Proceedings of the 5th International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/2026.38470

A Comparative Analysis of Volatility Predictions via EGARCH, Transformer and Hybrid EGARCH-Transformer Models: Changes in Forecasting Ability after Introducing Sentiment Factors

Yafei Wang[1,†], Qiujin Yu[2,*,†], Jiangchuan Peng[3,†], Bokai Yang[4,†], Ziqing Chen[5,†]

[1]*Department of Mathematics, University of Manchester, Manchester, UK*
[2]*School of Finance, Southwestern University of Finance and Economics, Chengdu, China*
[3]*Mathematics, Santiago Canyon College, Newport Coast, USA*
[4]*Statistical Science, University College London, London, UK*
[5]*International Business Management, Dongbei University of Finance and Economics, Dalian, China*
**Corresponding Author. Email: 42304439@smail.swufe.edu.cn*
†These authors contributed equally to this work and should be considered as co-first authors

Abstract. The application of cryptocurrency asset volatility forecasting is relevant for real-world risk-management procedures and derivatives pricing in the digital-asset market. This paper assesses out-of-sample forecast performance for three model families: standalone EGARCH, vanilla Transformer, and an EGARCH-Transformer combined hybrid architecture for forecasting Bitcoin volatility. Another goal is to measure whether there is any actual marginal forecasting benefit from investor sentiment obtained from the Reddit posts. We use information from Reddit with the subreddit r/Bitcoin from 25th March 2021 to 30th December 2024. We collaborate with the open-source ProsusAI/finbert language model that has been trained for financial text tasks to construct a daily sentiment score based on the number of upvotes given to the post-comment. In the hybrid model we constructed, the one-step-ahead conditional-variance values produced by EGARCH are used as other features as input to the Transformer network. The results from controlled feature-ablation tests show that, in general, using structured volatility inputs (volatility realized from volatility inputs) and EGARCH conditional volatility outputs yield better forecasts. Social media sentiment data of investors, by comparison, has mixed results. Under specific combinations of features, sentiment boosts the accuracy of predictions, but in other experimental settings, it hinders the accuracy of predictions. We have problems consistently filtering subjective signals from forum text by retail investors with our hybrid EGARCH-Transformer implementation. Results indicate that the performance of hybrid models is much more influenced by the composition of the input features and quality of the raw data than by adding additional predictor variables to the network.

Keywords: EGARCH, Transformer, volatility, Bitcoin, sentiment

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.38470

1. Introduction

The cryptocurrency markets are inherently volatile, exhibit non-linear price dynamics, and have obvious feedback with retail investor sentiment. Prediction of volatility is thus one of the fundamental problems of quantitative studies on digital assets. Volatility estimates feed into asset-valuation routines, risk-alert workflows and derivatives-pricing pipelines. The accuracy of volatility forecasts is crucial for the performance of investment strategies and risk management measures under real market conditions.

ARCH- and GARCH-type specifications have long been used as econometric tools in practice to fit financial time-series returns. They are good models if conditional heteroskedasticity and well-known volatility-clustering effects [1, 2] are assumed. Nevertheless, there are real constraints to fixed-parameter econometric frameworks. They are not effective at decomposing the complex nonlinear dynamics, sudden breakpoints in the regimes, and the simultaneous evolution of numerous predictors. Relevant predictors here include trading-signal technical indicators, derivatives-market readings, macroeconomic series, and investor-sentiment proxies [3, 4].

The multi-head self-attention Transformer architecture was introduced to the time-series research field through the progress of deep-learning. It alleviates the known defects of earlier recurrent-network designs. The model can capture long-range nonlinear dynamics from high-dimensional financial time series as well as volatility jumps that occur abruptly in time, providing a very real advantage for volatility forecasting applications [5]. The suitability of Transformer-class models for these tasks has already been shown in previous cryptocurrency research. Different variants of Transformers were created: Synthesizer-style and sentiment-aware. Those studies combined on-chain data, exchange transaction data, and social-media sentiment data to predict significant fluctuations in bitcoin. The usefulness of attention-based layers and sentiment-related features for the modelling of digital-asset volatility has been confirmed in published work [6, 7]. This adoption of deep-learning models is also reported in the broad literature of systematic reviews of financial-time-series modelling, where such models are able to extract nonlinear patterns from noisy real-world financial time-series data [8]. The Informer architecture uses sparse-attention to improve efficiency when handling long-horizon sequence-forecasting tasks with a long input window [9].

More recent studies start to combine classical econometric tools and deep-learning encoders to build hybrid GARCH-Transformer set-ups. This design option is made to maintain the economic interpretability of traditional time-series designs while maintaining the machine-learning capabilities of nonlinear pattern fitting. An illustrative study by Gebreslase et al. [10] passed conditional-variance predictions from different GARCH-family models into a Transformer encoder. They offer a hybrid workflow, which is able to capture volatility persistence, leverage-effect responses, and long-run temporal dependencies.

There are still some empirical gaps in the existing hybrid-model literature. A lot of previous experiments use merely inputs from price and return time-series. Despite the fact that investor sentiment can lead to extreme volatility in the crypto market, they mostly ignore such behavioural-finance factors. Realized-volatility level, skewness, and kurtosis change with indicators like the Fear-and-Greed Index, VIX, and put-call ratios. There are cross-feedback dynamic relationships between sentiment and higher-order volatility moments [11, 12].

There are many studies that regard sentiment measurement, fitting GARCH-type volatility, and constructing deep-learning models as independent research areas. There are few papers that conduct side-by-side comparisons of the aforementioned standalone EGARCH, vanilla Transformer, and

DOI: 10.54254/2754-1169/2026.38470

hybrid EGARCH-Transformer under the same experimental conditions. Few controlled ablation tests that measure exactly how much out-of-sample predictive gain sentiment variables bring are available. However, behavioural-finance theory provides sound conceptual motivation for sentiment inputs. Baker and Wurgler [13] record sentiment-related asset-valuation changes, especially for hard-to-arbitrage assets. Bollen et al. [14] demonstrate that the aggregated, social-media mood conveys predictive signals for financial outcomes at the broader-market level.

In this context, the multi-dimensional sentiment-augmented volatility-prediction framework is developed for Bitcoin in this paper. The three families of models that serve as baseline models are the standalone EGARCH, standalone Transformer, and hybrid EGARCH-Transformer, and modified versions of each of those with sentiment inputs. The features that we use include the Bitcoin log-returns, realized volatility, implied volatility readings, technical trading indicators, macro-financial time series, and our own investor-sentiment proxy. We maintain the interpretability of the EGARCH model and the nonlinear representation ability of the Transformer model, and embed the behavioural-sentiment features in the forecasting framework. The RMSE, MAE, R^2 and QLIKE are used to evaluate the model performance. We wanted to perform Diebold-Mariano significance tests for the comparison of forecasts, but due to computational constraints, we were not able to do this additional statistical analysis. Thus, we simply use numeric-metric comparison for performance ranking. This restriction will be discussed later. Our contributions bridge the gap in the empirical literature of unified controlled multi-model comparisons for sentiment-enhanced crypto-volatility forecasting and provide a hands-on reference for digital-asset risk assessment and derivatives-market work.

We have four concrete outputs of our empirical research. We construct a uniform experimental setup, including the same train-test split, the same feature pool, and the same evaluation metrics, to compare standalone EGARCH, vanilla Transformer, and hybrid EGARCH-Transformer. Second, we build a Reddit-sourced daily sentiment index weighted by the number of upvotes on the content consumed by retail investors and perform controlled ablation tests to quantify the out-of-sample predictive power of text sentiment from retail investors. Overall, our empirical findings indicate that social-media sentiment is not a universal mechanism to enhance the forecasting accuracy of Bitcoin volatility. The quality of hybrid models depends significantly on the sets of input features in its vicinity and is not solely defined by the network structure. Finally, we make practical implications for the practice of digital-asset risk-modelling. Better forecasting is not guaranteed simply by putting behavioural-sentiment proxies on the end. The number of raw features isn't as important as feature quality and complementarity between features.

This paper is structured as follows. A review of pertinent literature is presented in Section 2. Section 3 details the sources of data, the determination of sentiment-indexes and the complete specification of the models. Section 4 presents the results of empirical tests. In Section 5, the results are interpreted in conversation with existing literature. Limitations of the study and suggestions for future research are acknowledged at the end of Section 6.

2. Literature review

2.1. Research on traditional GARCH family volatility models

The study of traditional GARCH family volatility models is reviewed in this section. Standard benchmark specifications for the estimation of financial-asset volatility are GARCH-family specifications which are based on ARCH-theoretical foundations. They are good at modelling conditional heteroskedasticity and volatility-clustering effects in return time series [1, 2]. A variety

DOI: 10.54254/2754-1169/2026.38470

of GARCH variants have been tested and extended to equity-market and cryptocurrency data, with various residual-distribution assumptions to optimize their performance.

In the cryptocurrency-market research, Gyamerah [15] uses daily observations of Bitcoin and compares the performance of the sGARCH, iGARCH, and tGARCH models with different residual distributions. His findings suggest that tGARCH is able to account for the asymmetric news-response leverage effects of crypto-asset volatility. The NIG distribution is a good fit for leptokurtic and skewed data, which is typical of Bitcoin returns data [16]. In the realm of single-variant GARCH models, the tGARCH combined with NIG distribution is the most flexible in modelling the volatility of digital-assets based on the best in-sample fit. Pichl and Kaizoji [17] extended the basic GARCH model with jump components, the HAR-RV-J model. That model is flexible and is able to capture a price-jump event occurring frequently in intraday price action of Bitcoin.

Even the well-specified GARCH-variant models bring in constraints for complex cryptocurrency-market settings. Linear-parametric boundaries restrict their capability to capture long-range nonlinearities and abrupt market-regime changes [10]. Forecasting quality is significantly affected during big market moves and volatility shocks. The usual GARCH-style implementations also rely nearly exclusively on endogenous price and return data. Their simple built-in mechanisms aren't easily applicable to explain abnormal crypto-asset-market volatility in the presence of exogenous behavioural factors like investor sentiment.

2.2. Transformer model and sentiment-aware volatility prediction research

However, the multi-head self-attention mechanism in Transformer addresses some drawbacks of previous classical time-series models, including problems of signal decay over long time horizons and inefficient nonlinear-feature extraction [5]. This architecture is good at automatically mapping complicated correlations within long financial sequences and sudden volatility movements, and is a popular modern tool for volatility-forecasting tasks. Transformer eliminates iterative (cyclical) computation, unlike setups with recurrent neural networks, allowing parallel-processing time-series fitting. Multi-head attention layers are capable of attending to different feature subspaces, making them well-suited for multi-factor risk-driven volatility-prediction scenarios. The Informer model is another variant of self-attention that is tailored for forecasting future events from long sequences and is more efficient in terms of computational cost than self-attention with large historical input windows [9]. There is broad evidence of the adoption of deep learning in financial time-series work, as it can model complex nonlinear data patterns [8].

Within Bitcoin-volatility forecasting, researchers combine Transformer architectures with sentiment inputs to embed behavioural-market signals. Herremans and Low [7] implement a Synthesizer Transformer fed with on-chain indicators, exchange-transaction statistics, and miner-behaviour features to predict large Bitcoin-volatility spikes. Biabdillah et al. [6] develop a sentiment-aware Transformer that ingests technical indicators plus social-media sentiment scraped from Twitter, Reddit, and financial-news sources. Their pipeline jointly models market fundamentals and investor sentiment for crypto-volatility prediction. Those prior studies validate Transformer-model usefulness and point to potential value in sentiment-related inputs for digital-asset volatility work. Their textual-data-modelling approaches align with behavioural-finance findings linking sentiment to asset valuation [13] and research showing aggregated social-media mood carries market-predictive content [14].

DOI: 10.54254/2754-1169/2026.38470

2.3. Research on hybrid GARCH-transformer fusion model and research gap

To overcome weaknesses found in standalone econometric and deep-learning approaches, researchers build hybrid GARCH-deep-learning fusion pipelines. These form the theoretical foundation for our present study. Multiple papers show volatility outputs fitted within GARCH workflows serve as informative input features for machine-learning encoders, allowing each modelling paradigm to compensate for the other's blind spots [16]. Large-scale forecasting-competition outcomes back hybrid-modelling thinking. Results from the well-known M4 Competition show combined or hybrid set-ups regularly outperform isolated single-model baselines. Statistical-plus-machine-learning hybrid designs often deliver superior forecast accuracy [18]. For Bitcoin-volatility forecasting specifically, Kristjanpoller and Minutolo [19] observe improved nonlinear-volatility dynamics modelling when GARCH-family outputs are combined with neural-network learning.

Gebreslase et al. [10] innovate a multi-GARCH-Transformer hybrid built for stock-market volatility forecasting. Conditional-variance outputs generated by sGARCH, EGARCH, GJR-GARCH and alternative variants are passed into a Transformer encoder. Self-attention layers dynamically weight different volatility-signal sources. Their framework retains volatility-persistence and leverage-effect properties coming from financial-econometric theory, while harnessing deep-learning nonlinear-fitting power. Empirical tests in their work demonstrate hybrid-model advantages in out-of-sample forecasting and market-regime-switch responsiveness over pure GARCH or pure Transformer benchmarks. Nevertheless, the current hybrid-model studies are largely based on the equity markets. They largely disregard multi-dimensional investor-sentiment drivers, which are relevant in explaining cryptocurrency volatility.

Based on the literature synthesis, three research gaps are identified for cryptocurrency-volatility forecasting.

First, sentiment-measurement work, GARCH volatility modelling, Transformer deep-learning research, and hybrid-fusion developments mostly run as separate streams of research. However, there are few consolidated head-to-head studies comparing the performance of standalone GARCH, standalone Transformer, and hybrid GARCH-Transformer, with and without sentiment factors.

Second, few controlled paired experiments (model runs with and without sentiment inputs) have been reported. Marginal effects of sentiment on out-of-sample forecast error are similarly under-developed.

Third, most published previous work is based on static single-feature settings and relatively limited empirical viewpoints. There is a lack of papers that bring together several sentiment proxies and test the marginal forecasting contribution of different behavioural-finance-related inputs in a systematic manner.

This paper fills those gaps by building sentiment-augmented EGARCH-Transformer hybrid specifications. We perform cross-model comparisons to assess the value of investor-sentiment data as a true measure of its usefulness for Bitcoin-volatility forecasting.

3. Data and methodology

3.1. Sentiment data collection and sample selection

We've scraped the text data from Reddit's open API to create sentiment-factor inputs. We have a strictly defined sampling window, which is the period of market data availability – 25 March 2021 to 30 December 2024.

DOI: 10.54254/2754-1169/2026.38470

Only material from the r/Bitcoin subreddit is used as text-based sources. There were three practical reasons for choosing this community. We concentrate on discussion threads directly related to Bitcoin, aligning with our forecasting target variable. When compared to bigger subreddits such as r/CryptoCurrency, r/Bitcoin filters out some of the "noise" generated by discussions about numerous other cryptocurrencies. Trading-focused subreddits (such as r/BitcoinMarkets) are also off-limits. Those communities are smaller, with fewer total users, and traders tend to be more professional. They do not consistently represent the market consensus of retail-investors as a whole.

The volumes of posts on Reddit are very large every day. Due to our practical computational constraints, we used targeted sampling. Each calendar day, we take the top-100 highest-upvoted posts. For each selected post, we further collect its top-10 highest-upvoted comment threads. This sampling is a compromise between computational feasibility and noise filtering logic. It is worth noting that the more popular (higher upvotes) posts and comments tend to better match the consensus of retail investors, as compared to lesser-seen and less-upvoted posts. Hyperlinks, special symbols, empty records, and corrupted text entries are removed at the preprocessing step.

3.2. Sentiment factor construction

Cleaned posts and comments are merged together into a single consolidated dataset. Each individual text entry receives a cross-sectional weight equal to its public upvote count, which proxies for its relative market influence among forum participants.

Sentiment scoring for unstructured text makes use of ProsusAI/finbert, a publicly-available pre-trained language model fine-tuned for financial-domain textual analysis. For each text sample, FinBERT outputs probability estimates for positive, negative, and neutral sentiment categories. The text-level sentiment score is defined:

$$\mathrm{Score_i} = \mathrm{P(Positive)_i} - \mathrm{P(Negative)_i}$$

We aggregate individual text-level scores to build a daily market-sentiment index. Calculation uses upvote-weighted averaging over all valid observations available on day: $\mathrm{Daily_Sentiment_t} = \frac{\sum_{i=1}^{N}(\mathrm{Score_{i,t}} \times \mathrm{Upvotes_{i,t}})}{\sum_{i=1}^{N}\mathrm{Upvotes_{i,t}}}$

Where N_t gives the total valid sampled texts (posts plus comments) for each trading day. The resulting $\mathrm{Daily_Sentiment_t}$ time-series acts as our primary sentiment-feature input for subsequent volatility-prediction model runs.

3.3. EGARCH specification

The daily log return on Bitcoin is defined as where:

$$\mathrm{r_t} = \ln(\mathrm{P_t}) - \ln(\mathrm{P_{t-1}}) = \sqrt{\mathrm{h_t}}\varepsilon_t \quad (1)$$

Where P_t is the closing price, h_t is the conditional variance, and ε_t follows a standardized Student-t distribution.

The GARCH(1,1) model is specified as

$$h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1} \quad (2)$$

DOI: 10.54254/2754-1169/2026.38470

α captures the impact of new market shocks; β measures volatility persistence.

GJR-GARCH(1,1) incorporates an asymmetric shock-response term:

$$h_t = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 I_{\{\varepsilon_{t-1}<0\}} + \beta h_{t-1} I_{\{\varepsilon_{t-1}<0\}} \quad (3)$$

When $\varepsilon_{t-1} < 0$, $I_{t-1} = 1$. Statistically significant γ signals asymmetric volatility reactions to positive versus negative shocks of identical magnitude.

The Exponential EGARCH uses log-conditional variance to guarantee non-negativity without parameter constraints:

$$\ln(h_t) = \omega + \sum \alpha_i g(\varepsilon_{t-i}) + \sum \beta_j \ln(h_{t-j}) \quad (4)$$

Estimation starts with full specifications containing asymmetric response terms. Asymmetric coefficients are retained only when they reach statistical significance. Guided by information-criterion outputs and residual-diagnostic checks, EGARCH(2,1) is chosen as our baseline econometric model. All GARCH-family estimations apply maximum likelihood estimation under Student-t distribution assumptions. For out-of-sample forecasting, we implement an expanding window scheme: at every time step, estimation re-runs over all historical data available up to that date, without discarding earlier observations.

3.4. Setup transformer model

All experimental settings are kept constant across different variants of the Transformer models. The forecasting target is one-day-ahead realized volatility (RV_1D). The length of the input look-back window is set to 20 time steps. The entire dataset was divided into a 70 %training/30% out-of-sample test split. A random seed of 42 is used to facilitate reproducibility. The raw data that we use for the baseline are the OHLCV price data for Bitcoin. These evaluation metrics are RMSE, MAE, $R^2$and the QLIKE loss function.

Table 1. Hyper-parameter configuration of transformer model

Category	Parameter	Value
Architecture	Lookback window	20 days
	d_{model}	32
	Multi-head attention heads	4
	Encoder layers	2
	Dropout rate	0.1
	Output activation	Softplus
Training	Loss function	MSE
	Optimizer	Adam
	Learning rate	0.001
	Batch size	32
	Maximum epochs	100
	Early stopping	patience=10

DOI: 10.54254/2754-1169/2026.38470

Nine different model setups (Model 1 to Model 9) are designed by controlled feature ablation. The network architecture is the same, and differences in input-feature combinations are made between specifications.

Table 2. Feature composition for Model1 to Model9

Model	Full set of input features
Model1	transformer(OHLCV)
Model2	OHLCV+Realized Volatility (RV)
Model3	OHLCV + RV + VIX
Model4	OHLCV + RV + VIX + EGARCH conditional-variance forecast
Model5	OHLCV + RV + VIX + S&P500 + DXY
Model6	Model 4 feature set + Reddit-based daily sentiment index
Model7	OHLCV + RV + EGARCH conditional-variance forecast
Model8	Model 7 feature set + Reddit-based daily sentiment index
Model9	OHLCV + RV + VIX + S&P500 + DXY + EGARCH conditional-variance forecast

4. Empirical results

4.1. EGARCH estimation results

4.1.1. Descriptive statistics and model selection.

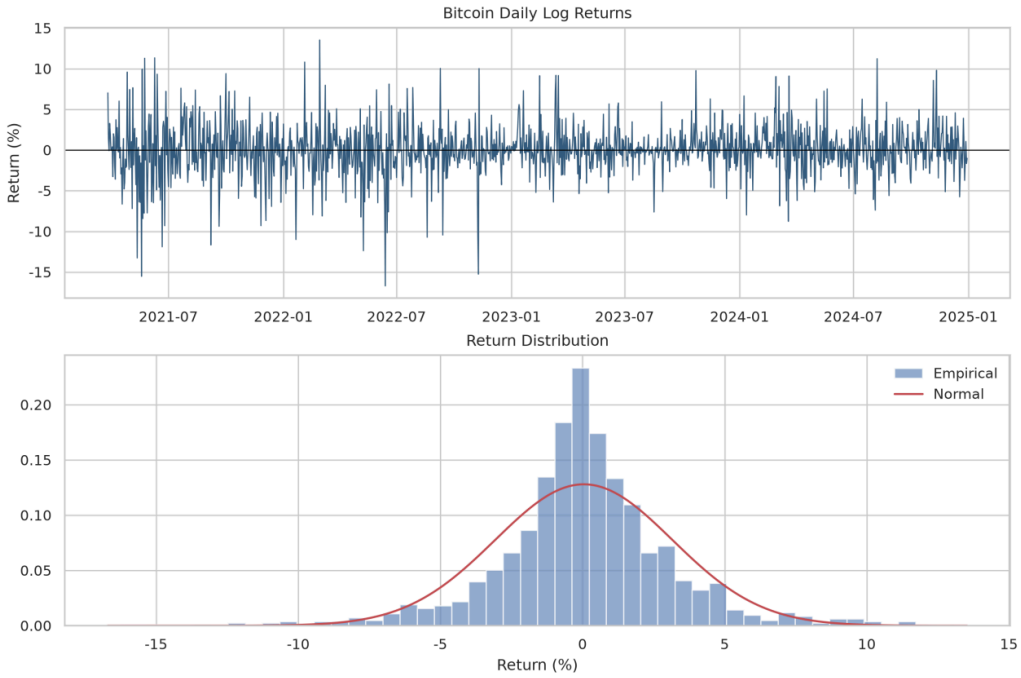


Figure 1. Bitcoin daily returns and their distribution

Our price-sample time range runs from 25 March 2021 through 30 December 2024. The first usable log-return observation corresponds to 26 March 2021, delivering 1376 total daily

DOI: 10.54254/2754-1169/2026.38470

observations. The sample mean return reaches 0.0431%, standard deviation = 3.1171 %, skewness= −0.2509, excesskurtosis=3.4156.

Figure 1 displays the behaviour of the return series and its empirical return distribution. Alternating periods of high and low volatility market conditions are observed in the data set. The empirical return distribution has leptokurtic characteristics and is fatter-tailed than a normal distribution.

ADF unit-root test statistic equals −11.8356 ($p < 0.001$), confirming return-series stationarity. ARCH-LM test with 10 lags returns statistic = 61.7093 ($p < 0.001$), providing clear statistical evidence for conditional heteroskedasticity.

Model estimation uses 1173 in-sample observations ending 10 June 2024. AIC values read 5757.32 (GARCH(1,1)), 5759.29 (GJR-GARCH(1,1)), and 5749.44 (asymmetric-term EGARCH(1,1)). Asymmetric-coefficient *p*-values for GJR-GARCH and EGARCH hit 0.996 and 0.969 respectively. Neither coefficient reaches statistical significance in this particular sample. That means we cannot confirm stable asymmetric shock-response patterns for our studied time window. Among these three candidates, EGARCH yields the lowest AIC and BIC values and is selected as the baseline econometric specification.

When comparing different EGARCH lag orders, EGARCH(3,3) delivers minimal AIC = 5735.04, while EGARCH(2,1) produces minimal BIC = 5769.95. For EGARCH(2,1), 10-lag Ljung-Box test on squared residuals yields $p = 0.536$; ARCH-LM test gives $p = 0.516$. No remaining ARCH-type effects show up within model residuals. We finally select EGARCH(2,1). Although EGARCH(3,3) achieves better in-sample AIC, EGARCH(2,1) demonstrates cleaner residual diagnostics and superior out-of-sample forecasting performance.

4.1.2. Out-of-sample forecasting and hybrid model input preparation

The hold-out test-sample spans 11 June 2024 to 30 December 2024 and includes 203 daily observations. With our expanding-window procedure, the model gets re-estimated each calendar day using all historical data available up to the prior trading day, producing one-step-ahead conditional-variance forecasts. Realized volatility (RV) acts as our target-volatility measure. Both volatility metrics are reported in percentage terms and remain non-annualized.

MAE, RMSE and R^2 metrics evaluate volatility-forecast quality. QLIKE loss is calculated over variance-forecast outputs.

Table 3. Out-of-sample performance of EGARCH-family models

Model	MAE	RMSE	R^2	QLIKE
EGARCH(2,1)	0.8959	1.2015	-0.1528	0.3180
EGARCH(3,3)	0.9022	1.2524	-0.2527	0.3410

EGARCH(2,1) outperforms EGARCH(3,3) across all four evaluation metrics.

From Figure 2 we can observe that EGARCH conditional-variance forecasts stay relatively smooth. The model substantially underestimates large volatility spikes that arrived in early August 2024.

For the fitted EGARCH(2,1), parameters $\alpha_1 = 0.3857$, $\alpha_2 = -0.2543$, and $\beta_1 = 0.9979$ are all significant at the 1 % significance level. The β_1 estimate sits very close to 1, signalling strong volatility-persistence behaviour. Student- t degrees-of-freedom parameter equals 2.9189, consistent with heavy-tailed return-distribution traits.

DOI: 10.54254/2754-1169/2026.38470

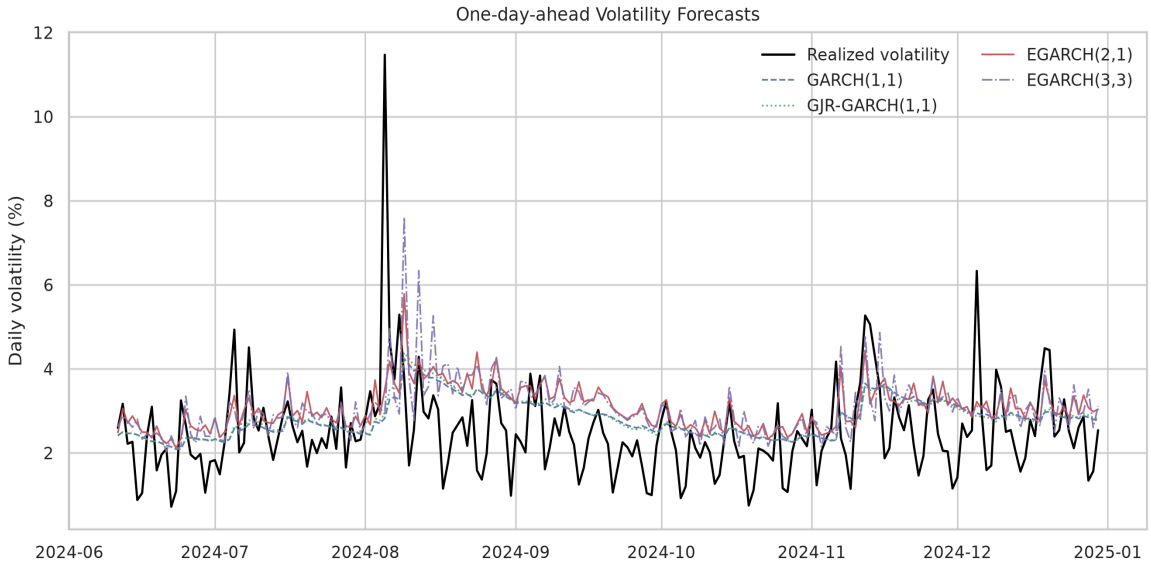


Figure 2. Out-of-sample forecast comparison of GARCH-type models

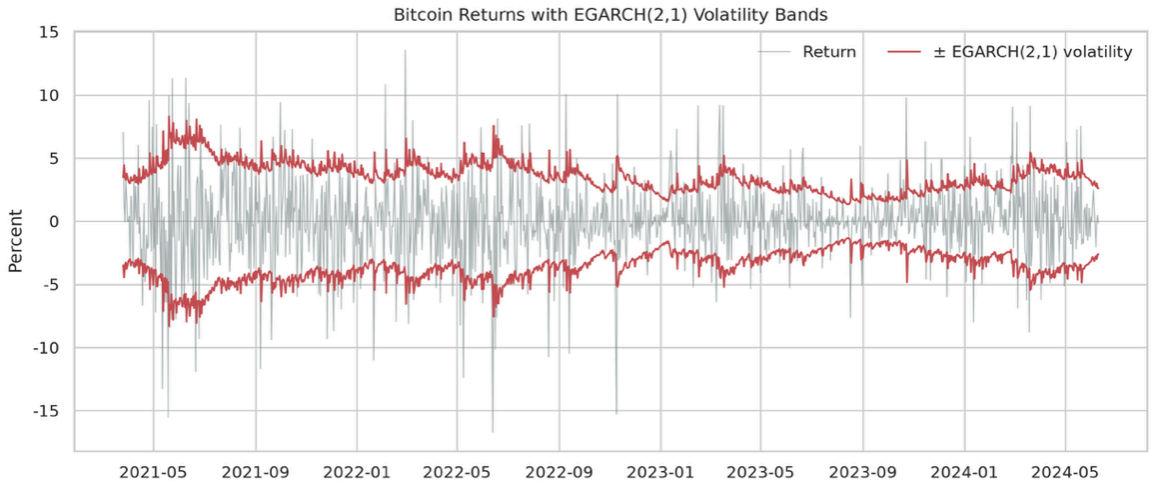


Figure 3. Conditional volatility from EGARCH(2,1)

Figure 3 visualizes Bitcoin return series together with EGARCH(2,1) conditional-volatility bands. In summary, the EGARCH baseline captures volatility-persistence dynamics reasonably well, yet responds poorly to abrupt volatility-regime shifts. We extract its one-step-ahead conditional-variance forecasts and feed those as supplementary input features for building Transformer-based hybrid models.

4.2. Baseline transformer and realized volatility

Table 4. Forecasting performance of transformer-based Model 1- Model 9

Mode	features	RMSE	QLIKE	R^2	MAE
Model1	transformer(OHLCV)	1.4738	0.4369	-0.7347	0.4369
Model2	model1+RV	1.1734	-0.0997	0.3327	0.8758

DOI: 10.54254/2754-1169/2026.38470

Table 4. (continued)

Model3	mdoel2+VIX	1.1789	0.3343	-0.1099	0.8862
Model4	model3+EGARCH	1.3959	0.3863	-0.5562	1.1120
Model5	mdoel3+S&P500+DXY	1.2804	-0.3092	0.3707	1.0052
Model6	mdoel4+sentiment	1.1572	0.3318	-0.0695	0.8334
Model7	model2+EGARCH	1.1191	0.2975	-0.0001	0.8063
Model8	model7+sentiment	1.2018	0.3242	-0.1534	0.8702
Model9	model5+EGARCH	1.1132	0.3029	0.0104	0.7399

Note. Negative R^2 values appear for several specifications. In volatility-forecasting contexts, negative R^2 indicates model predictions perform worse than a simple historical-mean benchmark. Negative QLIKE values are permissible given the mathematical properties of this loss function and do not imply computational errors.

Model 1 serves as our pure-OHLCV Transformer baseline, supplying reference benchmark results for all subsequent feature-augmented model variants. Adding lagged realized-volatility inputs inside Model 2 brings RMSE down from 1.4738 to 1.1734 and lifts R^2 from −0.7347 to 0.3327. This observation confirms that lagged realized-volatility time-series contain meaningful predictive information for future Bitcoin volatility. MAE simultaneously rises from 0.4369 up to 0.8758. Improvements therefore do not manifest uniformly across every single loss metric. Realized volatility is retained as a core predictor in all follow-up experimental runs.

Model 3 introduces VIX alongside OHLCV and RV inputs. RMSE barely shifts, while R^2 deteriorates. Under this specific configuration, VIX supplies no consistent incremental forecasting value. Model 5 further adds S&P 500 and DXY macro-market indicators. Results make it clear: adding extra external market variables will not mechanically deliver better forecast quality. Marginal predictive contributions from new predictors are highly dependent on what other features already sit inside the input-feature pool. This observation motivates our hybrid-model tests inserting EGARCH-forecast outputs under several distinct feature environments.

4.3. Construction and evaluation of the EGARCH–transformer hybrid model

Our hybrid-model specifications take EGARCH conditional-variance forecasts and treat those as supplementary input features fed into Transformer encoders. The core empirical question becomes: can econometric volatility-model outputs deliver complementary predictive signals for nonlinear temporal-pattern learning carried out by Transformer networks? We test EGARCH-forecast inputs embedded within multiple alternative feature sets.

4.3.1. Model 4: Model 3 plus EGARCH forecasts

Adding EGARCH outputs on top of the feature pool of Model 3 degrades multiple readings of the evaluation metrics. RMSE climbs from 1.1789 (Model 3) up to 1.3959 (Model 4). This result indicates that, not surprisingly, output from EGARCH models in general will not enhance the forecasting performance of Transformer models.

4.3.2. Model 7: EGARCH forecasts versus Model 2 forecasts

Model 7 is a model that removes the VIX inputs from the feature space and only shows EGARCH's contribution. RMSE drops from 1.1734 to 1.1191, and MAE decreases from 0.8758 down to 0.8063.

DOI: 10.54254/2754-1169/2026.38470

Comparing Model 4 against Model 7 delivers one key take-away: how useful EGARCH volatility-forecast inputs turn out to be critically depends on surrounding input-feature composition.

4.3.3. Model 9: Model 5 plus EGARCH forecasts

Model 9 acts as a combination of the EGARCH forecasts and RV, VIX, S&P 500, and DXY time series. It achieves RMSE = 1.1132 and MAE = 0.7399. These are the best error-based performance numbers among all the tested model set-ups.

Combined, the experimental results in the hybrid models demonstrate that EGARCH volatility signals can provide valuable supplementary information to the predictions in Transformer workflows. However, that performance improvement is dependent on the feature-set design and is not guaranteed for arbitrary sets of features provided as input. Hybrid architectures can provide useful additional value, but there are still a few important considerations when choosing which features to pool.

4.4. Sentiment-factor experiments and model-dependent performance

In our final set of ablation experiments, we find that even after having included data on market volatility and macro-financial characteristics, Reddit investor-sentiment data adds additional predictive power. The following two sentiment-augmented specifications are tested: Model 6 and Model 8.

The addition of sentiment variables as inputs into Model 4 (creating Model 6) results in a decrease in both RMSE (from 1.3959 to 1.1572) and MAE (from 1.1120 to 0.8334). In this specific feature-pool environment, sentiment is proven to deliver measurable performance improvement.

The results vary with Model 8, which has features built on Model 7 with sentiment added. Forecast quality worsens: RMSE increases from 1.1191 to 1.2018; MAE rises from 0.8063 to 0.8702; and the correlation R^2 drops to −0.1534.

The simple hypothesis of always using sentiment inputs in the forecasting of Bitcoin-volatility is thus rejected by empirical evidence. Sentiment has a varying predictive role depending on overall-model architecture and surrounding-feature configuration.

The most obvious one is simply because of intrinsic properties of the sentiment measure that we derived from Reddit. Reddit posts and comments, on the other hand, represent subjective retail-investor viewpoints, clearly distinct from more structured financial time-series (RV, VIX, equity-index returns, EGARCH forecasts). There are mixed and even contradictory interpretations of the same market events in forum discussions. They are also accompanied by short-run speculative expectations, individual cognitive biases, and a lot of noise, which is not always related to realized market volatility in the real world.

After Transformer is trained on the structured volatility patterns it has identified from RV, macro-market indices and EGARCH-model outputs, the aggregated Reddit-sentiment readings might not provide the necessary new predictive signals. In some feature-pool setups, sentiment can provide some small complementary data. In other experimental configurations, the subjectivity of its information and noise in embedded features pollute already learned useful feature representations. This offers a plausible route to understanding the variation in sentiment performance by model specification. We do not claim that our experimental setup can provide conclusive causal evidence for this proposed mechanism; this reading remains a possible empirical interpretation.

DOI: 10.54254/2754-1169/2026.38470

5. Discussion

This section discusses our empirical results and compares observations with previous published work.

First, our findings show that structured volatility measures, such as lagged realized volatility, plus well-conditioned EGARCH-model outputs, are indeed high-value predictors of Bitcoin volatility. This is consistent with previous hybrid-model studies [10, 16]. Those papers consider that the econometric-model volatility outputs contain information on volatility-persistence and leverage-effect, which can be used to improve deep-learning forecasting pipelines. Our work adds nuance to that conclusion: performance gains do not automatically hold for arbitrary feature-pool combinations. Simply concatenating EGARCH-forecast outputs together with random macro-variables can degrade prediction quality, as we see with Model 4.

Second, we document unstable forecasting effects coming from Reddit-source investor sentiment. This partially diverges from Biabdillah et al. [6], whose sentiment-aware Transformer finds sentiment brings forecasting improvements for crypto-volatility prediction. Divergent results can originate from multiple sources: differing text-data platforms, alternative sentiment-computing workflows, distinct sample-time windows, and different baseline-feature pools. Many existing sentiment-augmented crypto-volatility papers do not run systematic ablation tests toggling sentiment on and off under multiple competing feature environments. They cannot easily capture this conditional behaviour we observe. Our controlled-ablation experiments demonstrate that sentiment-related performance improvements cannot be generalized across settings. Social-media textual data carries potentially useful behavioural signals alongside substantial embedded noise.

Third, our work shows adding more input features will not guarantee better forecasting outcomes. Including VIX, S&P 500 or DXY does not universally lower forecast-error metrics. Model 9 achieves the best RMSE and MAE readings. Its strength comes from carefully selected complementary structured predictors rather than maximizing total feature quantity. In Bitcoin-volatility forecasting tasks, the quality of the raw data and feature complementarity become more important than the complexity of the model and the number of features.

Researchers and practitioners in the digital-assets risk management and digital-asset derivatives-modelling world should not take for granted that sentiment from social media will always improve the volatility outputs. When behavioural proxies are added to forecasting pipelines, controlled ablation testing is highly recommended to establish marginal improvements in prediction in that specific feature-pool environment.

6. Conclusion, limitations and future research

6.1. Main conclusions

This paper compares out-of-sample volatility-forecasting performance of EGARCH standalone, vanilla Transformer, and hybrid EGARCH-Transformer models for Bitcoin. We also create a FinBERT sentiment index based on Reddit upvotes and conduct controlled-ablation tests to see if there are any incremental gains from sentiment inputs from investors. There are some important empirical findings.

The historical structured volatility information is very useful for Bitcoin-volatility prediction. The use of lagged realized-volatility inputs leads to a significant improvement in the Transformer-model RMSE and R^2performance, but not necessarily the other evaluation metrics.

DOI: 10.54254/2754-1169/2026.38470

The addition of additional exogenous predictor variables does not mechanically increase the precision of forecasting. The pre-existing feature-pool composition is crucial to the marginal influences of VIX etc. macro-market indicators. It is feature-set selection that is important rather than model complexity.

EGARCH conditional variance forecasts can provide additional predictive information for Transformer encoders. That advantage is feature-environment dependent and can not be taken for granted for any possible model configuration. For all the tested specifications, our best RMSE and MAE are achieved by Model 9, which are 1.1132 and 0.7399, respectively. It comprises Transformer-based pattern learning, realised volatility readings, multi-market macro-financial indicators, and EGARCH-model volatility outputs.

Forecasting benefits from Reddit-based investor sentiment, which is volatile. Sentiment boosts the output of one feature pool model (Model 6), but decreases the prediction accuracy of another model (Model 8). Sentiment from social media is full of potentially useful behavioural information, yet it also has a lot of subjective noise. It is very sensitive with respect to the set of input features around it.

In total, the predictive power of Bitcoin-volatility forecasting depends on whether the various inputs really provide complementary signals. There is a need for more than simply stacking additional variables into Transformer-style deep-learning architectures. Structured volatility-based metrics provide much more consistent performance gains. The effectiveness of hybrid-models is influenced by network architecture, input-feature composition, and raw-source-data quality. New candidate predictor variables must be empirically validated, not assumed to add forecasting value to the model.

6.2. Limitations

There are a few significant practical constraints to this study.

(i)All sentiment signals were built with the r/Bitcoin Reddit community and with our own model, ProsusAI/finbert. Outcomes could differ substantially if we draw text data from alternative sources, such as X (Twitter) or mainstream-financial-news articles.

(ii)All tests are completed on daily-frequency data. Empirical patterns may vary if sentiment and volatility are measured using intraday, high-frequency data.

(iii)Analysis is only on Bitcoin. No other cryptocurrencies based on the concept of an alt-coin have been tested. The extent to which these conditional sentiment-related effects exist for other crypto-assets has not yet been tested.

(iv)Sentiment weighting is based only on the number of upvotes for a post and its comments. Other weighting methods (thread-engagement metrics, reply volume) could yield varied results for sentiment-index scores.

(v)We do not sample all of the Reddit text corpus; rather, we sample only the text corpus from the top-voted forum posts. There might be some sampling bias because of this targeted sampling.

(vi)All cross-model forecast comparisons are based purely on numeric values for RMSE, MAE, R^2 and QLIKE. Due to computational overhead, Diebold-Mariano statistical testing was not completed. So, it is not possible to formally test whether the performance differences that are observed between competing models are statistically significant.

6.3. Future research

There are a number of follow-on directions that are worth following.

DOI: 10.54254/2754-1169/2026.38470

Firstly, multi-source sentiment indicators, built on the basis of information from news-media, multiple social-media platforms, and forum-thread text could be integrated within hybrid forecasting pipelines. This would enable cross-validation of the conditional predictive value of sentiment. Second, future research might enable the separation of bullish-sentiment and bearish-sentiment components (rather than the aggregate sentiment index). Researchers can then examine if the sentiment represented by positive-text versus negative-text leads to asymmetric forecasting contributions for cryptocurrency volatility. Third, sub-sample analysis might be used to examine model performance in periods of high volatility in crisis situations compared to normal markets. This would explore if sentiment-related signals become more informative during market-turmoil periods. Fourth, other hybrid architectures that fuse the outputs of the GARCH-family with other time-series deep-learning architectures (e.g., TFT) would be interesting to explore for digital-asset volatility-forecasting tasks.

References

[1] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327. doi: 10.1016/0304-4076(86)90063-1.

[2] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987–1007.

[3] Wang, J., Ma, F., Bouri, E., & Guo, Y. (2022). Which factors drive Bitcoin volatility: Macroeconomic, technical, or both? *Journal of Forecasting, 42*(4). doi: 10.1002/for.2930.

[4] Zahid, M., Iqbal, F., & Koutmos, D. (2022). Forecasting Bitcoin volatility using hybrid GARCH models with machine learning. *Risks, 10*(12), 237. doi: 10.3390/risks10120237.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. [online] *Neural Information Processing Systems*. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [Accessed 26 Aug. 2026].

[6] Biabdillah, F., Triadi, F., Go, A. J. G., & Ramadhan, M. C. P. (2025). Sentiment-aware transformer for cryptocurrency volatility prediction using multi-source market and news sentiment. *Journal of Embedded Systems, Security and Intelligent Systems*, 701–714. doi: 10.59562/jessi.v6i4.10604.

[7] Herremans, D., & Low, K. W. (2025). Forecasting Bitcoin volatility spikes from whale transactions and CryptoQuant data using synthesizer transformer models. *IEEE Access, 13*, 117788–117807. doi: 10.1109/access.2025.3584243.

[8] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing, 90*, 106181. doi: 10.1016/j.asoc.2020.106181.

[9] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(12), 11106–11115. doi: 10.1609/aaai.v35i12.17325.

[10] Gebreslase, T. T., Mutua, K., Fissuh, Y. H., & Girmay, T. G. (2026). A hybrid multi-GARCH transformer model for stock price volatility forecasting. *Commun. Math. Biol. Neurosci., 2026*(0). doi: 10.28919/cmbn/9834.

[11] Ahadzie, R. M., Junior, P. O., & Woode, J. K. (2024). The impact of sentiment on realized higher-order moments in the S& P 500: Evidence from the fear and greed index. *Journal of Risk and Financial Management, 18*(1), 2–2. https://doi.org/10.3390/jrfm18010002

[12] Bandopadhyaya, A., & Jones, A. L. (2011). Measures of investor sentiment: A comparative analysis put-call ratio vs. volatility index. *Journal of Business & Economics Research (JBER), 6*(8). https://doi.org/10.19030/jber.v6i8.2458

[13] Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

[14] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

[15] Gyamerah, S. A. (2019). Modelling the volatility of Bitcoin returns using GARCH models. *Quantitative Finance and Economics, 3*(4), 739–753.

DOI: 10.54254/2754-1169/2026.38470

[16] Chinazzo, C., & Jeleskovic, V. (2024). Forecasting Bitcoin volatility: A comparative analysis of volatility approaches. *arXiv.org*. https://arxiv.org/abs/2401.02049
[17] Pichl, L., & Kaizoji, T. (2017). Volatility analysis of Bitcoin price time series. *Quantitative Finance and Economics, 1*(4), 474–485. https://doi.org/10.3934/qfe.2017.4.474
[18] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting, 34*(4), 802–808. https://doi.org/10.1016/j.ijforecast.2018.06.001
[19] Kristjanpoller, W., & Minutolo, M. C. (2018). A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis. *Expert Systems with Applications, 109*, 1–11. https://doi.org/10.1016/j.eswa.2018.05.011