Journal of Finance and Business Digital (JFBD)
Vol. 5, No. 1 2026: 33-52

ML-CryptoSI: A Multilingual Crypto Sentiment Index and its Role in Bitcoin and Ethereum Pricing

Ningyu Zhou
Al-Farabi Kazakh National university, Kazakhstan
Corresponding Author: Ningyu Zhou chzhou_ninyuy@live.kaznu.kz

A R T I C L E I N F O

Keywords: Cryptocurrency, Sentiment Analysis, Multilingual Text, Bitcoin, Ethereum

Received : 20, November
Revised : 22, January
Accepted: 24, March

A B S T R A C T

Cryptocurrency prices often move with narratives and investor sentiment. This paper builds a multilingual crypto sentiment index, ML-CryptoSI, using daily news text in six languages and Binance market data for BTC and ETH. We first aggregate language-level daily sentiment and then use PCA to extract the common component across languages. Next, we test whether ML-CryptoSI predicts next-day returns and volatility proxies after controlling for lagged market conditions, liquidity, and day-of-week fixed effects. The results show that ML-CryptoSI has incremental information for returns, especially for ETH, and the effect is stronger on high news-intensity days. In contrast, the evidence for volatility prediction is weak in this short sample. Overall, the findings suggest that the common factor in multilingual news sentiment matters for short-run crypto pricing and is state dependent.

INTRODUCTION

Over the past decade, the cryptocurrency market has seen repeated booms and crashes. Bitcoin (BTC) and Ethereum (ETH) are now major global assets, yet unlike traditional assets they lack stable cash flows and clear valuation anchors. As a result, prices can react strongly to expectations, narratives, and investor sentiment, and prior studies link sentiment and attention to crypto returns, volatility, and jump risk (Nakhli et al., 2022; Bouri et al., 2022; Koutmos, 2023). This paper asks one question: Does multilingual news sentiment add information for BTC and ETH pricing? Existing evidence shows that sentiment measures can predict returns and volatility in some settings and that effects can be stronger in extreme market states (Nakhli et al., 2022; Bouri et al., 2022). Related work reports strong links between sentiment and crypto prices (Farrugia & Deguara, 2025), short-run predictability from tweet sentiment (Lupu & Donoiu, 2025), and asymmetric effects of high-impact narratives on jumps and tail risks (Aysan et al., 2024). However, much of the evidence remains English-only or channel-specific, despite the global and cross-lingual nature of crypto information.

Recent NLP advances make it feasible to measure sentiment at scale and integrate it into forecasting and pricing tests. Prior research shows that adding sentiment features can improve forecasting performance and has examined deep learning architectures and multilingual text pipelines (Seabe et al., 2023, 2025; Hamayel et al., 2021; Yamak et al., 2019; Ponselvakumar et al., 2024; Natzir & Jatiprasetya, 2025; Amberkhani et al., 2025; Girsang & Stanley, 2022, 2023; Gurgul et al., 2023, 2024; García-Medina et al., 2025; Zhao et al., 2023). Newer lines also highlight multi-source signals, including on-chain text, while noting potential noise (Kleitsikas et al., 2025). From an asset-pricing perspective, composite sentiment factors such as "CryptoSent" suggest sentiment risk may be priced (John et al., 2024; Filippou et al., 2025), and recent methods propose improved labeling and tuning for noisy and multilingual corpora (Moradi-Kamali et al., 2025; Tiwari et al., 2025).

Despite this progress, three gaps remain. First, cross-lingual sentiment is still under-studied. Second, news, social media, and on-chain text are often analyzed separately, and unified frameworks remain rare (Kleitsikas et al., 2025; Gurgul et al., 2023). Third, systematic evidence on benchmark assets is limited, especially for BTC and ETH across returns, volatility, and liquidity (John et al., 2024; Filippou et al., 2025).

We address these gaps by building ML-CryptoSI (Multilingual Crypto Sentiment Index) and testing its pricing relevance for BTC and ETH. We collect multilingual news from professional media RSS feeds in 2025, including CoinDesk, Cointelegraph, Odaily, BlockBeats, TokenPost, ForkLog, and PANews (CoinDesk; Cointelegraph; Odaily; BlockBeats; ForkLog; PANews; TokenPost). We also incorporate extractable on-chain text from the Bitcoin and Ethereum blockchains (Kleitsikas et al., 2025). Market and liquidity data come mainly from the Binance public API and are cross-checked using CoinGecko's REST API (Binance; CoinGecko). We score text using multilingual pretrained language models and financial sentiment lexicons and use PCA to extract a cross-lingual common sentiment factor.

The results show that the multilingual common sentiment factor contains incremental information for next-day pricing, with clearer return effects for ETH in baseline tests, while volatility evidence is weaker in this short sample. The effect is also stronger when the news flow is heavier, consistent with state dependence. This paper contributes in three ways. First, it provides a reproducible pipeline for multilingual sentiment research in crypto markets. Second, it uses PCA to extract cross-lingual common sentiment factors and builds ML-CryptoSI. Third, it tests ML-CryptoSI for BTC/ETH returns and risk proxy variables and studies language heterogeneity and state dependence of information intensity.

LITERATURE REVIEW

Behavioral finance perspectives suggest that cryptocurrencies have weak fundamental anchors, making prices more sensitive to sentiment and attention. Empirical evidence using search indexes, survey sentiment, and text-based measures generally links sentiment to crypto returns, volatility, and tail risk (Dias et al., 2022; Mokni, 2022; Koutmos, 2023). Text-based studies further document strong associations between sentiment and crypto pricing, including large-scale sentiment measures and short-horizon signals from social platforms, and show that media tone and high-impact narratives can shape jumps and tail risk (Farrugia & Deguara, 2025; Lupu & Donoiu, 2025; Aysan et al., 2024).

From an asset-pricing view, sentiment may also act as a systematic factor rather than only a short-run correlate. Composite sentiment indexes such as "CryptoSent" are introduced into pricing models, showing heterogeneity in sentiment exposure across coins and links to expected returns (John et al., 2024; Filippou et al., 2025).

Methodologically, the literature increasingly combines sentiment with deep learning and pretrained language models. Forecasting studies often report that sentiment features improve performance and that deep architectures outperform linear benchmarks (Seabe et al., 2023, 2025; Shahid et al., 2021; Yamak et al., 2019; Ponselvakumar et al., 2024; Natzir & Jatiprasetya, 2025; Amberkhani et al., 2025). For sentiment extraction, FinBERT, RoBERTa, and XLM-R with financial lexicons are widely used, including hybrid designs that integrate embeddings with LSTM/GRU models (Girsang & Stanley, 2022, 2023). Multi-source frameworks also combine news, social media, and on-chain data and report gains from multimodal fusion, while reviews synthesize evidence that text-based sentiment features can improve prediction and trading strategies (Gurgul et al., 2023, 2024; García-Medina et al., 2025; Zhao et al., 2023). Related work on on-chain text constructs sentiment indicators and tests predictability for returns and volatility but emphasizes that on-chain text can be noisy and that simpler dictionary-based methods may be robust in practice (Kleitsikas et al., 2025).

Even with these advances, most studies remain English-centric or treat multilingual data through translation or pretraining without systematically comparing language communities. Composite indexes such as "CryptoSent" are also largely built from English news and social media, leaving the multilingual dimension under-incorporated in factor construction and tests on benchmark assets such as BTC and ETH (John et al., 2024; Filippou et al., 2025). Moreover, evidence on how multilingual sentiment relates jointly to BTC/ETH returns, volatility, and liquidity remains limited, especially beyond forecasting or event-specific settings (Kleitsikas et al., 2025; Gurgul et al., 2023). This paper builds a multilingual, multi-source sentiment index, ML-CryptoSI, and applies it to BTC and ETH. It connects the sentiment–return/volatility literature with cross-sectional factor ideas by extracting a common sentiment component from multilingual series and testing whether it adds information for pricing and risk in leading cryptocurrencies.

METHODOLOGY

To build a multilingual crypto market sentiment index (ML-CryptoSI) and test its explanatory power for Bitcoin (BTC) and Ethereum (ETH) returns and volatility, this paper combines daily-aligned market data with multilingual text data. All datasets are aligned to a single daily index using UTC-normalized timestamps. We apply standard field cleaning, deduplication, language-identification confidence checks (lang_prob), and a complete-day filter to reduce noise from incomplete scraping and missingness.

Market data come from Binance REST endpoints for BTCUSDT and ETHUSDT, providing OHLCV-style market information. Technical indicators from TAAPI.io (e.g., EMA, RSI, MACD) are used as controls. Inspired by the multi-agent large language model trading framework TradingAgents (Xiao et al., 2025), we also include microstructure/positioning proxies (e.g., order-book features, long–short ratios) as additional controls. All market timestamps are converted to UTC; if raw frequency is higher than daily, data are aggregated to daily measures. Returns are computed as r_t=\ln(P_t/P_{t-1}). Volatility is proxied by rolling-window standard deviations (and related realized measures), and liquidity/activity is captured by volume and related price–quantity measures.

News and media text are collected mainly from RSS feeds of professional crypto media over 2025-08-01 to 2025-11-21. We include CoinDesk, Cointelegraph, Odaily, BlockBeats, TokenPost, ForkLog, and PANews, and collect standard fields (title/summary, time, URL) while retaining source labels for checks and robustness (CoinDesk, n.d.; Cointelegraph, n.d.; Odaily, n.d.; BlockBeats, n.d.; TokenPost, n.d.; ForkLog, n.d.; PANews, n.d.). Scraping follows a rolling-update strategy to capture new items. Because duplication is common across channels, we apply deduplication using URLs and the (title, published time) pair. We collect 22,636 raw news items and retain 18,794 items after filtering to six target languages, removing 3,842 items that are in other languages or cannot be classified.

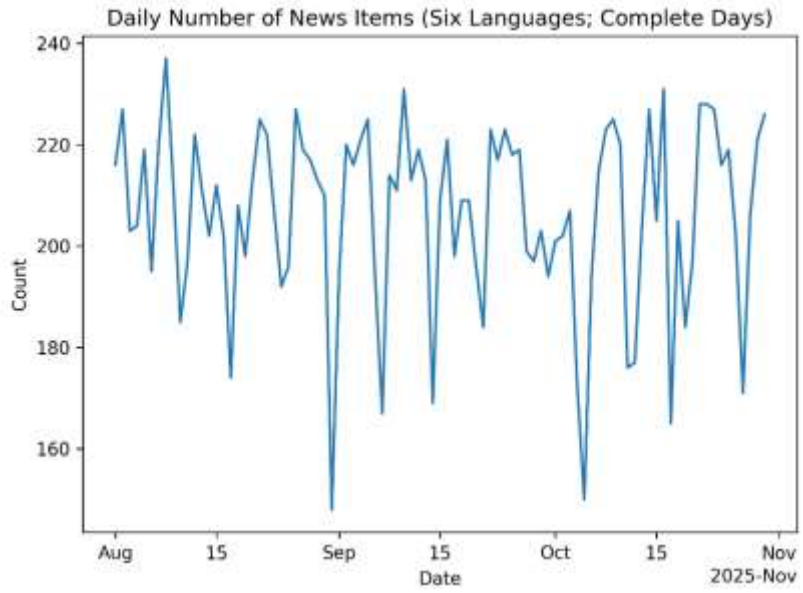

Figure 1. Daily News Counts in Six Languages (Complete-Day Sample)

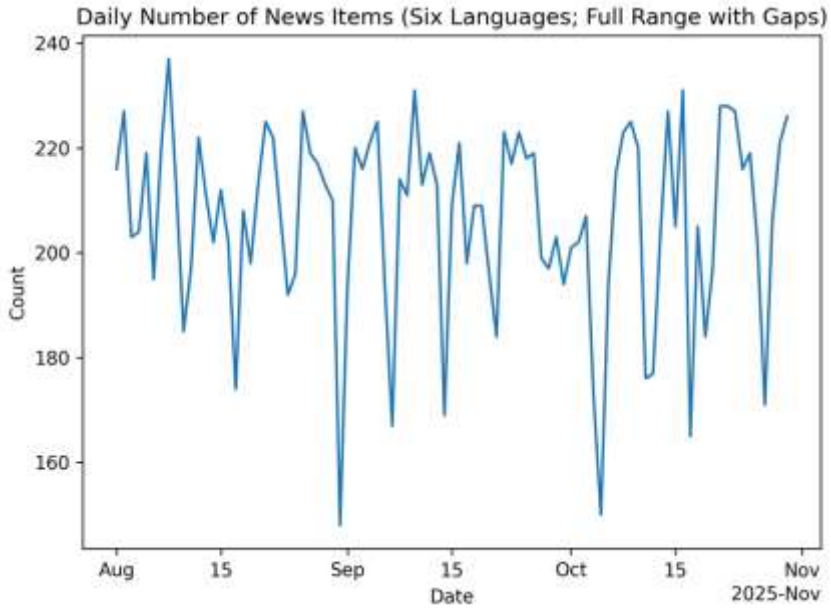

Figure 2. Daily News Counts in Six Languages (Full Window, Including Gaps)

Before sentiment inference, we remove HTML/control characters, normalize encoding to UTF-8, standardize punctuation/whitespace, and drop extremely short records. We then run automatic language identification to obtain a language label and confidence score lang_prob. The six target languages are English (en), Chinese (zh), Korean (ko), Japanese (ja), Spanish (es), and Russian (ru), with aliases such as zh-cn/zh-tw merged into zh. Coverage is highly imbalanced (en=17,473, ko=945, ru=190, zh=143, es=34, ja=9), so low-volume languages are treated as exploratory and robustness checks consider alternative aggregation/weighting rules.

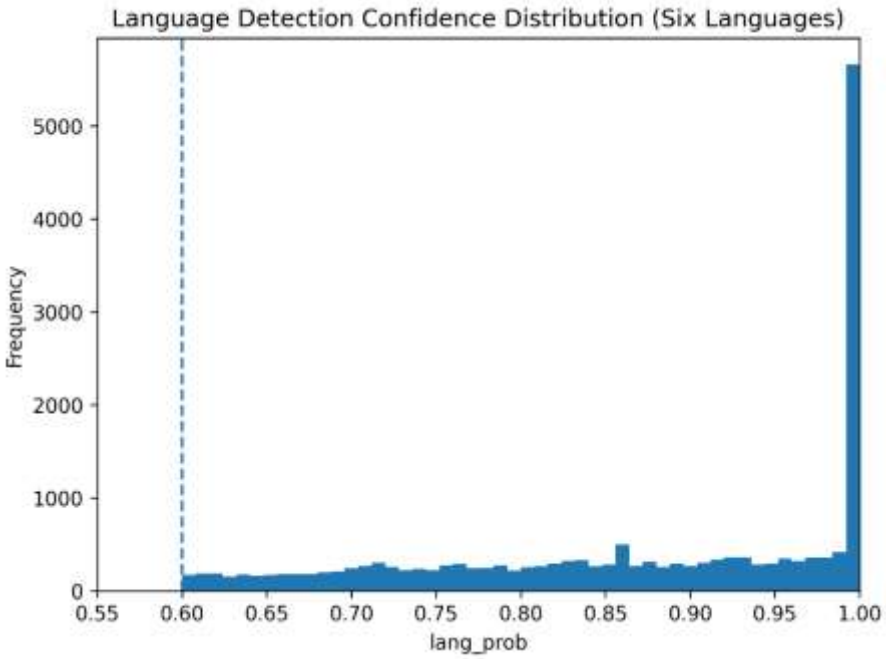

Figure 3. Distribution of Language-Identification Confidence (Six Languages)

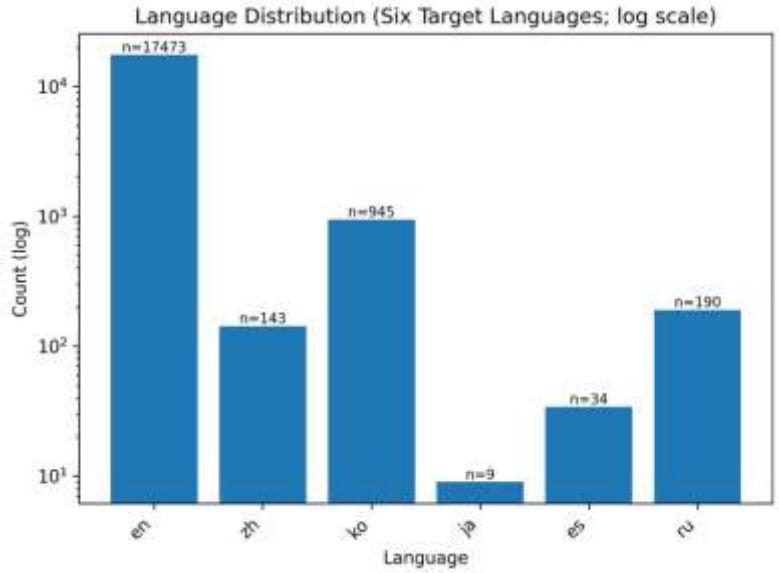

Figure 4. Sample Size By Language (Log Scale)

Besides news text, we build a social-media collection/cleaning module as a future extension; the baseline analysis focuses on professional news to avoid platform-policy and data-availability instability.

For each news item, we apply a multilingual sentiment model and output $p_{\text{pos}}, p_{\text{neg}}, p_{\text{neu}}$. We define a continuous sentiment score $\text{SentIndex}_i = p_{\text{pos}}^{(i)} - p_{\text{neg}}^{(i)}$, which is suitable for daily aggregation when most items are neutral. For descriptive reporting, we also assign discrete labels by the largest probability; in the six-language sample, label shares are neutral 83.6%, negative 10.2%, and positive 6.2. At the daily level, sentiment is aggregated as $\text{Sent}_t = \frac{1}{N_t}\sum_{i \in t} \text{SentIndex}_i$, where N_t is the number of items across the six languages; we also compute a 7-day rolling mean as a smoothed series.

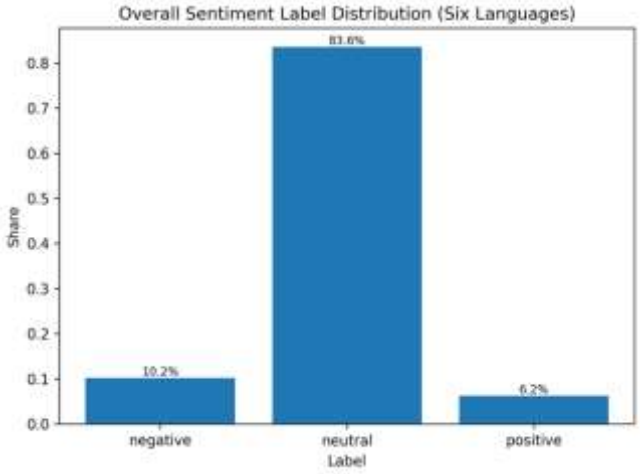

Figure 5. Overall Sentiment Label Distribution (Six Languages)

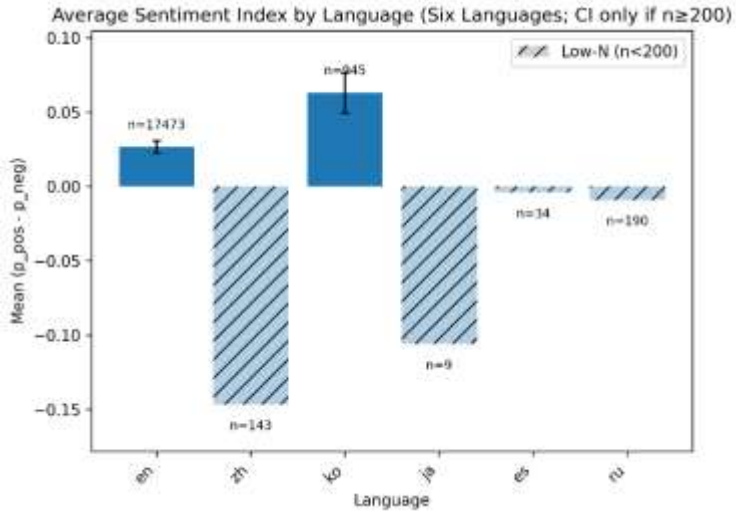

Figure 6. Average Sentiment Index By Language (With Confidence Intervals Under A Sample-Size Threshold)

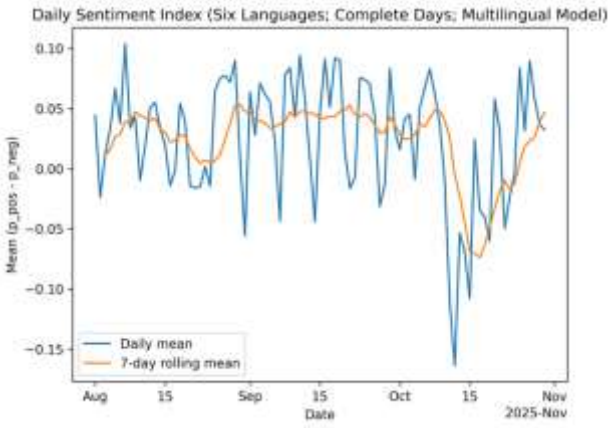

Figure 7. Daily Sentiment Index And 7-Day Rolling Mean (Complete-Day Sample, Six Languages)

To control for incomplete scraping days, we apply a "complete-day" filter. We compute the median daily news count across the six languages (about 209) and set the threshold at 50 using \max(50,0.2\times209). Days below the threshold are removed from the baseline daily series and main empirical sample. Under this rule, the complete-day sample ends on 2025-10-30. The six-language news window spans 2025-08-01 to 2025-11-21, but the main analysis uses 2025-08-01 to 2025-10-30, with the full window used for supplementary presentation.

We construct ML-CryptoSI from standardized group sentiment series. For a group g, we define $z_{g,t} = \frac{s_{g,t} - \overline{s_g}}{\sigma_g}$. A natural information-volume adjustment uses $w_{g,t} = \frac{n_{g,t}}{\sum_g n_{g,t}}$ and $\widetilde{s_t} = \sum_g w_{g,t} z_{g,t}$ (Zhang, 2025). We also consider dispersion/disagreement measures across sources or languages because disagreement is often linked to volume and volatility structure (Yoon and Takahashi, 2025). Because $n_{g,t} - weighting$ may mechanically amplify the dominant language under imbalanced coverage, the baseline feeds unweighted standardized multilingual series into PCA, while the weighted version is kept for robustness. After standardization, we form $Z_t = \left(z_{1,t}, \ldots, z_{G,t}\right)'$ and apply PCA, taking the first principal component as ML-CryptoSI$_t = a_1' Z_t$. We set the sign so that higher values correspond to more positive cross-lingual sentiment, using a reproducible alignment rule with overall daily sentiment (or English sentiment). Rolling-window PCA and exponentially weighted PCA are considered as extensions.

Sentiment extraction is model-dependent, so model choice is treated as a robustness/interpretation issue. We reference FinBERT as a financial-domain baseline (Huang et al., 2022), FinBERT2 as a benchmark in Chinese financial settings (Xu et al., 2025), and non-English "language × finance" specialization such as German FinBERT as an example of domain adaptation (Scherrmann, 2023). We also note that long documents can dilute sentiment and can be handled with key-sentence extraction and discourse simplification (Kim et al., 2025), and that sentiment direction can be domain-specific, motivating direction checks and calibration for narratives such as regulation or deleveraging (Barter et al., 2025). An optional calibration idea is based on market-derived labeling using price

paths and weak labels, followed by in-domain adaptation (Moradi-Kamali et al., 2025; Ider and Lessmann, 2022).

Return predictability is tested using $r_{t+h} = \alpha + \beta\,\text{ML-CryptoSI}_t + \Gamma' X_t + \varepsilon_{t+h}$, where X_t includes standard lagged market-state controls such as lagged returns, volatility proxies, day-of-week effects, and related controls. We use Newey–West standard errors for autocorrelation and heteroskedasticity (Zhang, 2025). The baseline focuses on one-step prediction using the one-lag form r_t on ML-CryptoSI_{t-1} to avoid information leakage. Time variation is examined via rolling-window coefficients $\widehat{\beta_w}$ (Zhang, 2025), and we emphasize incremental explanatory power and scenario heterogeneity (Davidovic and McCleary, 2025). We report out-of-sample RMSE and MAE and optionally test forecast improvement with the Diebold–Mariano test (Jin and Lin, 2025). To test whether sentiment enters variance dynamics, we consider a GARCH-X specification $\varepsilon_t \mid \mathcal{F}_{t-1} \sim (0, h_t) and h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1} + \delta\,\text{ML-CryptoSI}_{t-1} + \phi' X_{t-1}$, wher δ measures the marginal effect of sentiment on conditional volatility and disagreement terms can link the volatility channel to volume–volatility mechanisms (Yoon and Takahashi, 2025). Because GARCH-X benefits from longer samples or higher-frequency data, baseline results prioritize transparent regressions using realized-volatility proxies.

For richer cross-lingual dynamics, $\{z_{g,t}\}$ can be treated as a panel and analyzed using panel VAR or local projections (Jin and Lin, 2025). Interpretability can be added after establishing predictive relevance by applying SHAP or GroupSHAP and grouping features by language/keywords/platforms (Kim et al., 2025). As a complement to LM-based scores, structured narrative dictionaries can be constructed and used as grouped features, following related evidence (Chen, Hwang, and Peng, 2025).

To improve reproducibility and scalability, the pipeline is modularized and logged for auditing. For ethics and compliance, only official or publicly available APIs/RSS feeds are used and text is used only for aggregated statistics; outputs report distributions, time series, and regression results rather than individual items or accounts.

RESEARCH RESULT

This study aggregates multilingual news text and crypto market data to a daily frequency and constructs daily sentiment measures and market variables under a quality-controlled text coverage setting.

Figure 8 plots daily news count N_t. Within the complete-day sample, the daily mean of N_t is 206.43 with a standard deviation of 18.35, a minimum of 148, and a maximum of 237. The full window contains very low-news days on 2025-10-31 (N=6) and 2025-11-21 (N=3). We apply a complete-day threshold $N_t \geq 50$ and drop days with $N_t < 50$. The all-days sample spans 2025-08-01 to 2025-11-21 (93 days). The baseline complete-only sample spans 2025-08-01 to 2025-10-30 (91 days). The baseline sample contains 18,785 news items.

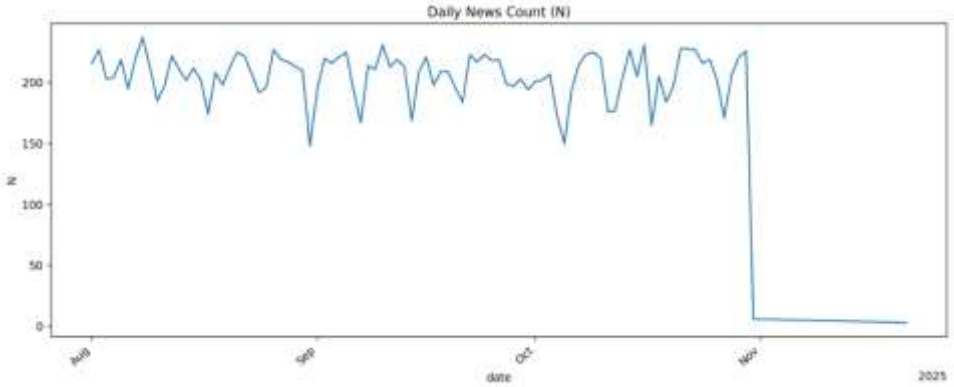

Figure 8. Daily News Count, N

Figure 9 reports average daily label shares. Neutral is about 83.45%, negative about 10.33%, and positive about 6.22%. Figure 10 plots the aggregated daily sentiment index Sent_t. In the baseline sample, Sent_t has mean 0.0249 and standard deviation 0.0513, with a minimum of -0.1638 on 2025-10-12 and a maximum of 0.1040 on 2025-08-07. The within-day textual consistency measure Disagree_std has mean 0.2792 and standard deviation 0.0308.

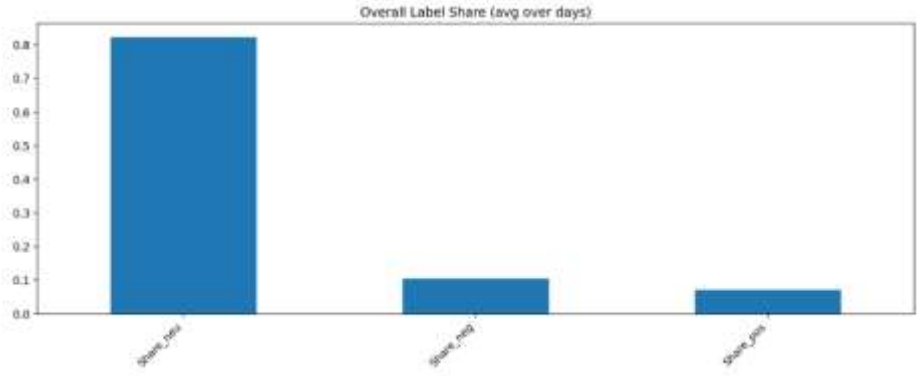

Figure 9. Overall Label Share

Figure 10. Daily Sentiment, Sent

Figures 11 and Figure 12 plot daily log returns for BTCUSDT and ETHUSDT. Because returns are first differences, the return sample size is 90. For BTC, the mean daily log return is -0.00050 and the standard deviation is 0.01863, with range [-0.07586, 0.03909]; the maximum return occurs on 2025-10-01 and the minimum occurs on 2025-10-10. For ETH, the mean daily log return is 0.00097 and the standard deviation is 0.03843, with range [-0.13153, 0.13418]; the maximum return occurs on 2025-08-22 and the minimum also occurs on 2025-10-10. The 7-day rolling realized volatility ($*\_vol\_7$) has mean 0.01735 for BTC and

0.03517 for ETH, with ranges [0.00605, 0.03677] and [0.01029, 0.07485], respectively. Because a 7-day window is used, the effective sample size for rolling volatility is 84. The mean log trading volume is 9.6123 for BTC and 13.0968 for ETH. We also construct the Amihud illiquidity measure as a liquidity control (see Table 1).

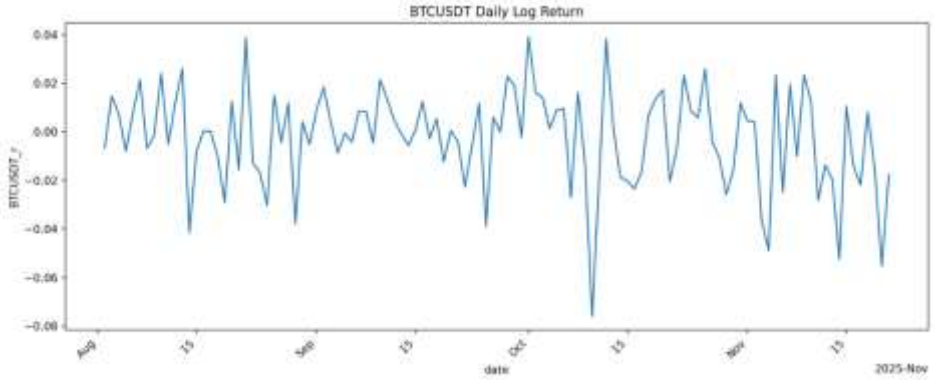

Figure 11. BTCUSDT Daily Log Return

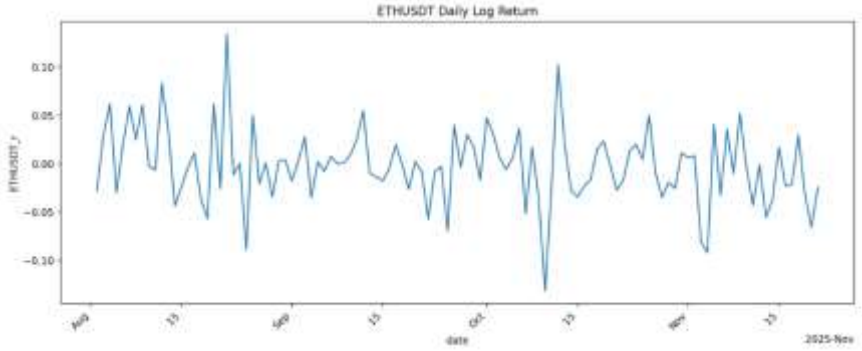

Figure 12.ETHUSDT Daily Log Return

Table 1. Descriptive Statistics (baseline: complete-only)

Variable	Mean	Std. Dev.	Min	Max
Sent	0.0249	0.0513	-0.1638	0.104
N	206.43	18.35	148	237
Disagree_std	0.2792	0.0308	0.2149	0.352
BTCUSDT_r	-0.0005	0.01863	-0.07586	0.03909
BTCUSDT_vol_7	0.01735	0.00719	0.00605	0.03677
BTCUSDT_log_volume	9.6123	0.5143	8.4144	11.0693
BTCUSDT_amihud	8.16×10^{-7}	5.50×10^{-7}	1.26×10^{-8}	3.02×10^{-6}
ETHUSDT_r	0.00097	0.03843	-0.13153	0.13418
ETHUSDT_vol_7	0.03517	0.01732	0.01029	0.07485
ETHUSDT_log_volume	13.0968	0.485	11.7813	14.2169
ETHUSDT_amihud	4.77×10^{-8}	3.35×10^{-8}	1.12×10^{-10}	1.46×10^{-7}

Note: BTCUSDT_r and ETHUSDT_r are daily log returns; vol_7 is the 7-day rolling standard deviation; amihud is the Amihud illiquidity measure; log_volume is log trading volume.

ML-CryptoSI is constructed from multilingual daily sentiment series. For each language $l \in$ {en, es, ja, ko, ru, zh}, we standardize $\text{Sent}_{l,t}$, apply PCA to the multilingual sentiment matrix, and use the first principal component score as $\text{ML-CryptoSI}_t = PC1_t$. We fix the sign so that $\text{corr}(\text{ML-CryptoSI}_t, \text{Sent}_t) > 0$. Figure 5.5 plots ML-CryptoSI (PC1). The standard deviation is 1.09, the minimum is -3.86 on 2025-10-12, and the maximum is 2.08 on 2025-08-14.

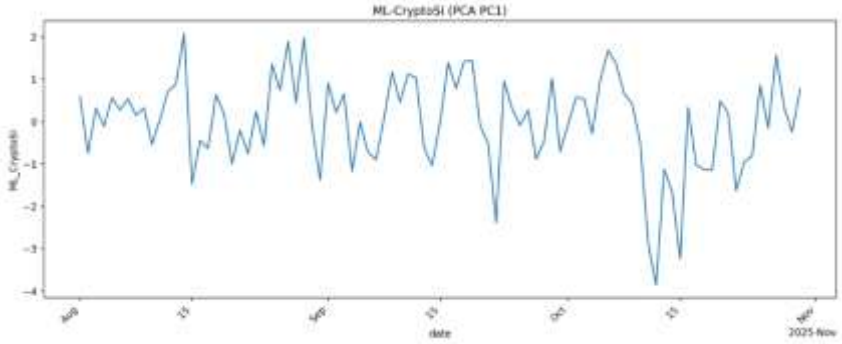

Figure 13. ML-CryptoSI, PCA PC1

Table 2 reports PC1 loadings, language coverage intensity (days_nonzero and share of total news), and correlations between language sentiment and ML-CryptoSI. English loading is 0.771, Korean is 0.546, and Russian is 0.323, while Spanish, Chinese, and Japanese are close to zero. English accounts for about 93.0% of total news volume, Korean about 5.0%, and Russian about 1.0%. Correlations with ML-CryptoSI are 0.84 (English), 0.63 (Korean), and 0.46 (Russian).

Table 2. PC1 Loadings and Language Coverage (Baseline: Complete-Only)

Language	PC1 loading	days_nonzero	share of total N	corr(ML-CryptoSI, Sent_lang)
en	0.7707	91	0.9298	0.8371
ko	0.5464	85	0.0503	0.6287
ru	0.3227	66	0.0101	0.4625
es	0.0499	26	0.0018	0.2443
zh	0.0288	55	0.0075	0.0536
ja	0.0087	7	0.0005	0.083

Note: Language share is the sum of $N_{l,t}$ over the sample divided by total news counts across all languages. "days_nonzero" counts days with $N_{l,t} > 0$.

Daily predictive regressions use $y_t = \alpha + \beta\,\text{ML-CryptoSI}_{t-1} + \Gamma' X_{t-1} + \text{DOW}_t + \varepsilon_t, where X_{t-1}$ includes $N_{t-1}, \text{Disagree\_std}_{t-1}$, lagged returns, lagged volatility (7-day rolling volatility), and lagged log volume. Standard errors use $Newey - West/HAC(maxlags = 3)$. The effective sample size is 83.

Table 3 reports return regressions. For ETH returns, ML-CryptoSI$_{t-1}$ has $\beta = -0.0088$ with $HAC\ s.e. = 0.0046$ and $p = 0.058$. The standard deviation of ML-CryptoSI is about 1.09, so a one-standard-deviation increase corresponds to about $-0.0088 \times 1.09 \approx -0.0096$ in next-day ETH log return. For BTC returns, $\beta = -0.0028$ with $p = 0.221$. For ETH, R^2 increases from 0.119 to 0.149 when adding ML-CryptoSI. For BTC, R^2 increases from 0.147 to 0.159.

Table 3. Baseline Regression: ML-CryptoSI and Daily Returns (HAC, 3 lags)

	(1) BTC r_t	(2) BTC r_t (no ML)	(3) ETH r_t	(4) ETH r_t (no ML)
ML-CryptoSI_{t-1}	-0.0028 (0.0023)	—	-0.00004048	—
Controls	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes
N	83	83	83	83
R^2	0.159	0.147	0.149	0.119
Adj. R^2	0.014	0.015	0.003	-0.018

Note: *Newey–West/HAC standard errors (maxlags=3) in parentheses. *, **, *** denote 10%, 5%, 1% significance.*

Table 4 reports volatility regressions using vol_7. For BTC volatility, ML-CryptoSI$_{t-1}$ is 0.000414 (0.000537). For ETH volatility, it is 0.000530 (0.000927). R^2 changes from 0.769 to 0.771 for BTC and from 0.814 to 0.815 for ETH.

Table 4. Baseline Regression: ML-Cryptosi and 7-Day Volatility (HAC, 3 Lags)

	(1) BTC vol_7	(2) BTC vol_7 (no ML)	(3) ETH vol_7	(4) ETH vol_7 (no ML)
ML-CryptoSI_{t-1}	0.000414 (0.000537)	—	0.000530 (0.000927)	—
Controls	Yes	Yes	Yes	Yes
Day-of-week FE	Yes	Yes	Yes	Yes
N	83	83	83	83
R^2	0.771	0.769	0.815	0.814
Adj. R^2	0.732	0.733	0.783	0.785

As a benchmark, the regression using the single aggregate sentiment series Sent$_{t-1}$ shows a significantly negative coefficient for ETH returns at the 5% level and a marginally significant negative coefficient for BTC at the 10% level.

Rolling regressions use $r_t = \alpha + \beta_t \, \text{Sent}_{t-1} + u_t$ with a 30-day window. Figures 5.6 and 5.7 plot rolling β_t.

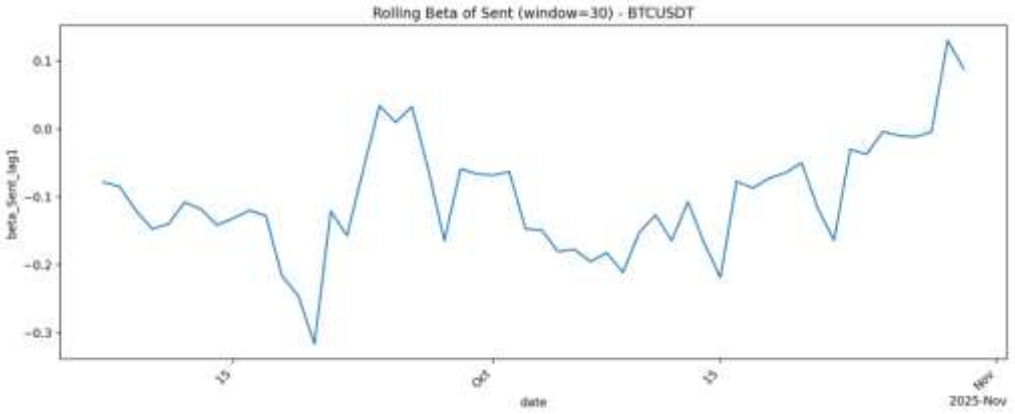

Figure 14. Rolling Beta of Sent, Window=30, BTCUSDT

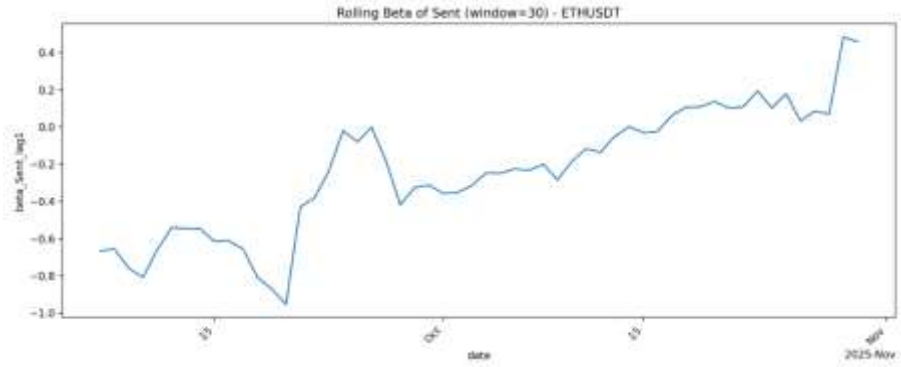

Figure 15. Rolling Beta of Sent, Window=30, ETHUSDT

For BTC, most rolling β_t values lie roughly between -0.22 and -0.05, with a low point near -0.31 and a high near 0.13 close to the end. For ETH, rolling β_t reaches a minimum around -0.95 early in the sample and rises to about 0.45–0.50 near the end, with a phase change toward mostly non-negative values after mid-to-late October. Robustness and heterogeneity checks show that the main sign and significance pattern is stable across sample definitions, sentiment constructions, and control sets, and that predictability is stronger on high news-intensity days. The summary table below reports the key robustness conclusions.

Table 5. Robustness and Heterogeneity

Robustness / heterogeneity test	**Samples**	**Main finding**	**Conclusion**
Complete-only vs all-days	complete-only / all-days	Similar sign and significance	Robust to sample rule
Alternative sentiment index	ML-CryptoSI vs Sent	ML-CryptoSI more significant	Structural aggregation helps
Language decomposition	By-language regressions	en/ko/ru drive effects	Language heterogeneity exists
Control set robustness	Add/remove controls	Coefficient stable	Not driven by controls

News-intensity split	High vs low N	Stronger effect in high N	State dependence present

DISCUSSION

This study examines how multilingual news sentiment relates to BTC and ETH returns and volatility at the daily frequency. After quality control, daily news coverage is sufficiently stable for daily aggregation, and the neutral-dominated label distribution supports using a continuous sentiment index to capture small shifts. A key finding is that ML-CryptoSI, constructed as PCA PC1 of standardized language-level sentiment series, is primarily driven by high-coverage languages. This is consistent with PCA extracting the strongest shared co-movement and down-weighting noisier low-coverage components. ML-CryptoSI therefore summarizes the common multilingual information environment rather than acting as a simple average across languages (Zhang, 2025).

On returns, the predictive regressions show clearer incremental information for ETH than for BTC. ML-CryptoSI is negative and marginally significant for next-day ETH returns, while the BTC coefficient is also negative but not significant. This cross-asset difference is economically plausible because ETH exhibits higher short-horizon volatility in the sample, making it more responsive to narrative and attention shifts. More broadly, the evidence supports a state- and asset-dependent view of predictability (Davidovic and McCleary, 2025). The negative sign should be interpreted as short-horizon dynamics rather than "positive news is bad news." A negative next-day response can reflect short-run reversal and overreaction, where sentiment-driven price moves partially mean-revert, or crowding effects in risk-on states that increase fragility and lead to pullbacks.

On volatility, evidence is weaker when volatility is measured by a 7-day rolling standard deviation: ML-CryptoSI is not significant for BTC or ETH and adds limited explanatory power beyond standard controls. A practical reason is persistence in rolling volatility, while a model-match explanation is that sentiment may enter the variance equation more naturally than reduced-form rolling proxies; disagreement measures may also connect more directly to volatility channels through uncertainty and volume–volatility mechanisms (Yoon and Takahashi, 2025). Rolling and robustness results reinforce the state-dependent interpretation. Sentiment exposure varies over time (especially for ETH), and the main coefficient pattern is stable across sample definitions and alternative sentiment constructions. Language-level evidence indicates that the signal is mainly associated with the high-coverage languages that drive PC1, and predictability strengthens on high news-intensity days, consistent with an information-load mechanism.

These results come with clear limitations. The sample is short and language coverage is highly imbalanced, so inference for low-coverage languages should remain cautious. The baseline index is built mainly from professional news, while social media and on-chain text are treated as extensions, and volatility conclusions depend on proxy choice. Even with these limits, the overall pattern is consistent: a cross-lingual common sentiment factor extracted from multilingual news carries incremental information for short-horizon crypto returns, with a stronger and more state-dependent effect for ETH.

CONCLUSIONS AND RECOMMENDATIONS

Using daily multilingual news text and Binance market data, this paper constructs ML-CryptoSI and evaluates whether a cross-lingual common sentiment factor adds information for BTC and ETH pricing. Five conclusions emerge. First, after quality control and a complete-day rule, daily text coverage is sufficiently stable and sentiment is neutral-dominated, supporting a continuous daily sentiment index. Second, PCA reveals clear language heterogeneity: ML-CryptoSI is mainly driven by high-coverage languages (especially English), with Korean and Russian as secondary contributors, while low-coverage languages contribute little in the current sample. Third, ML-CryptoSI contains incremental information for next-day returns, with stronger evidence for ETH than for BTC. Fourth, rolling estimates indicate time variation in sentiment exposure, consistent with state-dependent predictability rather than a constant relation. Fifth, evidence for volatility prediction is weaker under rolling realized-volatility proxies, suggesting that identifying a robust volatility channel may require longer samples and variance-focused specifications.

These findings imply practical and research recommendations. For trading and risk management, use sentiment as a conditional signal, increasing weight in high-news or high-information states and down-weighting it in low-news periods to reduce noise and overfitting. For asset allocation, the stronger and less stable ETH exposure suggests sentiment factors can help differentiate BTC/ETH risk profiles and guide more dynamic ETH risk budgeting and hedging during sharp sentiment shifts. For cross-lingual monitoring, prioritize data quality and timeliness in high-coverage languages while treating sparse languages with event-triggered attention (e.g., region-specific regulatory shocks). For model design and reporting, make language heterogeneity explicit via alternative aggregation rules (e.g., time-varying or calibrated weights) and language-contribution decompositions so the index remains interpretable.

ADVANCED RESEARCH

This study builds ML-CryptoSI and provides initial evidence on incremental information for returns/volatility and on language heterogeneity, but future work should extend the sample horizon to test stability across market regimes and reduce event-window dependence. It should also move beyond language-level daily aggregation by building a "language × platform" panel dataset that separates professional news from social media and allows clearer comparisons of diffusion mechanisms. Because volatility results based on rolling realized volatility proxies can be weak and unstable, future research should

adopt variance-focused frameworks such as GARCH-X and realized-volatility models such as HAR-RV, and compare multiple volatility definitions including absolute returns, squared returns, and intraday realized volatility. Interpretation can be improved by adding language interaction terms, grouped coefficients, or state-dependent weights and by using narrative dictionaries or topic-model features with dynamic tools such as panel VAR or local projections. Finally, multimodal fusion is a natural extension for extreme volatility and tail risk, especially text embeddings combined with technical indicators (Zou and Herremans, 2022) and, when feasible, acoustic/speech cues from podcasts and video commentary (Todd et al., 2025).

REFERENCES

Alnami, H., Mohzary, M., Assiri, B., & Zangoti, H. (2025). An integrated framework for cryptocurrency price forecasting and anomaly detection using machine learning. Applied Sciences, 15(4), 1864. https://doi.org/10.3390/app15041864.

Amberkhani, A., Bolisetty, H., Narasimhaiah, R., Jilani, G., Baheri, B., Muhajab, H., … Shubbar, S. (2025). Revolutionizing cryptocurrency price prediction: Advanced insights from machine learning, deep learning and hybrid models. In K. Arai (Ed.), Advances in Information and Communication: FICC 2025 (Lecture Notes in Networks and Systems, Vol. 1285, pp. 274–286). Springer. https://doi.org/10.1007/978-3-031-84460-7_18.

Aysan, A. F., Caporin, M., & Cepni, O. (2024). Not all words are equal: Sentiment and jumps in the cryptocurrency market. Journal of International Financial Markets, Institutions and Money, 91, 101920. https://doi.org/10.1016/j.intfin.2023.101920.

Barter, T., Gao, Z., Christodoulaki, E., Chen, J., & Cartlidge, J. (2025). BondBERT: What we learn when assigning sentiment in the bond market (arXiv preprint arXiv:2511.01869). https://arxiv.org/abs/2511.01869.

Binance. (n.d.). Binance API documentation. Retrieved September 25, 2025, from https://developers.binance.com/docs/binance-spot-api-docs/rest-api.

BlockBeats. (n.d.). BlockBeats RSS feeds [RSS]. Retrieved September 25, 2025, from https://api.theblockbeats.news/v2/rss/all.

CoinDesk. (n.d.). CoinDesk RSS feed [RSS]. Retrieved September 25, 2025, from https://www.coindesk.com/arc/outboundfeeds/rss.

CoinGecko. (n.d.). CoinGecko API documentation. Retrieved September 25, 2025, from https://docs.coingecko.com.

Cointelegraph. (n.d.). Cointelegraph RSS feeds [RSS]. Retrieved September 25, 2025, from https://cointelegraph.com/rss.

Dias, I. K., Fernando, J. M. R., & Fernando, P. N. D. (2022). Does investor sentiment predict Bitcoin return and volatility? A quantile regression approach. International Review of Financial Analysis, 84, 102383. https://doi.org/10.1016/j.irfa.2022.102383.

Farrugia, F., & Deguara, C. (2025). Sentiment analysis and cryptocurrency price correlation: A data-driven study. MCAST Journal of Applied Research & Practice, 9(2), 165–184. https://journal.mcast.edu.mt/api/files/view/2920104.pdf.

ForkLog. (n.d.). ForkLog [Website]. Retrieved December 14, 2025, from https://forklog.com/en.

Girsang, A. S., & Stanley. (2023). Hybrid LSTM and GRU for cryptocurrency price forecasting based on social network sentiment analysis using FinBERT. IEEE Access, 11, 120530–120540. https://doi.org/10.1109/ACCESS.2023.3324535.

Gurgul, V., Lessmann, S., & Härdle, W. K. (2023). Forecasting cryptocurrency prices using deep learning: Integrating financial, blockchain, and text data [Computer software]. GitHub. https://github.com/Humboldt-WI/CC-Price-Forecasting.

Gurgul, V., Lessmann, S., & Härdle, W. K. (2025). Deep learning and NLP in cryptocurrency forecasting: Integrating financial, blockchain, and social media data. International Journal of Forecasting, 41, 1666–1695. https://doi.org/10.1016/j.ijforecast.2025.02.007.

Hamayel, M. J., & Owda, A. Y. (2021). A novel cryptocurrency price prediction model using GRU, LSTM and bi-LSTM machine learning algorithms. AI, 2(4), 477–496. https://doi.org/10.3390/ai2040030.

Han, S. O. (2025). Investor sentiment and cross-section of cryptocurrency returns. Journal of Behavioral and Experimental Finance, 46, 101043. https://doi.org/10.1016/j.jbef.2025.101043.

Huang, A. H., Wang, H., & 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.

Ider, D., & Lessmann, S. (2022). Forecasting cryptocurrency returns from sentiment signals: An analysis of BERT classifiers and weak supervision (arXiv preprint arXiv:2204.05781). https://arxiv.org/abs/2204.05781.

Jin, X., & Lin, S.-L. (2025). An early prediction model on systemic risk under global risk: Using FinBERT and temporal fusion transformer to multimodal data fusion framework. The North American Journal of Economics and Finance, 76, 102361. https://doi.org/10.1016/j.najef.2025.102361.

John, K., Li, J., & Liu, R. (2024). Sentiment in the cross section of cryptocurrency returns (Working paper; CryptoSent index). New York University Stern School of Business; Stevens Institute of Technology. (Working paper landing/info page) https://ronming1303.github.io.

Kleitsikas, C., Korfiatis, N., Leonardos, S., & Ventre, C. (2025). Bitcoin's edge: Embedded sentiment in blockchain transactional data. In IEEE International Conference on Blockchain and Cryptocurrency (ICBC 2025). (Preprint) https://arxiv.org/abs/2504.13598.

Koutmos, D. (2023). Investor sentiment and Bitcoin prices. Review of Quantitative Finance and Accounting, 60, 1–29. https://doi.org/10.1007/s11156-022-01086-4.

Lupu, R., & Donoiu, P. C. (2025). Sentiment matters for cryptocurrencies: Evidence from tweets. Data, 10(4), 50. https://doi.org/10.3390/data10040050.

Mokni, K. (2022). Investor sentiment and Bitcoin relationship: A quantile-based analysis. North American Journal of Economics and Finance, 60, 101657. https://doi.org/10.1016/j.najef.2021.101657.

Moradi-Kamali, H., Rajabi-Ghozlou, M.-H., Ghazavi, M., Soltani, A., Sattarzadeh, A., & Entezari-Maleki, R. (2025). Market-derived financial sentiment analysis: Context-aware language models for crypto forecasting (arXiv preprint arXiv:2502.14897). https://arxiv.org/abs/2502.14897.

Natzir, S. M., & Jatiprasetya, H. (2025). Prediksi harga cryptocurrency XLM menggunakan metode deep learning LSTM dan GRU [Predicting XLM cryptocurrency prices using LSTM and GRU deep learning models]. HOAQ: Jurnal Teknologi Informasi, 16(1), 49–58. https://doi.org/10.52972/hoaq.vol16no1.p49-58.

Odaily. (n.d.). Odaily RSS feeds [RSS]. Retrieved September 25, 2025, from https://rss.odaily.news/rss/newsflash.

PANews. (n.d.). PANews RSS feed [RSS]. Retrieved September 25, 2025, from https://rss.panewslab.com.

Ponselvakumar, A. P., Giri Shankar, V. P., Iniyan, G., & Logesh, B. (2024). Improving the cryptocurrency price prediction using deep learning. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), Intelligent Systems Design and Applications (ISDA 2023) (Lecture Notes in Networks and Systems, Vol. 1048, pp. 145–153). Springer. https://doi.org/10.1007/978-3-031-64650-8_14.

Seabe, P. L., Moutsinga, C. R. B., & Pindza, E. (2023). Forecasting cryptocurrency prices using LSTM, GRU, and bi-directional LSTM: A deep learning approach. Fractal and Fractional, 7(2), 203. https://doi.org/10.3390/fractalfract7020203.

Seabe, P. L., Moutsinga, C. R. B., & Pindza, E. (2025). Sentiment-driven cryptocurrency forecasting: Analyzing LSTM, GRU, Bi-LSTM, and temporal attention model. Social Network Analysis and Mining, 15, 52. https://doi.org/10.1007/s13278-025-01463-6.

Tiwari, D., Bhati, B. S., Nagpal, B., Al-Rasheed, A., Getahun, M., & Soufiene, B. O. (2025). A swarm-optimization based fusion model of sentiment analysis for cryptocurrency price prediction. Scientific Reports, 15, 8119. https://doi.org/10.1038/s41598-025-92563-y.

Todd, A., Bowden, J., Cummins, M., & Su, Y. (2025). A multimodal sentiment classifier for financial decision making. International Review of Financial Analysis, 105, 104322. https://doi.org/10.1016/j.irfa.2025.104322.

TokenPost. (n.d.). TokenPost RSS feed [RSS]. Retrieved December 14, 2025, from https://www.tokenpost.kr/rss.

Xiao, Y., Sun, E., Luo, D., & Wang, W. (2024). TradingAgents: Multi-agents LLM financial trading framework (arXiv preprint arXiv:2412.20138). https://arxiv.org/abs/2412.20138.

Xu, Z., Wang, L., & Zhou, X. (2025). FinBERT2: Advancing financial language understanding with domain-adaptive large models (arXiv preprint arXiv:2506.06335). https://arxiv.org/abs/2506.06335.

Yamak, P. T., Yujian, L., & Gadosey, P. K. (2019). A comparison between ARIMA, LSTM, and GRU for time series forecasting. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2019) (pp. 49–55). Association for Computing Machinery. https://doi.org/10.1145/3514262.3514331.

Zhang, J., Cai, K., & Wen, J. (2024). A survey of deep learning applications in cryptocurrency. iScience, 27(1), 108509. https://doi.org/10.1016/j.isci.2023.108509.

Zou, Y., & Herremans, D. (2023). PreBit — A multimodal model with Twitter FinBERT embeddings for extreme price movement prediction of Bitcoin. Expert Systems with Applications, 233, 120838. https://doi.org/10.1016/j.eswa.2023.120838.