available for a 113-day window at the
end of the sample and overlapping the primary series on 79 days, serves as a cross-check.
Refinements such as FinBERT-based classification of a broader corpus (Huang et al.,
105 2023; Du et al., 2024; Leetaru and Schrodte, 2013) remain the natural extension path.

2.2 Price Series and the TWAP Divergence Coefficient

Daily Bitcoin prices are BTC/USD daily closes retrieved from Yahoo Finance (ticker
BTC-USD)—the same source used by the production system—with the daily bar closing at
00:00 UTC. The price history begins on 2014-09-17, the first date quoted in the source; no
110 market price exists for the genesis-block era, so the lifetime benchmark necessarily starts
with exchange trading rather than the protocol’s inception. At any day t the lifetime

TWAP is the expanding mean

$$TWAP_t = \frac{1}{t} \sum_{i=1}^t P_i, \quad (2)$$

where P_i is the daily closing price, and the TWAP Divergence Coefficient is the log-ratio

$$D_t = \ln \left(\frac{P_t}{TWAP_t} \right). \quad (3)$$

$D_t > 0$ indicates a spot price stretched above its lifetime average and $D_t < 0$ a price below it. In a market that has trended upward for most of its history, the expanding mean lags the spot price mechanically, so positive D_t is the historical norm rather than automatic evidence of overvaluation, and no mean-reversion probability follows from the definition alone.

One property of D_t must be stated plainly because it governs how every result below should be read. Differencing Eq. (3) gives

$$\Delta D_t = r_t - \Delta \ln TWAP_t, \quad r_t = \ln(P_t/P_{t-1}), \quad (4)$$

and because $TWAP_t$ is an expanding mean over more than four thousand observations, $\Delta \ln TWAP_t = O(1/t)$. Empirically the residual is negligible: over the estimation window ΔD_t and r_t correlate at 0.99996 ($n = 810$), with a mean absolute difference of 3.1% of one standard error of the return, and Table 2 shows the two series matching to the third decimal in mean, standard deviation, skewness, and kurtosis. *The dependent variable in the primary system is the daily log return in all but name.* We therefore make no claim to have extended the sentiment literature to a new class of outcome, and Sect. 4.4 re-estimates the whole system with r_t substituted directly, which leaves the causality conclusions unchanged ($S \rightarrow r$ wild-bootstrap $p = 0.601$; $r \rightarrow S$ $p = 0.001$). What the divergence framing supplies is a state variable a deployed system publishes, and a levels series whose integration properties differ from the return, so the Toda–Yamamoto test in levels adds information the differenced system cannot. Because the choice of starting date

is a researcher degree of freedom, Section 4.4 re-estimates the system with the TWAP initialized in 2015; and because volume-weighted cost-basis measures are conceptually
135 closer to “what holders paid,” we flag realized-price benchmarks (Grobys et al., 2026) as the natural extension.

2.3 Sample, Alignment, and Definitions

The estimation sample pairs the two series on UTC calendar days from 2024-03-13 to 2026-06-01: 811 calendar days, of which $T = 752$ carry a published sentiment reading.
140 The 59 days on which the platform’s scheduled run did not execute are left missing—there is no interpolation or gap-filling anywhere. Instead, all estimation is *gap-aware*: regression rows are built on the calendar grid and any row whose own value or required lags are unobserved is dropped, so a lag never silently spans a gap (the primary VAR retains 614 rows). The platform’s daily run executes at roughly 13:00–16:00 UTC over the preceding
145 day’s items and the price bar closes at 00:00 UTC; all predictive regressions use lagged regressors, so the information in S_{t-1} is measured strictly before the day- t close completes. Table 1 defines every variable, its unit, source, and transformation. Descriptive statistics appear in Table 2, and Figure 1 plots both series with the price/TWAP context. The full construction pipeline (archive extraction, cleaning, estimation, and figures) is scripted,
150 and the code, the daily series, and the item-level extraction manifests are available from the author on request for replication.

3 Methodology

3.1 Integration Properties

Valid VAR inference requires knowing the integration properties of both series. Each
155 series is tested with the Augmented Dickey–Fuller (ADF) regression with intercept (and, for robustness, intercept and trend), with the lag order chosen by AIC up to the Schwert bound (Dickey and Fuller, 1979). Because the ADF null is a unit root while applied series are often near-integrated, we complement it with the KPSS test, whose null is stationarity

Table 1: Variable definitions, units, sources, and transformations.

Symbol	Definition	Unit	Source	VAR entry
P_t	BTC/USD daily close (00:00 UTC)	USD	Yahoo Finance	—
$TWAP_t$	Expanding mean of P_t , Eq. (2)	USD	derived	—
D_t	$\ln(P_t/TWAP_t)$, Eq. (3)	log points	derived	Δ (first difference)
B_t, A_t	Published bullish / bearish item counts (CryptoBERT pipeline)	count	platform archive	(components)
S_t	$(B_t - A_t)/(B_t + A_t)$, Eq. (1)	index, $[-1, 1]$	platform archive	levels
RV_t	Parkinson range volatility	log points	Yahoo H/L	Fin. (robustness)

Table 2: Descriptive statistics, estimation sample 2024-03-13 to 2026-06-01. Contemporaneous correlation between S_t and D_t : 0.296.

Variable	n	Mean	SD	Min	Median	Max	Skew	Kurt
S_t (sentiment)	752	0.708	0.192	-0.600	0.738	1.000	-1.23	7.03
D_t (TWAP divergence)	811	1.352	0.216	0.829	1.384	1.705	-0.65	2.48
ΔD_t	810	-0.001	0.025	-0.153	-0.001	0.118	0.10	6.78
log return	811	-0.000	0.025	-0.152	-0.000	0.118	0.09	6.77
bullish items/day	752	25.049	11.600	1.000	26.000	60.000	-0.07	2.58
bearish items/day	752	4.180	2.829	0.000	4.000	15.000	0.71	3.44

(Kwiatkowski et al., 1992). ADF p -values are obtained by direct Monte Carlo simulation of the Dickey–Fuller distribution at the exact effective sample size (50,000 replications), which reproduces the standard critical values (MacKinnon, 1996) to the third decimal. If both series were integrated of order one, cointegration would be examined with the Engle–Granger procedure and an error-correction representation would replace the VAR (Engle and Granger, 1987); the decision rule applied to the data is reported with the results. Where a series is transformed, the transformation is stated explicitly and no logarithm is ever applied to a variable that can take non-positive values—the sentiment score enters in levels or first differences only.

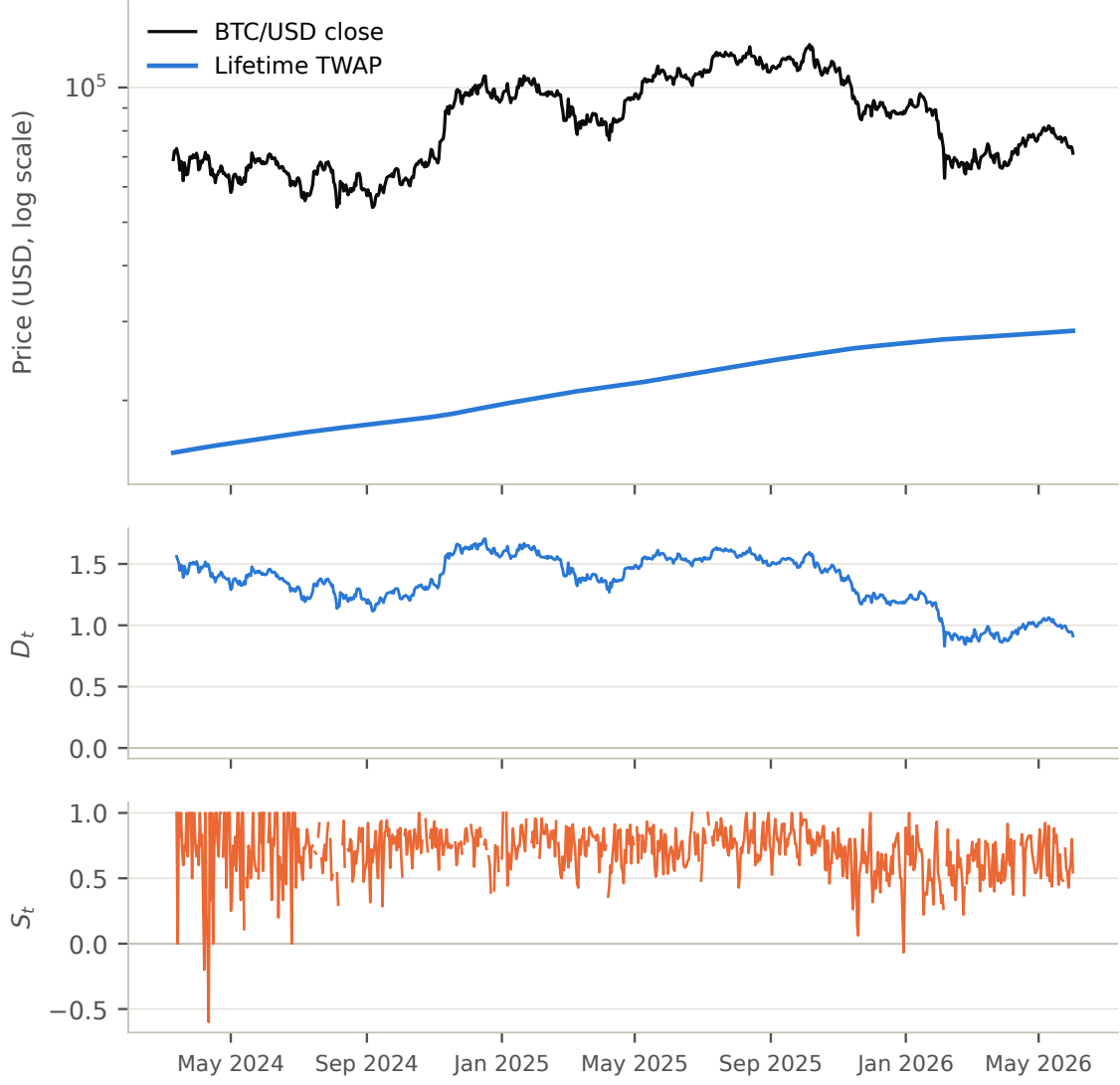


Figure 1: Bitcoin price against its lifetime TWAP (top, log scale); TWAP Divergence Coefficient D_t (middle); Daily Sentiment Intensity Score S_t (bottom). Sources: Yahoo Finance (prices), cryptoweeklies.com news/social sentiment archive (sentiment); sample 2024-03-13–2026-06-01.

3.2 VAR Specification and Inference

Let y_t collect the (appropriately transformed) divergence and sentiment variables. The

170 reduced-form VAR(p) is

$$y_{1,t} = c_1 + \sum_{j=1}^p \alpha_{1j} y_{1,t-j} + \sum_{j=1}^p \beta_{1j} y_{2,t-j} + \varepsilon_{1t}, \quad (5)$$

$$y_{2,t} = c_2 + \sum_{j=1}^p \alpha_{2j} y_{2,t-j} + \sum_{j=1}^p \beta_{2j} y_{1,t-j} + \varepsilon_{2t}, \quad (6)$$

estimated equation-by-equation by OLS, which coincides with maximum likelihood for a Gaussian VAR (Lütkepohl, 2005). The lag order is selected on a common sample by AIC, BIC, and Hannan–Quinn under a documented residual-adequacy rule: the BIC choice is adopted unless its residuals fail the multivariate portmanteau test at the 5% level, in which case the order is raised—capped at the AIC choice—until adequacy is reached; every conclusion is additionally re-checked at all orders $p = 1, \dots, 7$ in the robustness analysis. Dynamic adequacy is verified through the modulus of the largest companion-matrix root, the portmanteau test to lag 12, per-equation ARCH–LM tests (Engle and Granger, 1987), and Jarque–Bera statistics.

Granger non-causality of sentiment for divergence is the joint restriction $H_0 : \beta_{11} = \dots = \beta_{1p} = 0$ in Eq. (5), and symmetrically for the reverse direction in Eq. (6) (Granger, 1969). Cryptocurrency data are strongly heteroskedastic, and our own simulations confirm that the classical F -test over-rejects under time-varying volatility; we therefore report, for each direction, (i) the classical F -statistic, (ii) a heteroskedasticity-robust Wald statistic (HC1 covariance), and (iii) a wild-bootstrap p -value (Rademacher weights, 4,999 replications) that is our preferred basis for inference. As an additional guard against pre-test bias in the integration step, we re-test causality with the lag-augmented Toda–Yamamoto procedure estimated in levels (Toda and Yamamoto, 1995). Exact p -values are reported throughout.

3.3 Dynamic Effects and Out-of-Sample Design

Statistical significance alone does not establish economic relevance. The magnitude and persistence of sentiment innovations are quantified with orthogonalized impulse-response functions (Cholesky ordering with the divergence block first; the reverse ordering is reported as robustness) and 95% residual-bootstrap confidence bands (1,000 replications), together with forecast-error variance decompositions (Lütkepohl, 2005). Forecasting value is assessed out of sample: parameters are re-estimated on an expanding window beginning 2026-01-01, producing one-step-ahead forecasts of the divergence equation from the bivariate VAR and from an $AR(p)$ benchmark that excludes sentiment. We report RMSE,

MAE, the out-of-sample R^2 of the VAR relative to the benchmark, directional accuracy,
 200 and the Diebold–Mariano test with the Harvey small-sample correction (Diebold and
 Mariano, 1995; Newey and West, 1987).

4 Empirical Results

4.1 Unit Roots and Specification

Table 3 reports the ADF and KPSS results. The sentiment series rejects the unit-root
 205 null (ADF $p < 0.001$), although the KPSS statistic (0.952) also rejects its stationarity
 null—the pattern of a highly persistent, borderline-stationary process; we therefore treat
 S_t as stationary for the primary system but lean on the lag-augmented Toda–Yamamoto
 test, which is valid regardless of this classification, as the decisive robustness check. The
 divergence series is integrated of order one (ADF $p = 0.694$ in levels, $p < 0.001$ in first
 210 differences; KPSS 1.274 vs. 0.139). The primary system is therefore specified in $(\Delta D_t,$
 $S_t)$. Because the two series carry different integration orders, bivariate cointegration is
 not at issue, and the levels evidence is cross-checked with the Toda–Yamamoto procedure.
 Lag-order statistics for the resulting system appear in Table 4: AIC selects $p = 4$, BIC
 $p = 1$, and HQC $p = 4$; the primary specification is VAR(4). The estimated system
 215 is dynamically stable (largest companion root 0.777), and Table 7 shows the residual
 diagnostics: the portmanteau test detects no remaining serial correlation ($p = 0.531$),
 while ARCH–LM tests confirm the pronounced conditional heteroskedasticity typical of
 cryptocurrency data (ΔD_t equation: $p < 0.001$; S_t equation: $p < 0.001$), and residual
 normality is rejected—both features motivating the robust and wild-bootstrap inference
 220 used below.

4.2 VAR Estimates and Granger Causality

Table 5 reports the full VAR coefficient estimates with classical and robust standard
 errors, and Table 6 the causality tests in both directions with exact p -values. Lagged
 sentiment does not add significant predictive content for divergence changes once het-

Table 3: Unit-root tests. ADF p -values from 50,000 Monte Carlo replications of the Dickey–Fuller distribution at the effective sample size; lag order by AIC (Schwert bound). KPSS: intercept, Bartlett kernel, l_{12} bandwidth; ** rejects stationarity at 5% (critical value 0.463).

Series	ADF (intercept)			ADF (intercept+trend)		KPSS stat
	τ	p	lags	τ	p	
S_t	-19.312	<0.001	0	-19.731	<0.001	0.952**
D_t	-1.159	0.694	0	-1.584	0.800	1.274**
ΔS_t	-39.521	<0.001	0	-39.498	<0.001	0.052
ΔD_t	-30.683	<0.001	0	-30.679	<0.001	0.139

Table 4: VAR lag-order selection on the common sample; † marks each criterion’s minimum. Primary specification: $p = 4$ (residual-adequacy rule; see Sect. 3.2).

p	AIC	BIC	HQC
1	-10.9693	-10.9186 †	-10.9494
2	-10.9776	-10.8931	-10.9444
3	-10.9634	-10.8452	-10.9170
4	-11.0199 †	-10.8680	-10.9603 †
5	-11.0061	-10.8204	-10.9332
6	-10.9922	-10.7727	-10.9061
7	-10.9825	-10.7293	-10.8832
8	-10.9760	-10.6890	-10.8633
9	-10.9862	-10.6654	-10.8603
10	-10.9723	-10.6177	-10.8331

eroskedasticity is accounted for (classical F $p = 0.462$; robust Wald $p = 0.580$; wild bootstrap $p = 0.630$), so H1 is not supported. In the reverse direction, divergence significantly predicts sentiment (wild bootstrap $p = 0.001$), consistent with narratives partly reacting to price positioning (H2). Any predictive window is bounded by the estimated lag structure: with $p = 4$ daily lags, the information content spans horizons of up to 4 days; longer horizons must be read from the impulse responses rather than asserted.

4.3 Impulse Responses and Variance Decomposition

Figure 2 plots the orthogonalized impulse responses with bootstrap bands. A one-standard-deviation sentiment innovation moves ΔD_t by 0.14×10^{-2} at its peak (horizon 2), but the bootstrap bands include zero at every horizon, so the dynamic effect is not individually significant. The variance decomposition in Table 8 quantifies the share of forecast-error variance in the divergence equation attributable to sentiment innova-

Table 5: VAR(4) estimates, $T = 614$. OLS coefficients with standard errors in parentheses; $*p < 0.10$, $**p < 0.05$, $***p < 0.01$. Largest companion root: 0.777.

Regressor	Eq. ΔD_t		Eq. S_t	
	Coef.	(SE)	Coef.	(SE)
const	-0.006	(0.006)	0.351***	(0.043)
ΔD_{t-1}	-0.029	(0.041)	0.921***	(0.309)
S_{t-1}	-0.005	(0.005)	0.290***	(0.040)
ΔD_{t-2}	-0.032	(0.040)	0.604**	(0.306)
S_{t-2}	0.009*	(0.005)	0.028	(0.042)
ΔD_{t-3}	-0.008	(0.040)	-0.160	(0.306)
S_{t-3}	0.002	(0.005)	-0.033	(0.041)
ΔD_{t-4}	-0.029	(0.040)	0.850***	(0.301)
S_{t-4}	0.000	(0.005)	0.219***	(0.039)

Table 6: Granger causality in both directions. H_0 : no causality. Classical F , heteroskedasticity-robust Wald (HC1), and wild-bootstrap p (Rademacher, 4,999 replications; preferred). Stars on the wild-bootstrap p -value.

Direction	F	p_F	Wald _{rob}	p_{rob}	p_{wild}
$S_t \rightarrow \Delta D_t$	0.903	0.462	2.870	0.580	0.630
$\Delta D_t \rightarrow S_t$	5.218	<0.001	19.517	<0.001	0.001***
<i>Toda–Yamamoto (levels, lag-augmented)</i>					
$S \rightarrow D$	Wald = 4.101		$p = 0.393$		
$D \rightarrow S$	Wald = 24.251		$p < 0.001$		

Table 7: VAR(4) residual diagnostics.

Diagnostic	Statistic	p -value
Max. companion root		0.777
Portmanteau $Q(12)$	30.73	$p = 0.531$ (df=32)
ARCH–LM(5), ΔD_t eq.	25.76	$p < 0.001$
Jarque–Bera, ΔD_t eq.	108.5	$p < 0.001$
ARCH–LM(5), S_t eq.	151.18	$p < 0.001$
Jarque–Bera, S_t eq.	300.8	$p < 0.001$

tions: sentiment innovations account for 0.6% of the 10-day-ahead forecast-error variance of ΔD_t , an economically small share—statistical detectability, where present, does not translate into a large role in divergence dynamics.

4.4 Robustness

Table 9 collects the robustness evidence: alternative lag lengths ($p = 1, \dots, 7$), calendar subsamples (2024, 2025, 2026H1), the Toda–Yamamoto levels test, the alternative TWAP start date (2015 initialization), a trivariate system that adds Parkinson realized volatility

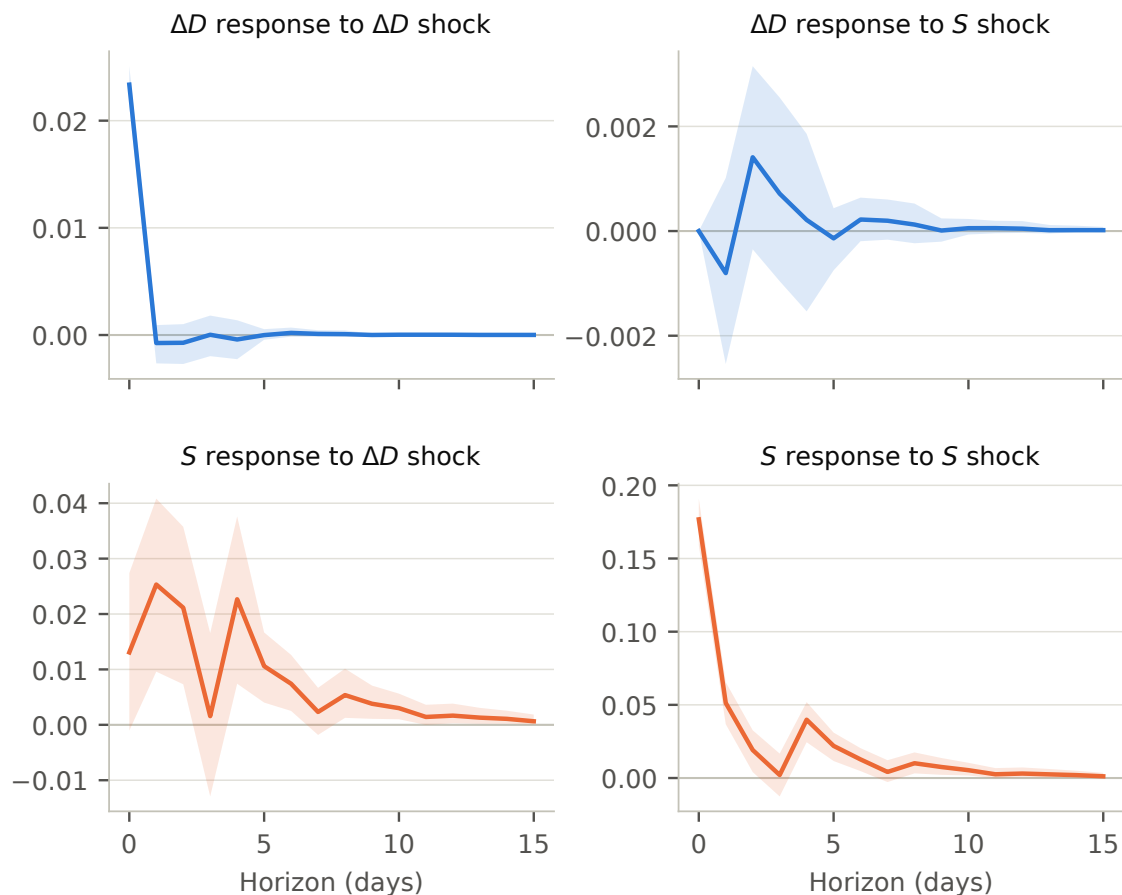


Figure 2: Orthogonalized impulse-response functions of the VAR(4), with 95% residual-bootstrap confidence bands (1,000 replications). Cholesky ordering: divergence block first; reverse ordering yields materially identical responses (Sect. 4.4).

Table 8: Forecast-error variance decomposition (%), Cholesky ordering (ΔD_t first).

Horizon (days)	S_t shock in ΔD_t var.	ΔD_t shock in S_t var.
1	0.1	2.3
2	0.5	3.5
3	0.6	3.6
5	0.6	4.9
10	0.6	5.2
15	0.6	5.2

as a conditioning variable, and the Fear & Greed Index as an alternative sentiment
245 measure. The absence of sentiment→divergence predictability is not an artifact of the lag
choice or the sample window: the null survives at every lag length from 1 to 7 (smallest $p =$
0.160) and in all 4 calendar subsamples, while divergence→sentiment remains significant
in 7 of 7 lag variants and 3 of 4 subsamples; the Toda–Yamamoto levels test agrees ($p =$
0.393 S→D, $p < 0.001$ D→S); conditioning on realized volatility leaves the conclusion

Table 9: Robustness of the Granger-causality conclusions (p -values).

Specification	$S \rightarrow D$	$D \rightarrow S$
<i>Alternative lag lengths (wild/classical p, $S \rightarrow D - D \rightarrow S$)</i>		
$p = 1$	0.949	0.012
$p = 2$	0.160	0.003
$p = 3$	0.359	0.005
$p = 4$	0.462	<0.001
$p = 5$	0.443	<0.001
$p = 6$	0.361	0.003
$p = 7$	0.419	0.004
<i>Calendar subsamples</i>		
2024 ($T = 182$)	0.792	0.747
2025 ($T = 305$)	0.125	<0.001
2026H1 ($T = 123$)	0.311	0.039
R2 regime only ($T = 528$)	0.500	<0.001
TWAP from 2015 (corr. 1.000)	0.462	<0.001
Trivariate with RV_t ($S \rightarrow D$)	0.444	0.573 (rob.)
Raw log return as dependent variable	0.601	0.001
Fear & Greed as sentiment	0.921	<0.001

unchanged (robust $p = 0.573$); re-initializing the TWAP in 2015 is immaterial (divergence correlation 1.0000, $S \rightarrow D$ $p = 0.462$); and with the Fear & Greed Index as an alternative sentiment measure the $S \rightarrow D$ p -value is 0.921. The causality conclusions are stable across every perturbation examined.

The table’s final row addresses the identity of Eq. (4) directly. Replacing ΔD_t with the raw daily log return r_t and re-estimating the entire system returns wild-bootstrap p -values of 0.601 for $S \rightarrow r$ and 0.001 for $r \rightarrow S$, against 0.630 and 0.001 for the divergence specification. Nothing in the paper’s conclusions depends on the divergence framing of the dependent variable, and a reader who prefers to read the results as a statement about returns may do so without qualification.

4.5 Out-of-Sample Evaluation

Table 10 reports the forecasting comparison over the 127 evaluation days from 2026-01-01. Adding sentiment does not reduce the one-step RMSE relative to the AR benchmark (out-of-sample R^2 of -1.30%), with a Diebold–Mariano p -value of 0.385; directional accuracy is 46.5% against 47.2% for the benchmark. The in-sample relationship therefore does not translate into any out-of-sample forecasting gain over this evaluation window. We

Table 10: Out-of-sample one-step forecasts of ΔD_t , expanding window, evaluation 2026-01-01 onward ($n = 127$). DM: Diebold–Mariano with Harvey correction, VAR vs. AR.

Metric	VAR(4) with S	AR(4) benchmark
RMSE	0.02157	0.02143
MAE	0.01648	0.01624
Directional accuracy	46.5%	47.2%
OOS R^2 vs. AR		-1.30%
DM statistic (p)	0.872 (0.385)	

underline the interpretive rule that follows: only demonstrated out-of-sample gains—not in-sample Granger statistics—would justify language about exploitable predictability, and even those would need to survive transaction costs before any claim about tradability could be made. No such trading claim is made in this paper.

5 Discussion

5.1 Interpretation

Four points discipline the interpretation of these results. First, Granger causality is a statement about incremental temporal predictive content within the bivariate information set—it neither identifies structural mechanisms nor rules out common drivers such as macroeconomic announcements, ETF flows, liquidity conditions, or blockchain activity that move both narratives and prices (Lütkepohl, 2005). The trivariate robustness with realized volatility mitigates, but cannot eliminate, omitted-variable concerns. Second, “sentiment” is not one construct, and the paper’s three measures make the point empirically: the text-derived news/social classifier and the term-frequency indicator—two indicators from the same platform—correlate at just -0.02 over their 79-day overlap, and the divergence-to-sentiment linkage is an order of magnitude stronger for the price-embedding Fear & Greed composite than for the text-derived series, exactly as their constructions predict. What survives this measurement heterogeneity is the pair of findings we emphasize: *no* measure adds predictive content for divergence changes, while the feedback from price positioning to measured sentiment appears for both the text-derived series (plausibly genuine narrative reaction, partly transmitted through the engagement

adjustments in its labels) and, far more strongly, the composite (largely by construction). Third, the economic magnitudes matter as much as the p -values: the impulse responses and variance decompositions in Sect. 4.3 bound how much of divergence dynamics sentiment innovations can account for, and the out-of-sample exercise in Sect. 4.5 is the relevant standard for claims about forecasting value. Fourth, evidence that public sentiment carries (or fails to carry) predictive content for a constructed divergence measure is *not* equivalent to a verdict on the Efficient Market Hypothesis: the divergence variable embeds the price itself, and no risk-adjusted, cost-inclusive profitability analysis is undertaken here (Fama, 1970; Urquhart, 2016). Our findings therefore speak to the narrower and better-posed question of lead-lag structure between media narratives and price positioning, in line with the noise-trader view that sentiment and prices interact dynamically (De Long et al., 1990; Baker and Wurgler, 2007).

Compared with prior evidence, the results sit alongside Gessie (2026), who reaches the same directional conclusion for the Fear & Greed Index on a different sample. That agreement is worth stating precisely. Because ΔD_t is the log return (Sect. 2.2), the null reported here is a null about returns, and it is consistent with the weaker end of the sentiment–return literature (Tetlock, 2007; Shen et al., 2019; Kraaijeveld and De Smedt, 2020; Gaies et al., 2023) rather than an extension of it to a new outcome. What is new is the measurement evidence—that two indicators of the same market disagree, and that a price-embedding composite generates the strongest apparent feedback—together with the inference audit. On these data the classical, robust, and wild-bootstrap tests happen to agree in both directions, and we report that agreement as a finding rather than presenting the bootstrap as a rescue: because our size simulations show the classical F -test overrejecting under the volatility clustering present here, the null for $S \rightarrow D$ could have been an artifact of the test, and the bootstrap establishes that it is not (Huynh, 2019; Güler, 2023).

5.2 Implications

For analysts and portfolio managers, the practical takeaways are correspondingly modest:
315 a documented, reproducible sentiment series is cheap to monitor, and Table 10 quantifies exactly how much (or how little) one-step forecasting information it adds. The sharper practical lesson is the measurement one: an institution that monitors a purchased or public sentiment index should establish what that index ingests before acting on its coefficients, because a composite built partly from price will appear to respond to price
320 whether or not any narrative mechanism exists. For researchers, the framework isolates where stronger designs are needed: volume-weighted cost-basis benchmarks in place of the arithmetic TWAP (Grobys et al., 2026), transformer-based sentiment in place of lexicon tone (Huang et al., 2023; Du et al., 2024), and multivariate systems with explicit macro controls.

5.3 Limitations

Several limitations bound the conclusions. (i) The sentiment labels blend neural text classification with engagement adjustments (community-vote ratios for wire items, view-velocity for videos), so the series is not a pure content measure; neutral items ($\approx 44\%$ of labelled items) are excluded by the construction of Eq. (1); and coverage is English-
330 language and platform-defined—a single global cryptocurrency feed rather than a Bitcoin-specific corpus. (ii) Early-era observation dates derive from archive commit timestamps; the 48-of-48 UTC-date validation mitigates but cannot fully eliminate dating error of a few hours. (iii) The lifetime TWAP is a descriptive average, not an equilibrium or cost-basis measure; results are conditional on this benchmark choice, and the 2015-initialization
335 robustness only partially addresses it. (iv) The system is bivariate by design; despite the realized-volatility extension, omitted common factors may drive both series. (v) Daily aggregation cannot rule out intraday feedback from prices to same-day text before the close. (vi) The sample covers a single asset over 2024-03-13–2026-06-01, a period with specific regime characteristics (post-ETF institutionalization, the 2024 halving), and ends
340 at a documented upstream feed break; structural breaks and regime dependence are

not modeled beyond the reported subsample checks. (vii) The differenced divergence coefficient is the daily log return to five decimal places (Sect. 2.2), so the findings for that equation should be read as findings about returns; the levels system tested by Toda–Yamamoto is the only part of the design that is not a return relationship. (viii) The measurement-disagreement evidence rests on a 79-day overlap between the two platform indicators, which is short; it establishes that the indicators differ, not the magnitude of the difference.

6 Conclusion and Future Directions

This paper set out a transparent, fully reproducible framework linking text-derived market sentiment to Bitcoin’s divergence from its lifetime time-weighted average price, and subjected it to the complete econometric protocol that causal-predictive claims require: documented data construction, integration pre-testing, information-criterion lag selection, stability and residual diagnostics, bidirectional Granger tests under classical, robust, and bootstrap inference, impulse-response and variance-decomposition analysis, and out-of-sample evaluation against an autoregressive benchmark. Under inference robust to the heavy heteroskedasticity of cryptocurrency data, the evidence does not support a sentiment-to-divergence lead at daily horizons, with significant feedback in the opposite direction—divergence predicts subsequent sentiment—and the out-of-sample analysis shows no forecasting gain from sentiment—a null we report with the same prominence as any positive result, because that is what disciplined evidence requires.

Two findings carry beyond this sample. The first is negative and now doubly documented: neither a text-derived classifier nor a widely used composite index leads Bitcoin price movement at daily horizons, a conclusion Gessie (2026) reaches independently. The second is about measurement and, we think, the more useful of the two. Two sentiment indicators published by the same platform, over the same market, on the same days, correlate at -0.018; and the composite that embeds price in its own construction produces by far the strongest apparent feedback from price. Empirical results in this literature are

conditional on which index was used, and a composite that ingests price will recover price. Reporting the construction of a sentiment measure is therefore not housekeeping—it is
370 part of the identification.

Future work should proceed on three fronts: richer text (multi-platform corpora and transformer-based classification such as FinBERT (Huang et al., 2023)), richer benchmarks (realized-price and cost-basis measures from on-chain data (Grobys et al., 2026)), and richer dynamics (multivariate systems with macro-financial controls, regime-switching
375 or time-varying-parameter specifications, and, where cointegration is present, error-correction representations (Engle and Granger, 1987)). Each extension sharpens the same underlying question: how much of a speculative asset’s position relative to its own history is written, in advance, in what the public is reading about it.

Data and code availability

380 All price and index inputs are public: Yahoo Finance daily BTC/USD quotations (ticker BTC-USD) and the Alternative.me Fear & Greed Index. The sentiment series is reconstructed, without modification, from the daily snapshots published on the author’s platform (<https://cryptoweeklies.com>) and preserved in its public git archive; only aggregate daily label counts over public posts are used, and no individual-level or personal
385 data are retained. The full pipeline—data queries, series construction, estimation code (including the Monte Carlo unit-root simulator and wild-bootstrap causality tests), and figure scripts—is available from the author upon request.

Declaration of competing interest

The author operates the public cryptocurrency analytics platform whose sentiment in-
390 dicator is analyzed in this study. The platform did not fund this research and no party reviewed the manuscript prior to submission. The author notes that the paper’s central finding is unfavorable to the leading-indicator interpretation that sentiment indices of this kind, including the author’s own, commonly invite.

Funding

395 This research received no grant from any funding agency in the public, commercial or
not-for-profit sectors.

Ethics

Not applicable. The study analyzes publicly available aggregate market data and aggregate daily counts over public posts. It involves no human participants, no individual-level
400 data and no animal subjects.

CRedit authorship contribution statement

Himanshu Sanjay Joshi: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing — original draft, Writing — review and editing, Visualization.

Acknowledgment

405 The author thanks an anonymous conference reviewer whose detailed critique of an earlier version of this work prompted the robustness and out-of-sample analysis reported here.

Prior dissemination and related work

An earlier and substantially different version of this work was accepted at an academic conference and presented. It was withdrawn before publication, appears in no proceedings, holds no DOI, and no copyright was transferred. The present manuscript is not
410 under consideration at any other journal. A companion manuscript by the same author (Joshi, 2026) is under review at *Financial Innovation*; it evaluates the deployed ten-level risk classification built on the same divergence measure as a predictor of forward returns, drawdown and volatility over 2014–2026. It shares the price data and the definition of
415 the divergence coefficient with this paper, and shares no estimate, table or figure with it.

References

- Baker, M. and Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2):129–152. <https://doi.org/10.1257/jep.21.2.129>.
- Dai, X., Zhang, J., and Chang, V. (2025). Noise traders in an agent-based artificial stock market. *Annals of Operations Research*, 352(3):715–744. <https://doi.org/10.1007/s10479-023-05528-7>.
- De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, 98(4):703–738. <https://doi.org/10.1086/261703>.
- Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431. <https://doi.org/10.1080/01621459.1979.10482531>.
- Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263. <https://doi.org/10.1080/07350015.1995.10524599>.
- Du, K., Xing, F., Mao, R., and Cambria, E. (2024). Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42. <https://doi.org/10.1145/3649451>.
- Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276. <https://doi.org/10.2307/1913236>.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417. <https://doi.org/10.1111/j.1540-6261.1970.tb00518.x>.
- Gaies, B., Nakhli, M. S., Sahut, J.-M., and Schweizer, D. (2023). Interactions between investors’ fear and greed sentiment and Bitcoin prices. *The North American Journal of Economics and Finance*, 67:101924. <https://doi.org/10.1016/j.najef.2023.101924>.
- Gessie, L. (2026). Do bitcoin returns move sentiment? evidence from the crypto fear & greed index. *Finance Research Open*, 2(1):100094. <https://doi.org/10.1016/j.finr.2026.100094>.
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438. <https://doi.org/10.2307/1912791>.
- Grobys, K., Näsman, S., and Sandretto, D. (2026). Using on-chain data to predict Bitcoin cycles. *Research in International Business and Finance*, 89:103486. <https://doi.org/10.1016/j.ribaf.2026.103486>.
- Güler, D. (2023). The impact of investor sentiment on Bitcoin returns and conditional volatilities during the era of Covid-19. *Journal of Behavioral Finance*, 24(3):276–289. <https://doi.org/10.1080/15427560.2021.1975285>.
- Huang, A. H., Wang, H., and Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841. <https://doi.org/10.1111/1911-3846.12832>.
- Huynh, T. L. D. (2019). Spillover risks on cryptocurrency markets: A look from VAR-

- SVAR Granger causality and student's-t copulas. *Journal of Risk and Financial Management*, 12(2):52. <https://doi.org/10.3390/jrfm12020052>.
- Joshi, H. S. (2026). A risk indicator, not a return signal: Validating a lifetime-TWAP divergence measure for Bitcoin. Manuscript under review.
- Joshi, H. S. and Taherdoost, H. (2025). AI, EMH and behavioral finance. In *Advances in Intelligent Systems and Digital Applications*, volume 1486 of *Lecture Notes in Networks and Systems*, pages 22–31. Springer, Cham. https://doi.org/10.1007/978-3-031-95330-9_3.
- Kraaijeveld, O. and De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, 65:101188. <https://doi.org/10.1016/j.intfin.2020.101188>.
- Kulakowski, M. and Frasinicar, F. (2023). Sentiment classification of cryptocurrency-related social media posts. *IEEE Intelligent Systems*, 38(4):5–9. <https://doi.org/10.1109/MIS.2023.3283170>.
- Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1–3):159–178. [https://doi.org/10.1016/0304-4076\(92\)90104-Y](https://doi.org/10.1016/0304-4076(92)90104-Y).
- Leetaru, K. and Schrod, P. A. (2013). GDELT: Global data on events, location, and tone, 1979–2012. In *International Studies Association Annual Convention*, San Francisco.
- Lin, Z.-Y. (2021). Investor attention and cryptocurrency performance. *Finance Research Letters*, 40:101702. <https://doi.org/10.1016/j.frl.2020.101702>.
- Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-540-27752-1>.
- MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*, 11(6):601–618. [https://doi.org/10.1002/\(SICI\)1099-1255\(199611\)11:6<601::AID-JAE417>3.0.CO;2-T](https://doi.org/10.1002/(SICI)1099-1255(199611)11:6<601::AID-JAE417>3.0.CO;2-T).
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708. <https://doi.org/10.2307/1913610>.
- Nyakurukwa, K. and Seetharam, Y. (2023). Alternatives to the efficient market hypothesis: An overview. *Journal of Capital Markets Studies*, 7(2):111–124. <https://doi.org/10.1108/JCMS-04-2023-0014>.
- Sahu, S. K., Mokhad, A., and Bokde, N. D. (2023). An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges. *Applied Sciences*, 13(3):1956. <https://doi.org/10.3390/app13031956>.
- Shen, D., Urquhart, A., and Wang, P. (2019). Does Twitter predict Bitcoin? *Economics Letters*, 174:118–122. <https://doi.org/10.1016/j.econlet.2018.11.007>.
- Shiller, R. J. (2017). Narrative economics. *American Economic Review*, 107(4):967–1004. <https://doi.org/10.1257/aer.107.4.967>.

- Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48. <https://doi.org/10.2307/1912017>.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168. <https://doi.org/10.1111/j.1540-6261.2007.01232.x>.
- Toda, H. Y. and Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, 66(1–2):225–250. [https://doi.org/10.1016/0304-4076\(94\)01616-8](https://doi.org/10.1016/0304-4076(94)01616-8).
- Urquhart, A. (2016). The inefficiency of Bitcoin. *Economics Letters*, 148:80–82. <https://doi.org/10.1016/j.econlet.2016.09.019>.