

Do Equities and Bitcoin Run on the Same Fuel? A Modality-Controlled Comparison of Price, Sentiment, and Activity Signals for Next-Day Return Prediction

Darsh Gupta
drshgpt09@gmail.com

ABSTRACT

Machine-learning research on equity return prediction and on cryptocurrency return prediction has developed as two largely disjoint literatures, each with its own datasets, feature conventions, and architectures. This methodological separation makes it impossible to determine whether the two asset classes are driven by the same informational modalities, or whether—as is widely conjectured—cryptocurrencies are disproportionately attention-driven while equities remain anchored to price momentum. A modality, in this paper, simply means a category of input information: how the price has moved, what the news says, or how much real economic activity underlies the asset.

We address this question through a modality-controlled experimental design, in which an identical modeling pipeline is applied to a large-capitalization equity (AAPL) and to Bitcoin (BTC-USD) over a common two-year daily sample (6 July 2024–6 July 2026; $n = 499$ and $n = 730$ usable observations, respectively). Three feature modalities are constructed symmetrically for both assets: *price* (moving averages, RSI, MACD, realized volatility), *sentiment* (news headlines scored for positive or negative tone and aggregated to a daily value), and *activity* (blockchain telemetry—active addresses, transaction count, hash rate—for Bitcoin, and volume-based order-flow proxies for the equity).

A stacked LSTM is trained under a price-only baseline and under three additive ablations. An LSTM, or Long Short-Term Memory network, is a neural network designed for data that arrives in time order; it processes each day in sequence while carrying forward a memory of earlier days, which lets it learn patterns that unfold over time. Additive fusion refers to the practice of combining several information types into one model by simply adding all of their features together, and an ablation is the procedure of adding or removing one component at a time to isolate how much that component contributes. A LightGBM surrogate with SHAP attribution ranks modality importance, and Granger causality tests assess cross-asset predictive spillover. SHAP is a method that assigns each input feature a share of the credit for a model's predictions, so that the features can be ranked by how much they actually mattered. A Granger causality test asks whether knowing one series helps predict another beyond what the second series' own past already explains.

Oct 2026
Vol 12. No 2.

Oxford Journal of Student Scholarship
www.oxfordjss.org

Three findings emerge. First, contrary to the prevailing framing of the multimodal fusion literature, additive fusion *degraded* out-of-sample performance for both assets: the price-only baseline attained the highest directional accuracy (52.9% for AAPL, 50.5% for BTC-USD) while full fusion attained the lowest (42.6% and 45.6%), with root-mean-square error inflating by 49.1% and 66.9% respectively. Directional accuracy is the share of days on which the model correctly predicted whether the price would rise or fall, so 50% is equivalent to guessing. We emphasize—and quantify—that none of these accuracies is statistically distinguishable from chance at conventional levels given the available test-set sizes, and we argue this null-result framing is itself the appropriate scientific conclusion.

Second, attribution is markedly asset-heterogeneous, but *not* in the direction the attention-driven hypothesis predicts: momentum and money-flow constructs (MACD, the money-flow proxy, relative volume, RSI) dominate for the equity, whereas the three highest-ranked features for Bitcoin are all on-chain activity measures (hash rate, transaction count, unique active addresses). On-chain data refers to statistics read directly from Bitcoin's public transaction ledger, a class of information that has no equivalent for a company's stock. For *both* assets the two sentiment features rank last, with attribution distributions collapsed near zero—an attribution-side result that is mutually consistent with the ablation-side failure of fusion.

Third, spillover is nonetheless asymmetric: Bitcoin news sentiment Granger-causes next-day equity returns at lags one and two ($p = 0.018$, $p = 0.034$), while the reverse direction is insignificant at every tested lag ($p > 0.24$). Taken together, these results argue that multimodal fusion should be treated as an empirical hypothesis to be tested per asset and per corpus rather than an assumed improvement; that Bitcoin's daily predictability, to the extent it exists, is grounded in settlement-layer activity rather than in news polarity; and that any sentiment channel between the two markets operates cross-sectionally rather than within-asset.

Keywords: return prediction; multimodal fusion; sentiment analysis; on-chain analytics; LSTM; SHAP; Granger causality; cross-market spillover; cryptocurrency

INTRODUCTION

Short-horizon return prediction is among the most intensively studied problems in empirical finance, and the introduction of machine learning has substantially expanded the space of hypotheses that can be tested. A return is simply the percentage change in an asset's price from one day to the next, and short-horizon prediction means forecasting that change one day ahead rather than over months or years. Gu et al. (2020) demonstrate that regularized ensembles and neural networks recover economically meaningful cross-sectional predictability in U.S. equities, and Fischer and Krauss (2018) show that recurrent architectures extract temporal structure from daily equity returns that linear specifications do not. In parallel, an expanding cryptocurrency-forecasting literature routinely combines historical price data, technical indicators, blockchain telemetry, and news or social sentiment into hybrid deep-learning pipelines.

These two literatures have, however, developed in near-complete isolation. A given study selects a single asset class, constructs a bespoke feature set, and evaluates a bespoke architecture. As a consequence, a foundational comparative question remains unresolved: *do equities and cryptocurrencies respond to the same informational modalities?* The widely repeated intuition—that crypto markets are narrative- and attention-driven while equity markets are anchored to price momentum—is asserted far more often than it is tested, precisely because no study places both asset classes under an identical experimental protocol.

Terminology

Because this paper draws on vocabulary from both quantitative finance and machine learning, the central terms are defined here for readers approaching from either direction.

- **Modality.** A category of input information. This paper uses three: price history, news sentiment, and real economic activity.
- **Additive fusion.** Combining several modalities in one model by pooling all of their features together. The multimodal fusion literature generally assumes this improves prediction; testing that assumption is the purpose of this study.
- **Ablation.** Adding or removing one component at a time and re-measuring performance, so that the contribution of each component can be isolated.
- **LSTM (Long Short-Term Memory).** A neural network built for sequential data. It reads observations in time order and maintains an internal memory, with learned gates that decide what information to retain and what to discard. This suits financial data, where the order of events carries meaning.
- **LightGBM.** A gradient-boosted decision tree model. It builds many small decision trees in sequence, each correcting the errors of the previous ones. It performs well on tabular data and is used here for interpretation rather than prediction.
- **SHAP (SHapley Additive exPlanations).** A method, adapted from cooperative game theory, that distributes credit for a model's predictions among its input features. It allows features to be ranked by how much they genuinely influenced the output.
- **Granger causality.** A statistical test of whether the past values of one series improve prediction of another series beyond that series' own history. Despite the name, it establishes predictive precedence, not true causation.
- **Directional accuracy.** The proportion of days on which the model correctly predicted the sign of the return. A value of 50% corresponds to random guessing.
- **On-chain data.** Statistics recorded directly on a public blockchain, such as the number of active addresses or transactions. Equities have no comparable public ledger.

Three Candidate Modalities

It is useful to organize the predictor space into three conceptually distinct modalities. The first is *price*: the trend, momentum, and volatility structure of the asset's own history, formalized in the technical-indicator tradition. The second is *sentiment*: the polarity of contemporaneous news coverage. Tetlock (2007) established that media tone carries incremental information for equity returns, and Bollen et al. (2011) extended the argument to social-media mood. The third is *activity*: whether observable economic usage supports a price move, or whether the move is unsupported by transactional substance. Bitcoin renders this modality unusually observable, since its settlement layer is public—active addresses, transaction counts, and aggregate hash rate are directly measurable. Equities admit no on-chain analogue, so volume-normalized and money-flow proxies constructed from OHLCV data serve as the closest available substitute.

Fusion studies that combine these modalities typically report improvements over price-only baselines. That empirical record must, however, be read against a well-documented publication asymmetry: negative fusion results are rarely reported, and almost all published evaluations are single-asset, making it impossible to separate a genuine modality effect from asset-specific idiosyncrasy or dataset-specific tuning.

Research Questions

We address three questions under a single controlled protocol:

- Does the incremental addition of sentiment and activity modalities improve out-of-sample next-day return prediction relative to a price-only baseline, and does the answer differ by asset class?
- Which features drive next-day directional predictions for each asset under model-agnostic attribution, and is the attribution profile asset-heterogeneous in the direction the attention-driven hypothesis predicts?
- Does sentiment in one market carry incremental predictive content for next-day returns in the other, and is any such relationship symmetric?

Contributions

This paper makes four contributions. First, it provides a *modality-controlled cross-asset design*: the architecture, feature-construction logic, training protocol, split scheme, and evaluation metrics are held fixed across an equity and a cryptocurrency, so that observed differences are attributable to the assets rather than to divergent experimental choices.

Second, it reports a *quantified negative result on additive fusion*. Fusion monotonically degraded prediction error and reduced directional accuracy below the coin-flip threshold for both assets. Unlike much of the fusion literature, every accuracy figure is accompanied by an exact binomial confidence interval, showing that no configuration is statistically separable from chance on the available test sets. This is treated as the paper's central methodological message rather than as an inconvenience.

Third, it supplies *modality-level evidence of asset heterogeneity that runs against the prevailing narrative*. SHAP attribution on a common surrogate shows momentum dominance for the equity and on-chain activity dominance for Bitcoin, with sentiment ranked last for both assets. This contradicts the widely asserted claim that cryptocurrency is hype-driven at the daily within-asset horizon, and replaces it with a sharper alternative: Bitcoin's distinguishing predictive modality is settlement-layer activity, a covariate class with no equity analogue.

Fourth, it documents an *asymmetric spillover within the same framework*. A statistically significant unidirectional Granger relationship runs from Bitcoin sentiment to next-day equity returns, complementing the econometric connectedness literature with a modality-resolved, prediction-oriented result.

RELATED WORK

Machine Learning for Return Prediction

The modern literature on machine learning in asset pricing is anchored by Gu et al. (2020), who benchmark a wide model class on U.S. equity returns and attribute much of the performance gain to the ability of trees and neural networks to capture nonlinear interactions among predictors. Fischer and Krauss (2018) apply LSTM networks to S&P 500 constituents and report that recurrent memory captures short-horizon dependence beyond that available to feedforward or logistic baselines. López de Prado (2018) provides the methodological counterweight, cataloging the ways in which naive cross-validation, sample overlap, and repeated backtesting inflate apparent performance in financial machine learning. Gradient-boosted decision trees, and in particular the histogram-based LightGBM implementation of Ke et al. (2017), remain the default for tabular financial prediction because of their tolerance for mixed scales, missingness, and high feature counts.

Textual Sentiment as a Predictor

Tetlock (2007) showed that pessimism in a widely read financial column predicts downward pressure on market prices with subsequent reversal, establishing media tone as a measurable state variable rather than noise. Bollen et al. (2011) extended the construct to aggregate social mood. Lexicon-based scoring remains widely used for headline-scale text: Hutto and Gilbert (2014) introduced VADER, a rule- and lexicon-based sentiment model calibrated on social-media register. VADER assigns each piece of text a single polarity score between -1 (strongly negative) and $+1$ (strongly positive) using a hand-built dictionary of sentiment-bearing words and grammatical rules, so it requires no training data of its own. It is attractive for short, high-volume, domain-general text because it requires no supervised fine-tuning and yields a bounded compound polarity score. Its limitations—weak handling of finance-specific idiom and no discourse-level modeling—are well documented and are relevant to our interpretation in the Discussion.

On-Chain Analytics

The public settlement layer of permissionless blockchains exposes a class of covariates with no equity analog. Active-address counts, transaction throughput, and hash rate have been studied as proxies for network adoption, transactional demand, and producer commitment respectively. Hash rate measures the total computing power devoted to securing the Bitcoin network, and therefore reflects how much capital miners have committed to it. Because these series are observable at daily frequency and are not derived from price, they constitute a genuinely independent modality rather than a transformation of the price process—which is precisely why the equity side of any symmetric comparison requires a deliberate proxy choice.

Cross-Market Connectedness and Spillover

Corbet et al. (2018) document that cryptocurrencies are relatively isolated from conventional assets under normal conditions, with integration intensifying episodically. Diebold and Yilmaz (2012) provide the canonical directional-connectedness framework, decomposing forecast-error variance to measure which market transmits and which receives. This literature is predominantly volatility- and return-based; it does not resolve which *informational modality* drives each market day to day, nor does it typically embed the spillover test inside the same predictive pipeline used to model each market. RQ3 is designed to close that gap at the level of sentiment specifically.

Interpretability

Because deep recurrent models resist direct feature attribution, surrogate-based explanation is standard practice. A surrogate model is a simpler, more interpretable model trained on the same inputs, used to explain a complex model that cannot be inspected directly. Lundberg and Lee (2017) unify additive feature-attribution methods under the SHAP framework, with an exact polynomial-time solution for tree ensembles. Our use of a LightGBM surrogate for attribution follows this convention and is explicitly disclosed as an approximation below.

PROBLEM FORMULATION

Target Variable

Let $P_t^{(a)}$ denote the adjusted closing price of asset $a \in \{AAPL, BTC - USD\}$ on trading day t . The simple next-day return is

$$r_{t+1}^{(a)} = \frac{P_{t+1}^{(a)} - P_t^{(a)}}{P_t^{(a)}},$$

and the prediction task is to learn f_θ mapping an information set available at the close of day t to an estimate $\hat{r}_{t+1}^{(a)}$. All features are constructed to be strictly measurable with respect to information available at or before t , so that no look-ahead leakage enters the design matrix.

Modality Partition

The feature space is partitioned into three disjoint modalities:

$$X = X^P \cup X^S \cup X^A, \quad X^P \cap X^S = X^P \cap X^A = X^S \cap X^A = \emptyset,$$

where X^P collects price and technical constructs, X^S the sentiment terms, and X^A the activity terms. Four nested configurations are evaluated:

$$C_1 = X^P, \quad C_2 = X^P \cup X^S, \quad C_3 = X^P \cup X^A, \quad C_4 = X^P \cup X^S \cup X^A.$$

The comparison of C_1 against C_2 , C_3 , and C_4 constitutes the ablation that answers RQ1. Crucially, the *same* four configurations are instantiated for both assets; only the concrete realization of X^A differs, by necessity.

Evaluation Metrics

Given a held-out set T of size n_{test} , we report root-mean-square error

$$RMSE = \sqrt{\frac{1}{n_{test}} \sum_{t \in T} (\hat{r}_{t+1} - r_{t+1})^2}$$

and directional accuracy

$$DA = \frac{1}{n_{test}} \sum_{t \in T} I[\text{sgn}(\hat{r}_{t+1}) = \text{sgn}(r_{t+1})].$$

In other words, RMSE measures how far the predicted return was from the true return on average, penalizing large misses heavily, while directional accuracy ignores magnitude entirely and asks only whether the predicted direction was correct. These metrics are not monotonically related, and the divergence between them is itself informative. A model that shrinks its predictions toward zero reduces (4) on quiet days while conveying no directional information, so improvements in RMSE unaccompanied by improvements in DA should be read as variance reduction rather than as signal extraction. We observe exactly this pattern below.

Because DA is a binomial proportion, we accompany each estimate with a Wilson score interval at the 95% level and an exact two-sided binomial test against the chance rate $\pi_0 = 0.5$. A Wilson score interval gives the plausible range within which the true accuracy lies, given how few test days were available; if that range includes 50%, the model cannot be said to have beaten random guessing. This is a deliberate departure from common practice in the applied fusion literature, where point accuracies are frequently reported without any indication of sampling uncertainty.

DATA AND METHODOLOGY

Assets and Sample

We study one large-capitalization equity, Apple Inc. (AAPL), and one cryptocurrency, Bitcoin (BTC-USD), at daily frequency from 6 July 2024 to 6 July 2026. After feature construction and the removal of rows with incomplete lag windows, 499 usable observations remain for the equity and 730 for the cryptocurrency. The difference reflects the fact that cryptocurrency markets trade continuously while equity markets observe weekends and exchange holidays; we retain the natural calendars rather than downsampling the crypto series, since forcing a common calendar would discard genuine weekend information that is part of what distinguishes the two markets.

Data Sources

All sources are public and require no paid subscription, which supports exact replication.

- **Price and volume.** Daily OHLCV series for both assets from Yahoo Finance via the yfinance interface. OHLCV denotes the open, high, low, and close prices together with the trading volume for each day.
- **On-chain telemetry (Bitcoin only).** Unique active addresses, daily transaction count, and network hash rate from the public blockchain.info endpoints.
- **Equity activity proxies.** Because equities have no settlement-layer analogue, we construct volume-based substitutes from OHLCV data.
- **News text.** Headlines retrieved from Google News RSS for each asset and scored with VADER. The present run captures 100 equity headlines and 100 cryptocurrency headlines; this corpus density is a material limitation and is revisited in the Limitations section.

Feature Construction

Price modality. Simple moving averages over windows $w \in \{10, 20, 50\}$ are computed as

$$MA_t^{(w)} = \frac{1}{w} \sum_{i=0}^{w-1} P_{t-i}.$$

The relative strength index over a 14-day window is

$$RSI_t = 100 - \frac{100}{1 + RS_t}, \quad RS_t = \frac{\bar{U}_t}{\bar{D}_t},$$

where $\bar{U}_t$ and $\bar{D}_t$ are the Wilder-smoothed averages of upward and downward price changes. The moving-average convergence–divergence line is the difference of exponential moving averages,

$$MACD_t = EMA_t^{(12)} - EMA_t^{(26)}.$$

Realized volatility over a ten-day window is the rolling standard deviation of log returns,

$$RV_t^{(10)} = \sqrt{\frac{1}{9} \sum_{i=0}^9 (\ell_{t-i} - \ell_t^-)^2}, \quad \ell_t = \log(P_t/P_{t-1}).$$

RSI and MACD are standard momentum indicators: RSI gauges whether recent gains outweigh recent losses, and MACD compares a fast and a slow moving average to detect shifts in trend.

Sentiment modality. Let H_t denote the set of headlines attributed to day t and $v(h) \in [-1, 1]$ the VADER compound polarity of headline h . Daily sentiment is the equal-weighted mean

$$s_t = \frac{1}{|H_t|} \sum_{h \in H_t} v(h),$$

and sentiment momentum is its first difference,

$$\Delta s_t = s_t - s_{t-1}.$$

Both the level and the first difference are included so that the attribution analysis can distinguish a market that responds to the standing polarity of coverage from one that responds to revisions in it.

Activity modality. For Bitcoin, X^A comprises active addresses A_t , transaction count N_t , and hash rate H_t , together with their log growth rates. For the equity, X^A comprises relative volume

$$RelVol_t = \frac{V_t}{\frac{1}{20} \sum_{i=1}^{20} V_{t-i}}$$

and a Chaikin-style money-flow proxy

$$CMF_t = \frac{\sum_{i=0}^{19} m_{t-i} V_{t-i}}{\sum_{i=0}^{19} V_{t-i}}, \quad m_t = \frac{(C_t - L_t) - (H_t^{hi} - C_t)}{H_t^{hi} - L_t},$$

where C_t , L_t , H_t^{hi} denote the close, low, and high. Both proxies are also computed for Bitcoin, so that the equity and crypto feature sets differ only by the three on-chain series and their growth rates.

We stress that this asymmetry is irreducible rather than a design flaw: no equity possesses a public settlement ledger, so any cross-asset comparison of an activity modality must accept proxy substitution. We therefore interpret RQ1 results for C_3 as a comparison of *the best available activity measure for each asset*, not of identical constructs.

Predictive Architecture

The primary model is a stacked LSTM (Hochreiter and Schmidhuber, 1997). For input x_t and hidden state h_{t-1} , the standard cell recursion is

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, h_t = o_t \odot \tanh(c_t),$$

with $\sigma(\cdot)$ the logistic sigmoid and $\odot$ the elementwise product. The three gates control the flow of information: the forget gate f_t decides how much of the previous memory to discard, the input gate i_t decides how much new information to store, and the output gate o_t decides how much of the memory to expose to the next layer. Two LSTM layers of 64 and 32 units are stacked, each followed by dropout with rate 0.2, and a linear read-out maps the terminal hidden state to $\hat{r}_{t+1}$. Training minimizes mean squared error,

$$\theta^* = \arg \min_{\theta} \frac{1}{|D_{train}|} \sum_{t \in D_{train}} (f_{\theta}(x_{t-L+1:t}) - r_{t+1})^2.$$

The design matrix is split *chronologically*, with the final segment of each series held out for testing, so that the evaluation period strictly postdates the training period and no future information can influence fitted parameters. The held-out evaluation sets comprise 68 observations for AAPL and 206 for BTC-USD; these are the sample sizes on which every reported accuracy and confidence interval is based. Continuous features are standardized using training-set moments only.

Attribution via a Tree Surrogate

Direct attribution of an LSTM's internal gating dynamics is not reliably interpretable, and exact Shapley computation for recurrent models is not standardized. Following common practice, we train a LightGBM

classifier (Ke et al., 2017) on the same feature set—flattened to remove the sequential dimension—to predict the sign of r_{t+1} , and compute exact tree SHAP values (Lundberg and Lee, 2017):

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [g(S \cup \{j\}) - g(S)],$$

where F is the feature set and $g(\cdot)$ the conditional expectation of the model output. Intuitively, this formula asks how much the prediction changes when feature j is added, averaged across every possible combination of the other features. We state explicitly that ϕ_j characterizes the *predictive content of the features under the surrogate*, not the literal internal computation of the LSTM. Because both assets are attributed with the identical surrogate procedure, the cross-asset *comparison* of rankings remains valid even where the absolute attribution is approximate.

Cross-Asset Spillover Testing

To assess RQ3 we estimate, for each direction, a restricted and an unrestricted autoregression. Writing y_t for the recipient market's return and x_t for the transmitting market's sentiment,

$$(restricted) \quad y_t = \alpha + \sum_{i=1}^p \beta_i y_{t-i} + \varepsilon_t, \quad (unrestricted) \quad y_t = \alpha + \sum_{i=1}^p \beta_i y_{t-i} + \sum_{i=1}^p \gamma_i x_{t-i} + u_t,$$

and test $H_0: \gamma_1 = \dots = \gamma_p = 0$ by the F statistic

$$F = \frac{(RSS_r - RSS_u)/p}{RSS_u/(T - 2p - 1)}.$$

The restricted model predicts a market's return from its own past alone; the unrestricted model adds the other market's past sentiment. If the second model fits significantly better, the added series carries predictive information. Tests are run at lags $p \in \{1, 2, 3\}$ in both directions on the calendar-aligned intersection of the two samples, and are complemented by a lagged cross-correlation function for visual inspection of lead-lag structure. Granger causality is a statement about incremental predictive content conditional on own history; it does not establish a structural causal channel, a distinction we maintain throughout.

RESULTS

Sample Description: Prices and News Polarity

Figure 1 plots daily closing prices against daily aggregated sentiment for both assets. Over the terminal portion of the sample AAPL appreciates from approximately \$282 to \$313, while BTC-USD oscillates in a band of roughly \$60,000–\$64,000. Sentiment for both assets fluctuates around a neutral level with

occasional sharp negative excursions—one AAPL headline-day scores approximately -0.43 —consistent with the bursty, event-driven arrival process characteristic of RSS-sourced news. This burstiness is the first indication that the sentiment modality may behave less like a smooth state variable than the fusion literature implicitly assumes.



Figure 1. Daily closing price (line) and daily aggregated news sentiment score (bars) for AAPL (top) and BTC-USD (bottom) over the sample period. Sentiment is the equal-weighted mean VADER compound polarity of headlines attributed to each day, per (10).

On-Chain Activity

Figure 2 shows the three Bitcoin on-chain series over the final ninety days of the sample. Hash rate exhibits a broadly steady-to-rising trend consistent with continued mining capital expenditure, while active addresses and transaction counts fluctuate more sharply with short-horizon network usage. The differing persistence of these series matters for modeling: hash rate is a slow-moving, near-deterministic covariate that contributes little cross-sectional variation at daily frequency, whereas the usage series are noisier but potentially more informative.

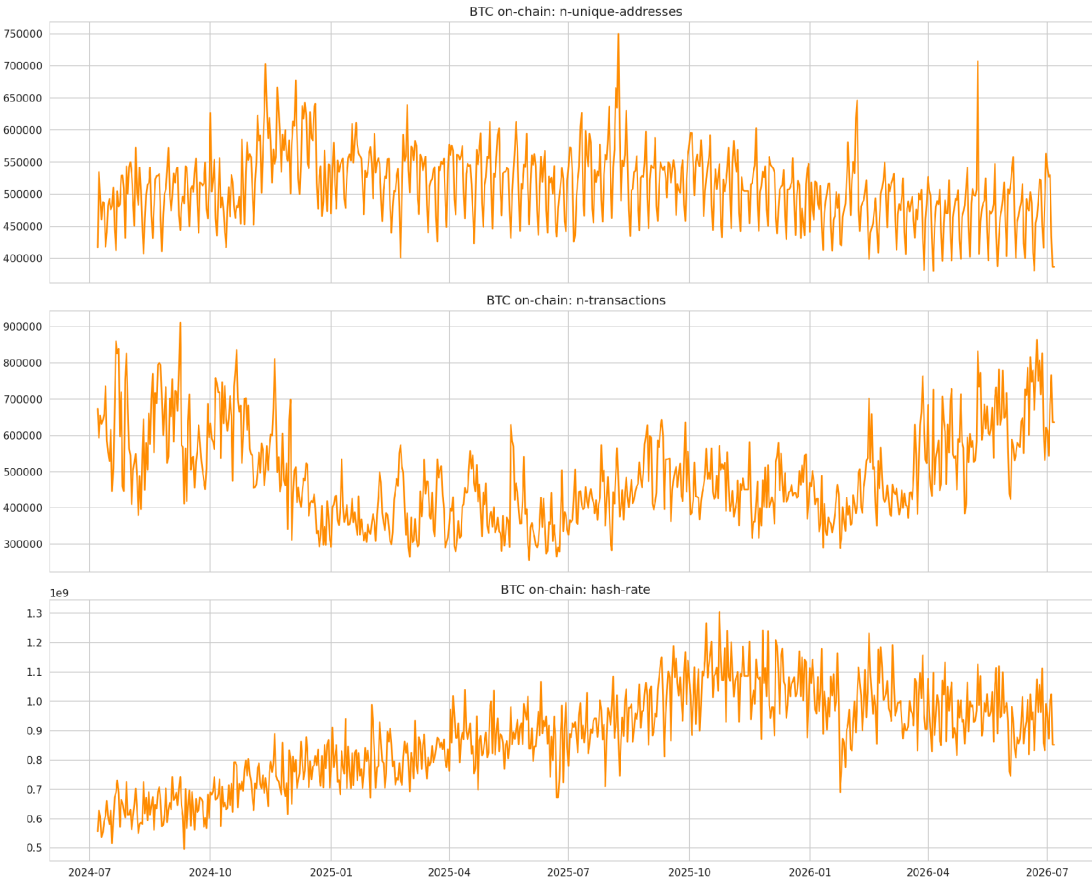


Figure 2. Bitcoin on-chain activity over the final ninety days of the sample: unique active addresses, daily transaction count, and network hash rate.

Correlation Structure

Figure 3 compares feature correlation matrices across the two assets. Three regularities are apparent. First, contemporaneous sentiment is only weakly correlated with contemporaneous returns for both assets, which already suggests that any sentiment signal is unlikely to be linear and contemporaneous. Second, relative volume and ten-day realized volatility are appreciably correlated in both panels, reflecting the well-known volume–volatility relation. Third, for Bitcoin, active addresses and transaction count are strongly mutually correlated—as expected, since both proxy aggregate network usage—which introduces redundancy into the activity modality and foreshadows the dimensionality problem analyzed in the Discussion.

Do Equities and Bitcoin Run on the Same Fuel? A Modality-Controlled Comparison of Price, Sentiment, and Activity Signals for Next-Day Return Prediction

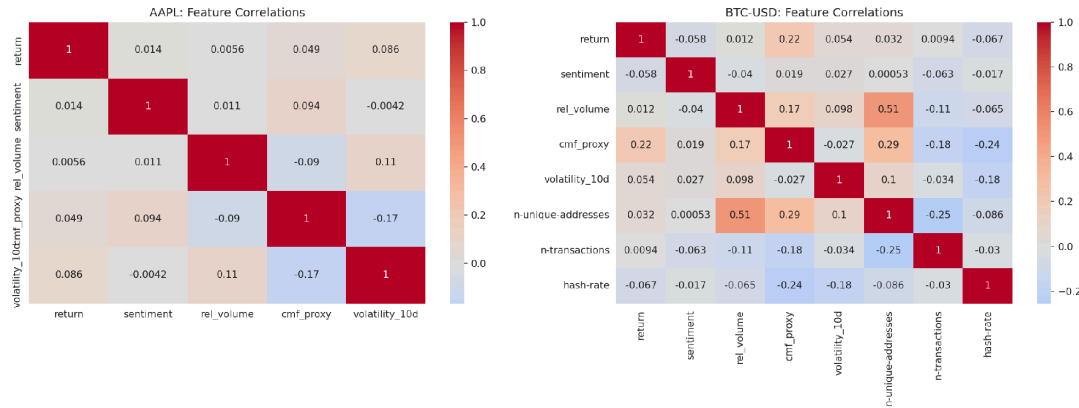


Figure 3. Feature correlation heatmaps for AAPL (left) and BTC-USD (right).

Modality Ablation

Table 1 and Figure 4 report RMSE and directional accuracy for all four configurations and both assets, together with Wilson 95% intervals and exact binomial p -values against chance.

Table 1. Modality ablation results. ΔRMSE is the percentage change relative to the price-only baseline for the same asset (lower is better). Wilson 95% confidence intervals and exact two-sided binomial p -values are computed against the chance rate $\pi_0 = 0.5$. Test-set sizes are $n_{\text{test}} = 68$ (AAPL) and $n_{\text{test}} = 206$ (BTC-USD).

Asset	Configuration	RMSE	ΔRMSE	Dir. Acc.	95% CI
AAPL	C_1 Price-only	0.02280	—	0.529	[0.412, 0.643] [0.716]
AAPL	C_2 + Sentiment	0.02125	− 6.8%	0.471	[0.357, 0.588] [0.716]
AAPL	C_3 + Activity	0.02911	+ 27.7%	0.515	[0.398, 0.629] [0.904]
AAPL	C_4 Full fusion	0.03399	+ 49.1%	0.426	[0.316, 0.545] [0.275]
BTC-USD	C_1 Price-only	0.02081	—	0.505	[0.437, 0.572] [0.944]
BTC-USD	C_2 + Sentiment	0.02849	+ 36.9%	0.476	[0.409, 0.544] [0.531]
BTC-USD	C_3 + Activity	0.02425	+ 16.5%	0.505	[0.437, 0.572] [0.944]
BTC-USD	C_4 Full fusion	0.03474	+ 66.9%	0.456	[0.390, 0.525] [0.236]

Three observations follow. First, the price-only baseline attains the lowest RMSE and the highest directional accuracy for both assets, and full fusion attains the highest RMSE and the lowest directional accuracy for both. The ordering is consistent across two assets with different sample lengths, different

microstructures, and different realizations of the activity modality, which makes a purely coincidental explanation less plausible than it would be in a single-asset study.

Second, RMSE and directional accuracy diverge in an instructive way. For AAPL, adding sentiment *lowers* RMSE from 0.02280 to 0.02125 while simultaneously lowering directional accuracy from 0.529 to 0.471. This is precisely the shrinkage pathology anticipated above: the augmented model learns to predict values closer to zero, which is rewarded by a squared-error criterion on quiet days but is uninformative—or actively misleading—about the sign of the move. Papers that report only RMSE would record this configuration as an improvement.

Third, and most importantly, *no configuration is statistically distinguishable from chance*. Every Wilson interval in Table 1 contains 0.5, and every binomial p -value exceeds 0.23. A two-proportion test comparing price-only against full fusion yields $z = 1.20$ ($p = 0.229$) for AAPL and $z = 0.99$ ($p = 0.324$) for BTC-USD. We therefore do not claim that fusion causes a statistically significant degradation; we claim the weaker and better-supported proposition that *the data provide no evidence that fusion helps*, with a consistently negative point-estimate direction across both assets. Given the prevailing publication norm of reporting fusion gains of similar or smaller magnitude without any uncertainty quantification, we regard this as the more scientifically defensible framing.

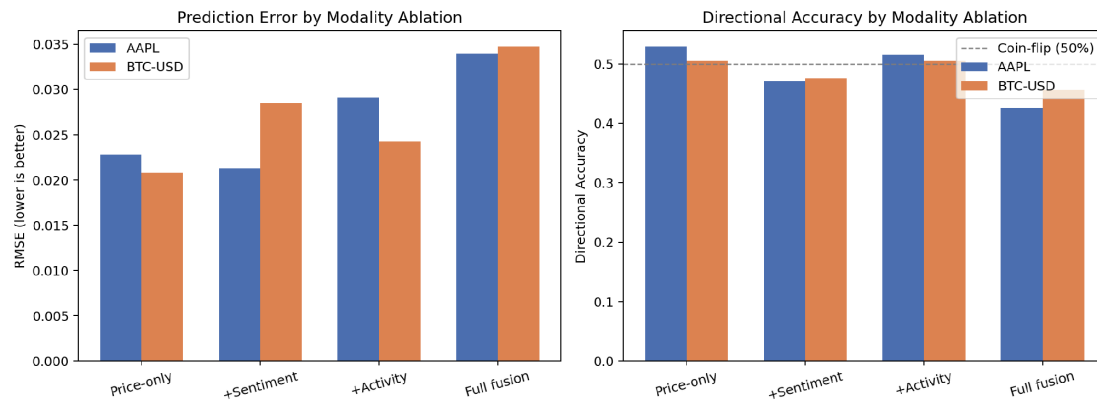


Figure 4. Prediction error (left) and directional accuracy (right) across modality ablations for AAPL and BTC-USD. The dashed line marks the 50% chance level. Note that the two panels do not rank the configurations identically, reflecting the metric divergence discussed above.

Feature Attribution

Figures 5 and 6 present SHAP summary plots for the LightGBM surrogate trained on the full-fusion feature set.

For AAPL (nine features displayed), the ranking in descending order of mean absolute attribution is MACD, the money-flow proxy, relative volume, RSI, ten-day realized volatility, MA10, MA20, and

finally the two sentiment terms. Momentum and money-flow constructs therefore dominate, and both sentiment features rank last.

For BTC-USD (twelve features displayed), the ranking is hash rate, transaction count, unique active addresses, MA10, realized volatility, RSI, MACD, the money-flow proxy, relative volume, MA20, and—again last—contemporaneous sentiment s_t and sentiment momentum Δs_t . The three highest-ranked features are thus *all* on-chain activity measures, while the technical block occupies the middle of the ranking.

Two observations follow, and the second is the paper's most consequential attribution result.

First, the assets are genuinely heterogeneous, but along an axis different from the one the popular narrative proposes. The equity is anchored to price-derived momentum and money flow; Bitcoin is anchored to settlement-layer activity. The distinguishing modality for Bitcoin is therefore not attention but *observable network usage*—a covariate class that has no equity counterpart whatsoever, which is precisely why the two markets cannot be said to run on the same fuel.

Second, and contrary to the framing common in the cryptocurrency-forecasting literature, *sentiment ranks last for both assets*. In Figure 6 the attribution distributions of s_t and Δs_t collapse to near-vertical lines at zero, indicating that the surrogate assigns them almost no predictive weight; the same collapse is visible for AAPL in Figure 5. This attribution-side finding is mutually corroborating with the ablation-side finding above: a modality that receives negligible attribution when included is exactly the modality whose inclusion should be expected to add dimensionality without adding signal, and therefore to degrade held-out performance. The two independent lines of evidence agree, which materially strengthens the fusion conclusion relative to either result alone. We caution, however, that this is a statement about *this* sentiment corpus, not about news sentiment as a construct.

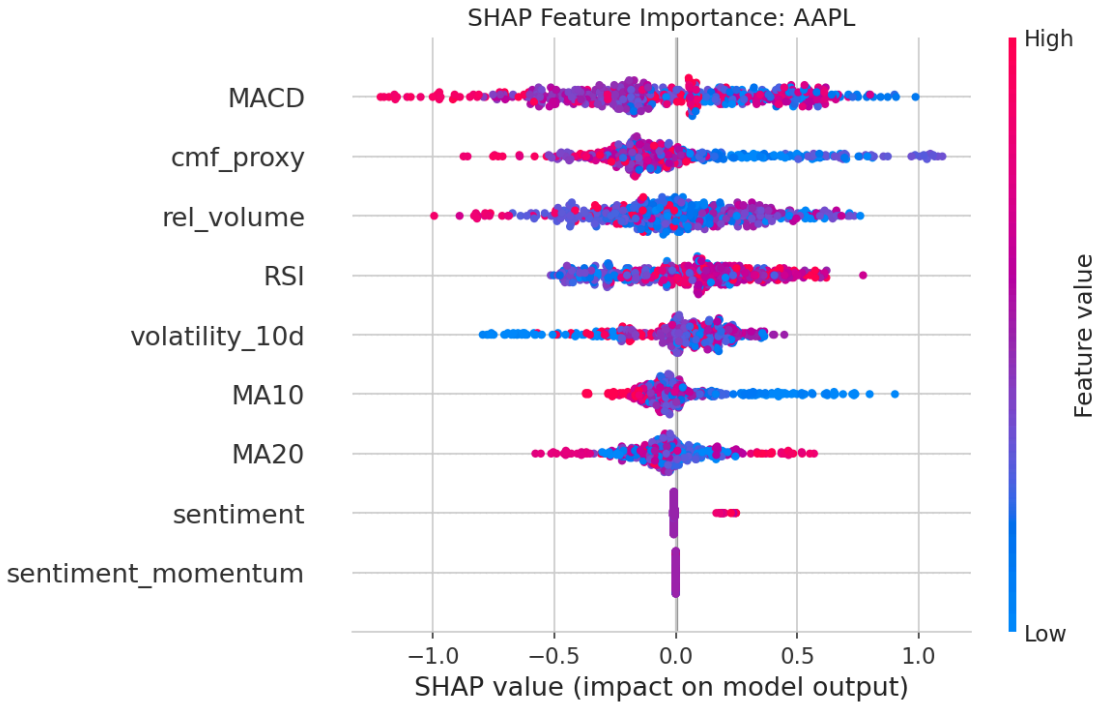


Figure 5. SHAP feature attribution for the AAPL full-fusion surrogate, ordered by mean absolute attribution. Momentum and money-flow constructs dominate; both sentiment features rank last with attribution collapsed near zero.

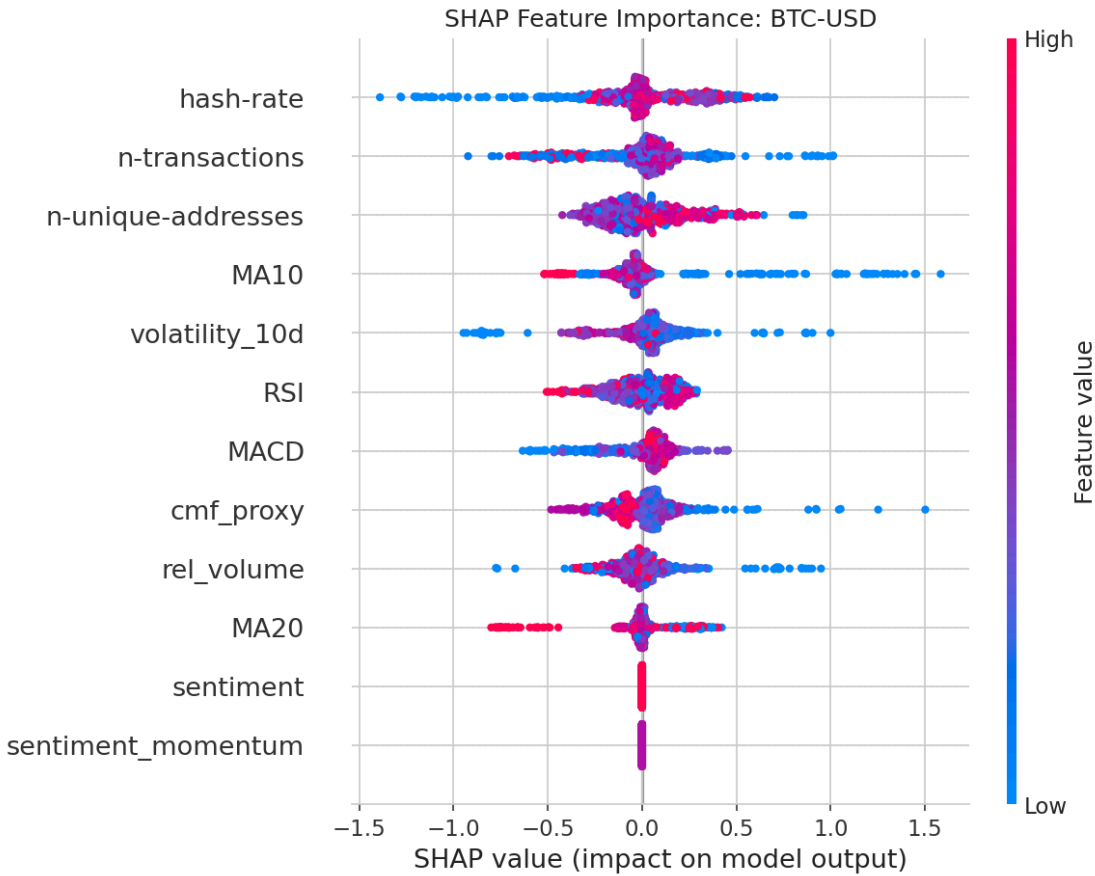


Figure 6. SHAP feature attribution for the BTC-USD full-fusion surrogate, ordered by mean absolute attribution. The three highest-ranked features—hash rate, transaction count, and unique active addresses—are all on-chain activity measures; both sentiment features rank last.

Cross-Asset Spillover

Table 2 reports Granger causality F -test p -values in both directions at lags one through three, and Figure 7 plots the lagged cross-correlation between Bitcoin sentiment and AAPL returns.

Table 2. Granger causality test results (F -test p -values) for (19). Bold entries are significant at the 5% level.

Direction	Lag 1	Lag 2	Lag 3
BTC sentiment → AAPL return	0.018	0.034	n/a [†]
AAPL sentiment → BTC return	0.242	0.324	0.515

[†] The lag-3 test in this direction could not be computed owing to constant-column degeneracy in the available sample.

Bitcoin sentiment carries statistically significant incremental predictive content for next-day AAPL returns at lags one and two, conditional on the equity's own return history. The reverse direction is insignificant at every tested lag, with p -values exceeding 0.24 throughout. The relationship is therefore *asymmetric*: within this sample, the crypto market's news polarity leads the equity market's returns, but not conversely.

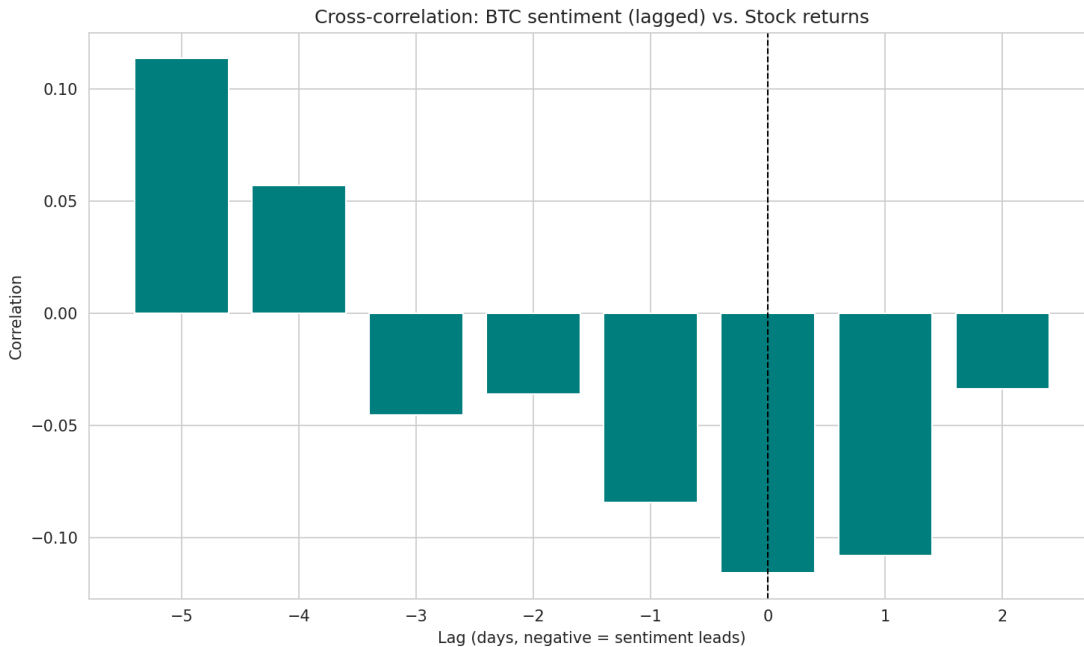


Figure 7. Lagged cross-correlation between BTC-USD sentiment and AAPL returns, lags -5 to $+5$ days. Negative lags indicate that sentiment leads returns.

Two interpretive cautions apply. Granger causality establishes incremental predictive content conditional on own history, not a structural causal mechanism; the observed relationship is equally consistent with both series responding to a common macro or news-cycle factor with differential absorption speeds. Moreover, the effect magnitudes implied by Figure 7 are modest. The finding is therefore economically suggestive and directionally consistent with the connectedness literature, but it is not, on its own, a tradable signal.

DISCUSSION

Why Additive Fusion Did Not Help

The fusion result runs counter to the framing of much recent applied work. We identify three non-exclusive mechanisms, ordered by how strongly the present evidence supports them.

Corpus density and signal-to-noise ratio. The sentiment series is constructed from 100 headlines per asset, retrieved from an RSS feed that returns only recent items rather than a dense historical archive. Under sparse coverage, the daily aggregate in (10) rests on very few documents. The measurement error in s_t is therefore large, and it varies unevenly from day to day. On days with no retrieved headline, the series is effectively imputed rather than observed. A feature whose measurement error dominates its informational content will, in expectation, degrade a model that must expend capacity fitting it. The attribution results support this mechanism directly. Both sentiment terms collapse to near-zero attribution for both assets, as Figures 5 and 6 show. That is the signature of a channel carrying no usable information, rather than of one carrying information the model fails to exploit. We regard this as the best-supported of the three mechanisms, and it is the one most directly addressable by future work.

Parameter-sample ratio. Moving from C_1 to C_4 expands the input dimension substantially while the training sample remains fixed at roughly 420 and 620 observations for the equity and cryptocurrency respectively. For a two-layer LSTM with 64 and 32 units, this ratio is unfavourable. The correlation structure in Figure 3 compounds the problem: active addresses and transaction count are near-collinear, so the extra dimensions supply redundant rather than independent information. Dropout at rate 0.2 mitigates but does not eliminate the resulting overfitting, which manifests exactly as observed: acceptable in-sample loss with degraded held-out directional accuracy.

Objective-metric mismatch. As formalized earlier and observed empirically for AAPL under C_2 , minimizing squared error does not align with maximizing directional accuracy. Additional noisy features give the optimizer more ways to reduce squared error by shrinking toward the unconditional mean, which is directionally uninformative. A model trained under a classification or sign-aware objective might exhibit a different—though not necessarily better—fusion profile.

Asset Heterogeneity in Attribution

The attribution results provide the paper's clearest affirmative answer to RQ2, and they answer it against the prevailing narrative. Equity predictions are anchored to price-derived momentum and money-flow structure. Bitcoin predictions are anchored to settlement-layer activity: hash rate, transaction throughput, and unique active addresses occupy the top three positions, ahead of every technical indicator. The two markets therefore do not run on the same fuel—but the difference is that Bitcoin possesses an *additional observable modality*, not that it is more responsive to narrative.

This finding is easy to rationalize ex post, and it was not the hypothesis we set out to confirm. On-chain telemetry measures realized economic usage of the network directly, with no estimation layer between the observable and the construct. News polarity, by contrast, is a noisy proxy for an unobservable attention state, mediated by editorial selection, corpus coverage, and a lexicon calibrated on a different linguistic register. Where a clean measurement of the underlying construct is available, a tree ensemble will prefer it. The equity has no such clean activity measurement—the money-flow and relative-volume proxies are themselves price-and-volume transforms—which is consistent with the equity ranking being dominated by price-derived features by default.

It is worth being explicit that this result inverts a claim frequently made in the applied cryptocurrency literature and repeated informally as market folklore. At the daily horizon and on this corpus, the hype channel is the *weakest* of the three modalities for Bitcoin, not the strongest. We do not read this as refuting attention-based accounts of cryptocurrency pricing in general. We read it as showing two things: that daily aggregated headline polarity is an inadequate instrument for the attention construct, and that on-chain data offers a substantially better-measured alternative.

The practical implication for model builders is concrete. Sentiment features should not be included by default on the assumption that they help. Where an asset exposes a directly measured activity channel, that channel should be prioritized over a proxied attention channel.

Interpreting the Asymmetric Spillover

The spillover result sits in instructive tension with the attribution result. Sentiment receives negligible attribution *within* each asset, yet Bitcoin sentiment carries significant incremental predictive content for equity returns *across* assets. Three readings are available, and we cannot discriminate among them with the present design.

The first is that the two findings operate at different margins. Within-asset attribution asks whether s_t improves prediction of the same asset's return, given a rich set of own-price features. The Granger test asks something weaker: whether s_t improves prediction of a *different* asset's return, given only that asset's own return history. Sentiment can be incrementally informative in the second setting while being redundant in the first.

The second is substantive: cryptocurrency markets may function as a high-frequency barometer of aggregate risk appetite, absorbing and expressing attention shocks before they propagate into more heavily intermediated traditional assets. Bitcoin also trades continuously, so it has the mechanical opportunity to price weekend and overnight news before equity markets reopen. That alone would generate a lead relationship, with no behavioural channel required.

The third is that the spillover result is a small-sample artifact. Two tests fall in the significant direction, at $p = 0.018$ and $p = 0.034$. Six tests were conducted in total across directions and lags, so a multiplicity adjustment would render neither significant at the 5% level; the Bonferroni threshold would be 0.0083. We report the uncorrected values for comparability with the connectedness literature, but we regard the multiplicity caveat as material and decline to treat this result as established. Distinguishing among the three readings would require intraday alignment, a denser sentiment corpus, and a cross-section of assets rather than a pair.

LIMITATIONS

We state limitations explicitly, since several of them materially bound the strength of the conclusions.

1. **Sentiment corpus density.** Google News RSS returns only recent headlines; the present run captures 100 per asset. A dense historical archive would likely change the fusion result and should be the first robustness check before any strong claim about sentiment's marginal value is drawn. The negative fusion finding is properly read as *conditional on this corpus*, not as a general property of sentiment features.
2. **Statistical power.** With $n_{test} = 68$ and 206, the study is underpowered to detect the effect sizes typical of daily return prediction. Every directional accuracy reported here is consistent with chance, and the consistency of the *ordering* across assets, rather than any individual point estimate, is what carries evidential weight.
3. **Two assets, one regime.** AAPL and BTC-USD over a single two-year window constitute a proof of concept for the design, not a characterization of equity or cryptocurrency markets in general.
4. **Surrogate attribution.** SHAP values are computed on a LightGBM classifier fitted to flattened features rather than on the LSTM itself. This is a disclosed and conventional approximation; it supports cross-asset comparison of feature rankings but not literal claims about recurrent internal dynamics.
5. **Single split, no walk-forward validation.** Results derive from a single chronological split. A rolling-origin scheme with multiple evaluation folds would yield distributions rather than point estimates and is the natural next step.
6. **Granger causality is not structural causality.** The spillover result may reflect common exposure to a shared news cycle, or the continuous-trading calendar mechanism discussed above, rather than a behavioural transmission channel. Six tests were conducted, and neither significant p -value survives a Bonferroni threshold of 0.0083.
7. **Sentiment features are near-degenerate in this sample.** The collapsed SHAP distributions indicate that the sentiment series contributes almost no variation the surrogate can exploit. We can conclude that *this* sentiment measurement is uninformative, not that news sentiment is uninformative.
8. **No transaction-cost or portfolio evaluation.** We report statistical accuracy only. Directional accuracies in the observed range would not survive realistic transaction costs, and no economic-significance claim is made.

COCLUSION

We applied a single modality-controlled pipeline to an equity and to Bitcoin in order to test whether the two asset classes are driven by the same informational inputs. Three conclusions follow. First, additive fusion did not improve out-of-sample prediction for either asset; the price-only baseline was the strongest

configuration on both metrics for both assets, and full fusion the weakest, though no configuration is statistically separable from chance on the available test sets. Second, attribution is strongly asset-heterogeneous: momentum and money-flow constructs dominate for the equity, whereas Bitcoin's three highest-ranked features are all on-chain activity measures—and sentiment ranks last for *both* assets, an attribution-side result that independently corroborates the ablation-side failure of fusion. Third, Bitcoin news polarity Granger-causes next-day equity returns at short lags while the reverse relationship is absent, although this result does not survive a multiplicity correction and should be treated as provisional.

The broader methodological message is that multimodal fusion should be treated as a hypothesis requiring per-asset, per-corpus validation rather than as an assumed improvement, and that fusion studies should report uncertainty intervals alongside point accuracies as a matter of course. The substantive message is that equities and Bitcoin do not run on the same fuel—but not for the reason usually given. The distinguishing feature of Bitcoin in this study is not that it is more responsive to narrative; it is that its settlement layer exposes a directly measured activity modality with no equity analogue, and that this modality outranks both price and sentiment features in attribution. At the daily horizon, on this corpus, the hype channel was the weakest of the three modalities for both assets.

Future work should prioritize, in order: a dense historical sentiment corpus, which is the single change most likely to alter the fusion and attribution conclusions; walk-forward validation across multiple folds to convert point estimates into distributions with adequate power; expansion to a cross-section of equities and cryptocurrencies to test whether on-chain dominance generalizes beyond Bitcoin; and intraday alignment to separate the behavioural interpretation of the spillover result from the continuous-trading calendar mechanism.

DATA AND CODE AVAILABILITY

All data sources used in this study are publicly accessible without subscription: daily OHLCV data from Yahoo Finance, on-chain telemetry from the public blockchain.info endpoints, and news headlines from Google News RSS. Sentiment scoring uses the open-source VADER lexicon. The author declares no competing financial or personal interests.

REFERENCES

- Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science*, vol. 2, no. 1, 2011, pp. 1–8, <https://doi.org/10.1016/j.jocs.2010.12.007>.
- Corbet, Shaen, et al. "Exploring the Dynamic Relationships between Cryptocurrencies and Other Financial Assets." *Economics Letters*, vol. 165, 2018, pp. 28–34, <https://doi.org/10.1016/j.econlet.2018.01.004>.

- Diebold, Francis X., and Kamil Yilmaz. "Better to Give than to Receive: Predictive Directional Measurement of Volatility Spillovers." *International Journal of Forecasting*, vol. 28, no. 1, 2012, pp. 57–66, <https://doi.org/10.1016/j.ijforecast.2011.02.006>.
- Fischer, Thomas, and Christopher Krauss. "Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions." *European Journal of Operational Research*, vol. 270, no. 2, 2018, pp. 654–669, <https://doi.org/10.1016/j.ejor.2017.11.054>.
- Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica*, vol. 37, no. 3, 1969, pp. 424–438, <https://doi.org/10.2307/1912791>.
- Gu, Shihao, Bryan Kelly, and Dacheng Xiu. "Empirical Asset Pricing via Machine Learning." *The Review of Financial Studies*, vol. 33, no. 5, 2020, pp. 2223–2273, <https://doi.org/10.1093/rfs/hhaa009>.
- Hochreiter, Sepp, and Jürgen Schmidhuber. "Long Short-Term Memory." *Neural Computation*, vol. 9, no. 8, 1997, pp. 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Hutto, C. J., and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014, pp. 216–225.
- Ke, Guolin, et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems 30*, 2017, pp. 3146–3154.
- López de Prado, Marcos. *Advances in Financial Machine Learning*. John Wiley & Sons, 2018.
- Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.
- Tetlock, Paul C. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance*, vol. 62, no. 3, 2007, pp. 1139–1168, <https://doi.org/10.1111/j.1540-6261.2007.01232.x>.
- Wilson, Edwin B. "Probable Inference, the Law of Succession, and Statistical Inference." *Journal of the American Statistical Association*, vol. 22, no. 158, 1927, pp. 209–212, <https://doi.org/10.1080/01621459.1927.10502953>.