ECU33143 Introduction to Big Data in Economics Research Report

Raj Aryan Upadhyaya

Content:

Introduction

Since 2016, cryptocurrencies have gained significant popularity and have rapidly evolved from niche digital tokens into assets with growing integration in modern technology. Their applications now extend beyond encrypted payment systems to broader uses such as blockchain infrastructure, decentralized finance (DeFi), and non-fungible tokens (NFTs). While much of this innovation has been technological, cryptocurrencies have also increasingly attracted attention as financial assets.

This research investigates the pricing behaviour of cryptocurrencies, focusing on the causal relationship between cryptocurrency returns, macroeconomic indicators, and social media sentiment. The aim is to determine whether cryptocurrencies continue to be primarily influenced by online sentiment or whether they have become more responsive to macroeconomic fundamentals, potentially reflecting a transition toward behaving like traditional financial assets.

To capture a representative view of the market, the study analyses ten major cryptocurrencies, including Bitcoin and Ethereum, against the U.S. dollar using both hourly (for macro) and daily (altered for sentiment) price data. As the analysis is conducted entirely from a U.S. perspective, macroeconomic variables are obtained via the Federal Reserve Economic Data (FRED) API, while sentiment measures are derived from U.S.-based social media platforms, primarily Reddit (While Twitter/X was initially considered for sentiment analysis, API accessibility constraints necessitated exclusive use of Reddit-based data). Together, these datasets enable an examination of the short-term causal dynamics between macroeconomic and sentiment factors in cryptocurrency markets.

Literature Review:

Alongside my research project, I have recently concluded a Literature Review (for the main reason of understanding what methods to look at, and why they would be useful), where I took 2 papers from each of the following topics (all sourced from Google Scholar search).

1) Cryptocurrency market efficiency
    a. Sahoo and Sethi (2022)
    b. Souza and Carvalho (2022)

    Findings:

    - Both papers found cryptocurrencies to be informationally efficient (i.e. the efficiency of the cryptocurrency market depended on the level of information present for the price to move), with some exceptions (as we see with Sahoo & Sethi).
    - Souza and Carvalho (2022) validates the need for utilising higher-frequency data in order to explain the cause of more abrupt shifts, as well as emphasising on looking at this from a multivariate perspective, as cryptocurrency exchanges show causality when compared amongst one another in the Granger sense and also allows to gain a broad understanding of the inherent risks associated with cryptocurrencies.
    - Sahoo & Sethi (2022) found exceptions for information efficiency in Ripple (one of my tested cryptocurrencies), DASH, and XMR, but the majority follow an efficient market hypothesis.

2) Cryptocurrency links to Macroeconomic Data
    a. Sakemoto and Nakagawa (2021)
    b. Baranovskyi et al (2021)

    Findings:

    - Both papers found a correlation between Macro and Cryptocurrencies.
    - Sakemoto and Nakagawa (2021) found strong correlations through the use of a factor model (using Output, Labour, housing, bonds prices, amongst other factors) previously discussed in S&W, where Inflation and Money supply don't tend to affect it too much. Primarily used regression to test their theory.
    - Baranovskyi et al (2021) found there is correlation, and that multi-regression paired with correlation analysis has shown the market's impact on modern investment vehicles such as cryptocurrencies.
      Their main methods used were multiple regression as well as machine learning techniques.

3) Cryptocurrency links to Social Media via Sentiment Analysis
    a. Colianni et al (2015)
    b. Wooley et al (2019)

    Findings:

    - Both tended to find a link between cryptocurrency price shifts with regards to sentiment analysis.
    - Colianni et al (2015) looked at this from a sentiment perspective on Twitter, where they found similar, well-correlated results, and had looked at Logistic Regression, with Bernoulli Naïve Bayes methods.
    - Wooley et al (2019) were looking at this from a sentiment perspective on Reddit, where they had looked at sentiment time series, amongst other forms of time series analysis in order to highlight where there is correlation, which was proven via Reddit. They had used VADER[21] as their Sentiment model, and had also utilised a rolling z-score for analysis, GARCH for volatility, and a Self-Organising Fuzzy Neural Network to look at non-linear complex relationships.

4) Existing work that combines both Cryptocurrency-Macro and Cryptocurrency-Sentiment analyses.
    a. Kabo et al (2025)
    b. Frendo (2025)

    Findings:

    - Both of these papers found that there was a need to further explore studies of this kind, and both of these papers approached the area in different ways, which carried out their own results. Both of these papers being new also indicates that there is indeed progression being made on amalgamating both cryptocurrency-macro and cryptocurrency-sentiment in analysis of cryptocurrencies.
    - Kabo et al (2025) were looking purely from a predictive standpoint, investigated primarily on bitcoin, and only tracked GDP, but otherwise showed that there is indeed feasibility for my work, as they had built a model by integrating ARIMA and LSTM, which had improved by combining both sentiment and GDP into their model.
    - Frendo (2025) looked at this primarily from the perspective of cryptocurrency-macro, but had integrated cryptocurrency-sentiment, allowing them to find meaningful insight into cryptocurrency price movement, but had suggested to consider regulations for sentiment, as well as more robust methodology in order to fully capture cryptocurrency price movements.

Overall, I found that a lot of the studies had utilised Granger Causality, allowing us to use that to pinpoint the relationships between the different factors involved, and similarly, had given insight towards how to better structure this project for methodology going forward in ways I may not have considered before (such as a self-organising Fuzzy Neural Network, which I did not use, but is indeed quite interesting nonetheless).

Project Implementation

My project highlights Cryptocurrency Pricing relationships between both Macroeconomic Data, as well as Sentiment Data, thus my implementation has been split between both.

Crypto-Macro:

Data:

- Hourly historical cryptocurrency pricing data for 10 major cryptocurrencies (including Bitcoin, Ripple, Ethereum, and DOGE) were obtained through the Binance API. Macroeconomic indicators were extracted using the FRED API and Python's yfinance library, including the Consumer Price Index (CPI), Federal Funds Rate, M2 Money Supply, 10-year Treasury bond yields, crude oil prices, and various equity market indexes.

Data Cleaning:

- The cryptocurrency data required minimal preprocessing as the Binance API provided clean, structured time-series data at hourly intervals.
- For macroeconomic variables, which are typically updated monthly, forward-fill interpolation was applied to maintain consistent values within each reporting period, ensuring alignment with the higher-frequency cryptocurrency data.

Model:

- **Initial OLS Regression Analysis:** Static linear regression models were initially implemented to examine relationships between cryptocurrency prices and macroeconomic indicators. However, diagnostic testing revealed fundamental violations of OLS assumptions. The models exhibited R^2 values ranging from 0.63 (AVA) to 0.96 (BTC), yet concerning diagnostic statistics emerged: Durbin-Watson statistics averaged below 0.1, and condition numbers exceeded 1.5×10^6. These diagnostics indicated severe autocorrelation and multicollinearity issues, rendering static OLS regression inappropriate for time-series data with non-stationary characteristics.
- **Vector Autoregression (VAR) Model:** Given the time-series nature of the data and evidence of changing relationships over time, a VAR model was subsequently employed. This approach explicitly accounts for temporal dependencies and allows for dynamic interactions between variables, making it well-suited for capturing the evolving relationships between cryptocurrency markets and macroeconomic conditions. The VAR model results are presented in Figures 1-10.

Interpretation:

- The analysis revealed heterogeneous integration patterns between cryptocurrencies and macroeconomic variables. Approximately half of the examined relationships demonstrated significant linkages, most notably with the USD Index, S&P 500, VIX, and gold prices. The M2 Money Supply (liquidity measure) exhibited mixed relationships with cryptocurrency prices.
- Overall, the results suggest that while cryptocurrencies have progressively integrated as risk assets since 2016, showing correlations with traditional macroeconomic indicators, the cryptocurrency-macro linkage remains in development.

Crypto-Sentiment:

Data:

- Crypto Data: Identical to the Crypto-Macro analysis dataset, covering 10 major cryptocurrencies (BTC, ETH, BNB, ADA, XRP, SOL, DOGE, DOT, AVAX, LINK) with daily OHLCV data.
- Sentiment Data: Reddit API (PRAW) was used to extract text data from 25+ cryptocurrency-focused subreddits for sentiment analysis.

Cleaning:

- Crypto Price Data: No additional cleaning required, as any necessary preprocessing was completed during the Crypto-Macro analysis phase.
- Sentiment Data: Reddit sentiment scores were aggregated to daily frequency, but presented data sparsity challenges due to irregular posting patterns across subreddits. Missing sentiment values were forward-filled to align with the continuous daily price data, with zero-sentiment days replaced by the most recent non-zero sentiment score. A 7-day moving average was calculated to smooth short-term noise and capture broader sentiment trends.

Model:

- The analytical framework builds upon the Crypto-Macro methodology, employing multiple complementary techniques:
  - Granger Causality Tests: Bidirectional tests (sentiment → returns and returns → sentiment) with AIC-based lag selection (optimal lag: 10 days across all assets) to assess temporal precedence and predictive relationships.
  - Value-at-Risk (VaR) Models: Two approaches were compared at 95% and 99% confidence levels:
  - Historical VaR: Rolling 250-day window using empirical return distributions
  - Conditional VaR: Quantile regression incorporating sentiment features (raw sentiment and 7-day MA) to generate sentiment-conditional risk estimates
- Backtesting: Kupiec Proportion-of-Failures test was applied to validate VaR model accuracy by comparing expected versus observed breach rates.
- Correlation Analysis: 30-day rolling correlations between sentiment and returns to capture time-varying relationships and regime shifts.

Interpretation:

Granger Causality Findings: The analysis reveals limited direct causal relationships between Reddit sentiment and cryptocurrency returns. Only 1 out of 10 cryptocurrencies demonstrated statistically significant bidirectional causality ($p < 0.05$), with most p-values exceeding 0.70. This suggests sentiment does not consistently precede price movements at daily frequencies across the broader crypto market.

VaR Model Performance: Despite weak Granger causality, sentiment-conditional VaR models significantly outperformed historical approaches. The Conditional VaR achieved a 100% Kupiec test pass rate across all assets and confidence levels, compared to 0% for Historical VaR. Exception rates for Conditional VaR (5.19% at 95%, 1.16% at 99%) closely matched theoretical expectations, demonstrating superior risk forecasting accuracy when sentiment is incorporated as a conditioning variable.

Asset-Specific Patterns:

- SOLUSDT exhibited a weak positive sentiment-return relationship ($\beta = 0.006$) with high correlation volatility (-0.6 to +0.6), suggesting regime-dependent sensitivity to social media sentiment.
- XRPUSDT showed a slight negative relationship ($\beta = -0.004$) with more stable correlations, indicating sentiment may reflect contrarian indicators or post-event reactions rather than forward-looking signals.

Rolling Correlation Dynamics: Time-varying correlations reveal that sentiment's predictive power is not constant but fluctuates with market regimes. Periods of high volatility (2020-2021, 2023) show stronger correlations, implying sentiment plays a more pronounced role during uncertainty.

Practical Implications: While Reddit sentiment alone is insufficient for directional price prediction, it adds meaningful information to risk models, particularly for tail-risk estimation. The VaR improvements suggest sentiment captures latent market stress or positioning dynamics not reflected in historical returns alone.

Results:

Conclusion:

This research examined the pricing dynamics of ten major cryptocurrencies through macroeconomic fundamentals and social media sentiment analysis.
The central question, Whether cryptocurrencies respond systematically to traditional macroeconomic indicators, or are primarily driven by speculative sentiment, yielded nuanced findings that challenge the binary characterisations of crypto markets.

Macroeconomic Integration is heterogeneous. Vector autoregression analysis and Granger causality tests revealed inconsistent cryptocurrency responsiveness to macro factors.
While certain assets showed statistically significant relationships with specific indicators (as seen in figure 11), some Macro variables showed evidently weak relationships to cryptocurrencies as a whole.
Impulse response functions (Figures 1-10) confirmed temporary but significant effects from macro shocks, with most responses decaying within 2-5 days, suggesting cryptocurrencies occupy a spectrum between macro-integrated assets and speculative instruments rather than forming a homogeneous asset class.

Sentiment Improves Risk Forecasting, not Return Prediction. Reddit sentiment analysis produced a counterintuitive result: only 1 of 10 cryptocurrencies exhibited significant Granger causality between sentiment and returns ($p < 0.05$).

However, when incorporated into Value-at-Risk models through quantile regression and Kupiec Tests, sentiment dramatically improved tail-risk forecasting.
Conditional VaR models achieved 100% Kupiec test pass rates across all assets versus 0% for historical VaR. Exception rates closely matched expectations (5.19% vs 5.00% at 95% confidence; 1.16% vs 1.00% at 99% confidence).

This divergence reveals that sentiment operates as a risk factor rather than a leading indicator. Social media sentiment captures latent market stress and herding tendencies that manifest in tail events rather than mean returns, distinguishing risk conditioning from directional prediction.

Practical Implications:

Regarding risk, these findings justify incorporating sentiment into downside risk frameworks and regulatory capital calculations. The 100% Kupiec pass rate demonstrates sentiment improves VaR models.
However, sentiment should not drive directional trading strategies given its weak causal relationship with Cryptocurrencies.

If one were to apply this in a portfolio strategy, limited crypto-macro integration suggests cryptocurrencies may provide diversification benefits, though this varies substantially across individual coins. The heterogeneous macro sensitivities inform relative value strategies pairing macro-sensitive assets (e.g. BTC) against macro-insensitive ones (e.g. DOGE).

Discussion:

Methodological Limitations: Daily data aggregation may obscure intraday sentiment-price dynamics. Daily aggregation potentially attenuated true causal relationships. The Reddit-only approach, while justified by API constraints, limits platform diversity.
Multi-platform sentiment incorporating Twitter/X, Weibo, etc could improve signals. VADER sentiment analysis, though efficient, misses cryptocurrency-specific linguistic nuances like "to the moon" or "diamond hands" that fine-tuned models (FinBERT, crypto-specific transformers) could capture.

Forward-filling macro data introduces **look-ahead bias** by assuming end-of-period values (monthly CPI, quarterly GDP) were known throughout periods.
While this may be standard practice, this overstates apparent correlations. State-space models or MIDAS regressions would provide theoretically sounder mixed-frequency approaches.

Additionally, the 2016-2025 sample period spans multiple distinct regimes (ICO bubble, pandemic liquidity surge, crypto winter, post-halving recovery), and pooling these assumes stable relationships that rolling correlations suggest is violated. Regime-switching models could reveal strengthening macro integration during risk-off periods or amplified sentiment effects during market manias.

Comparison to Literature: When we look at what the referenced literature, both Sakemoto & Nakagawa (2021), as well as Baranovskyi et al (2021) had agreed that there was Macro integration which we found through our clear linkage of variables for crypto-macro. Crypto-Sentiment research showed good correlation through Reddit (Wooley et al (2019)), however it was evident via figure 13 that this may not necessarily hold now, but we show an alternative, where it is evidently a very good risk tool via figures 14-24, where the VaR's quantile regression evidently seemed apt.

Broader Implications: Weak Granger causality tentatively supports cryptocurrency market efficiency at daily frequencies. If sentiment reliably forecasted returns, arbitrage should eliminate this inefficiency. However, persistent sentiment-risk relationships suggest limits to arbitrage from capital constraints, short-selling limitations, or noise traders influencing short-term volatility. The heterogeneous macro integration challenges "cryptocurrencies as an asset class" narratives.

For example: Bitcoin behaving differently from Dogecoin suggests separate analytical frameworks may be required.

Figures

Figure 1

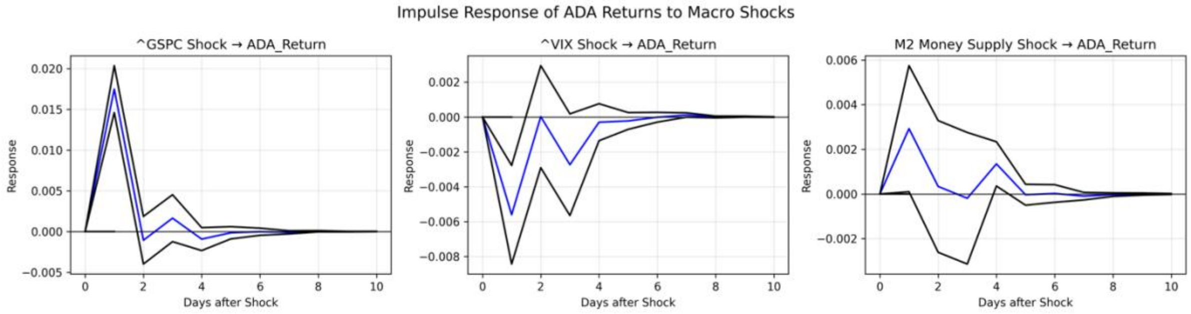


Figure 2

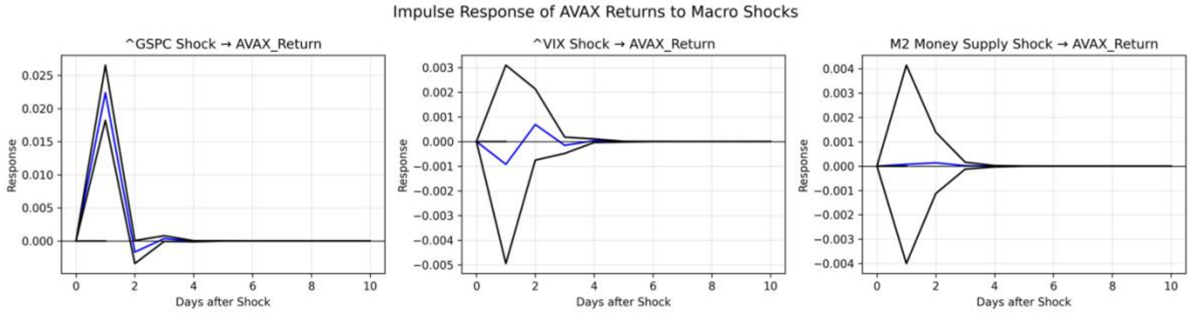


Figure 3

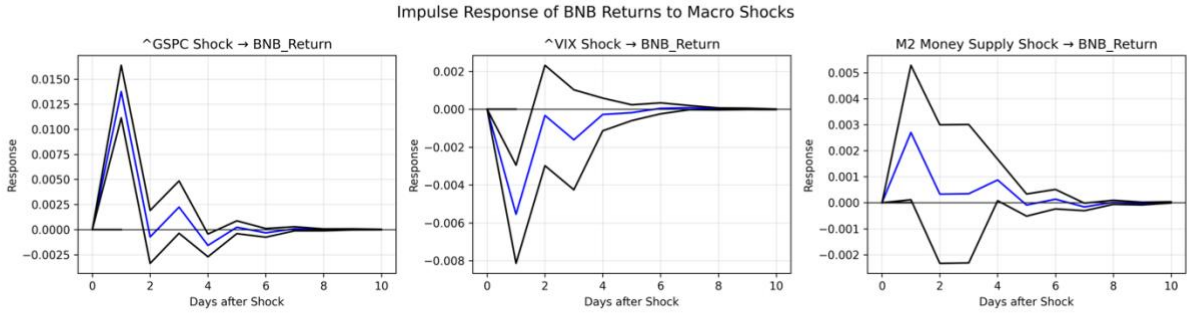


Figure 4

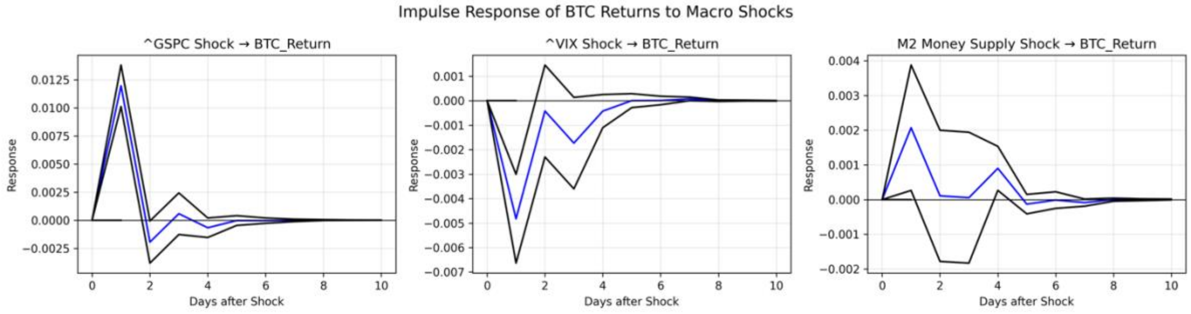


Figure 5

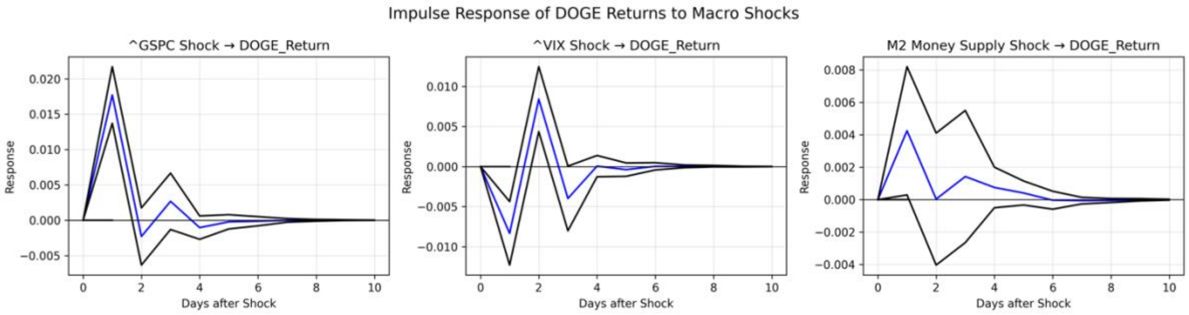


Figure 6

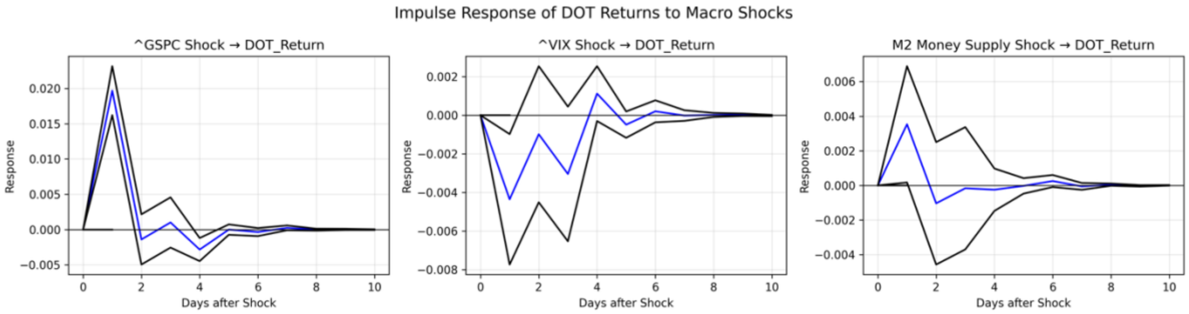


Figure 7

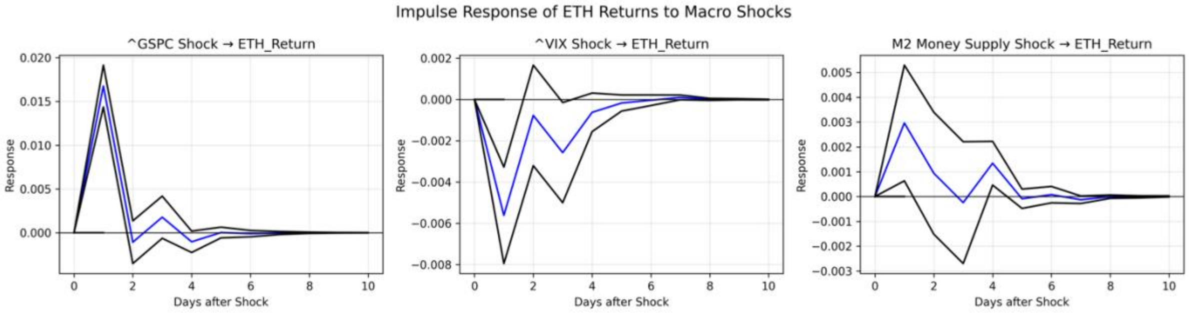


Figure 8

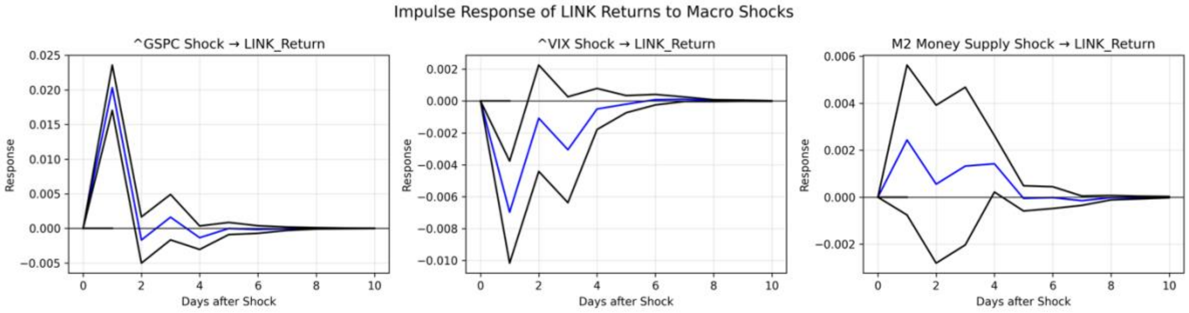


Figure 9

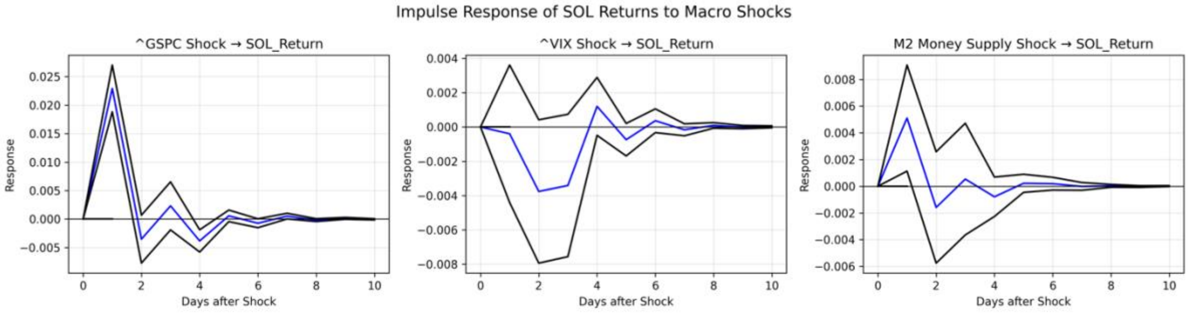


Figure 10

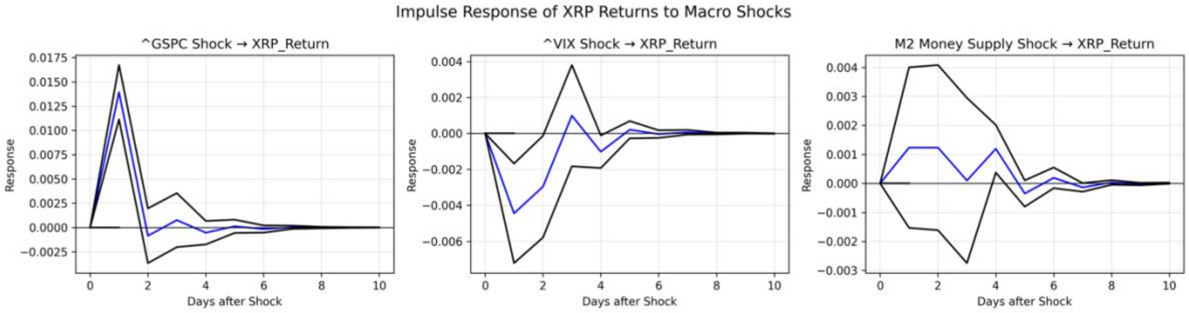
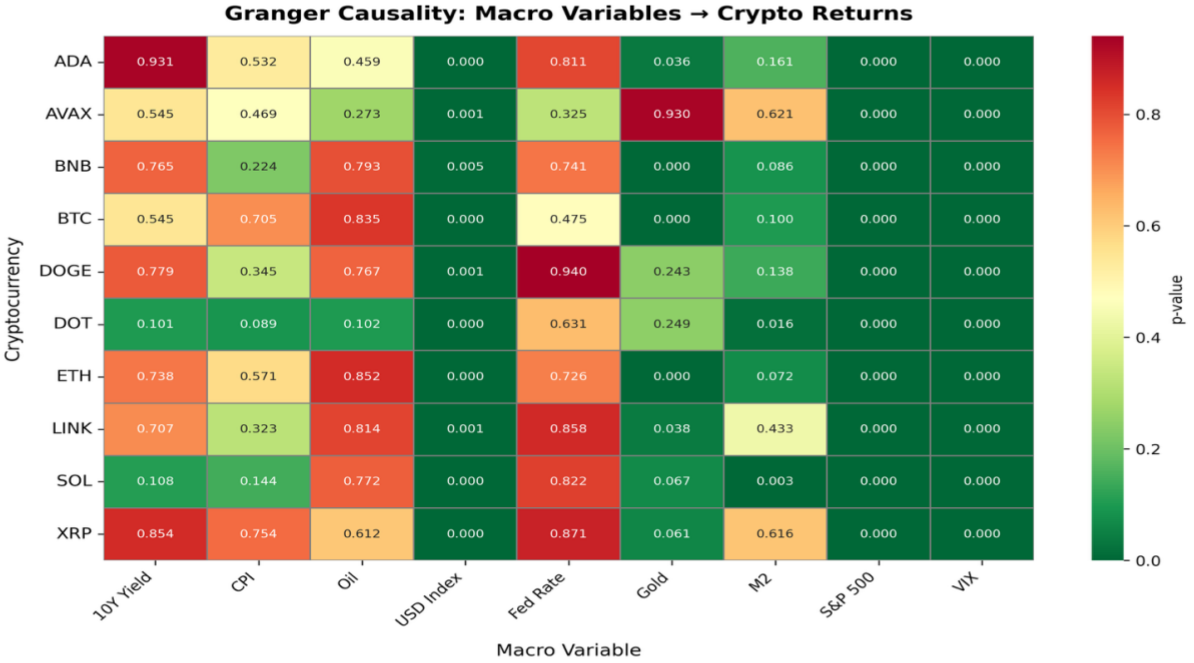
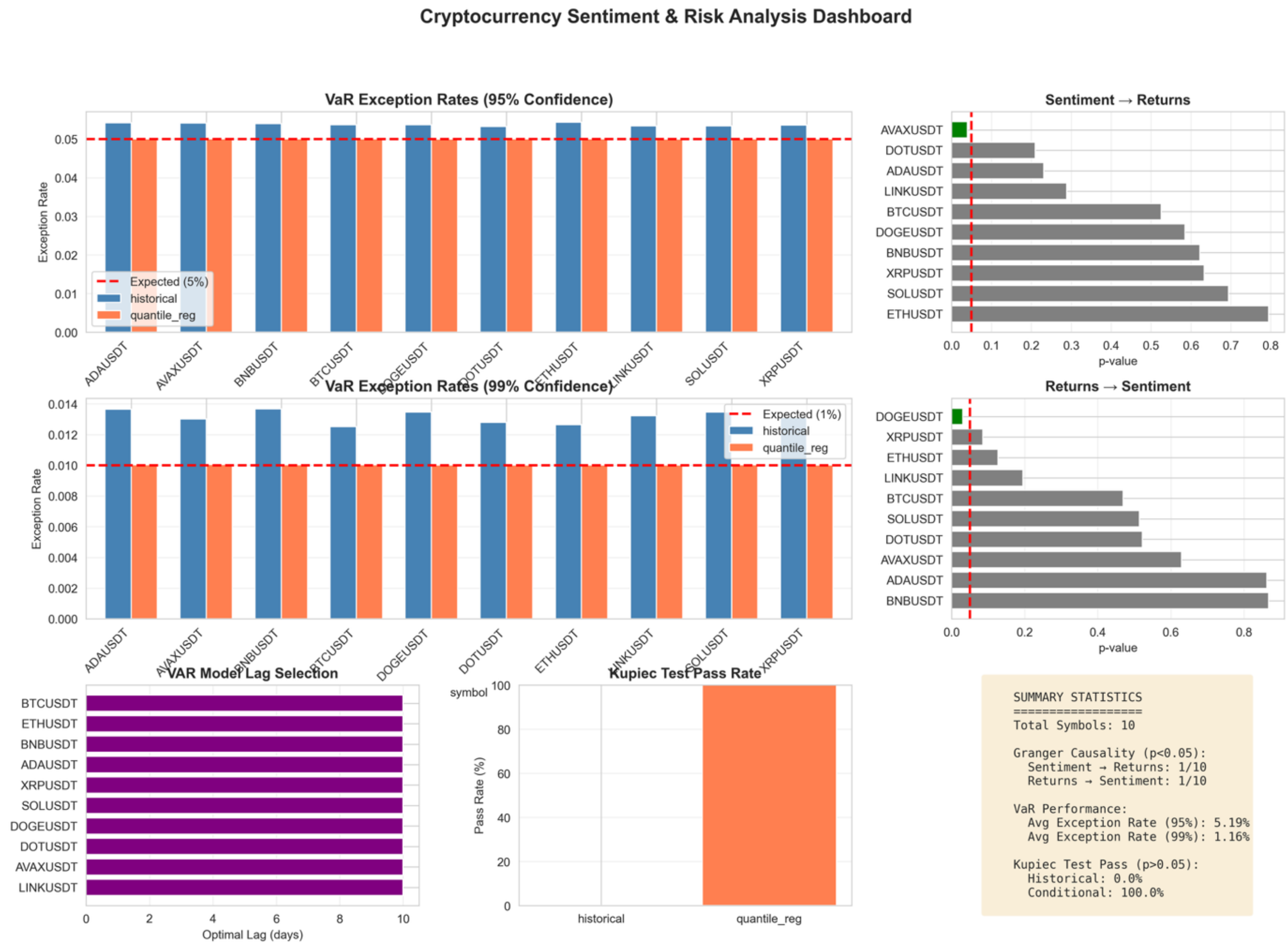


Figure 11



	10Y Yield	CPI	Oil	USD Index	Fed Rate	Gold	M2	S&P 500	VIX
ADA	0.931	0.532	0.459	0.000	0.811	0.036	0.161	0.000	0.000
AVAX	0.545	0.469	0.273	0.001	0.325	0.930	0.621	0.000	0.000
BNB	0.765	0.224	0.793	0.005	0.741	0.000	0.086	0.000	0.000
BTC	0.545	0.705	0.835	0.000	0.475	0.000	0.100	0.000	0.000
DOGE	0.779	0.345	0.767	0.001	0.940	0.243	0.138	0.000	0.000
DOT	0.101	0.089	0.102	0.000	0.631	0.249	0.016	0.000	0.000
ETH	0.738	0.571	0.852	0.000	0.726	0.000	0.072	0.000	0.000
LINK	0.707	0.323	0.814	0.001	0.858	0.038	0.433	0.000	0.000
SOL	0.108	0.144	0.772	0.000	0.822	0.067	0.003	0.000	0.000
XRP	0.854	0.754	0.612	0.000	0.871	0.061	0.616	0.000	0.000

Figure 12



```
SUMMARY STATISTICS
==================
Total Symbols: 10

Granger Causality (p<0.05):
  Sentiment → Returns: 1/10
  Returns → Sentiment: 1/10

VaR Performance:
  Avg Exception Rate (95%): 5.19%
  Avg Exception Rate (99%): 1.16%

Kupiec Test Pass (p>0.05):
  Historical: 0.0%
  Conditional: 100.0%
```

Figure 13

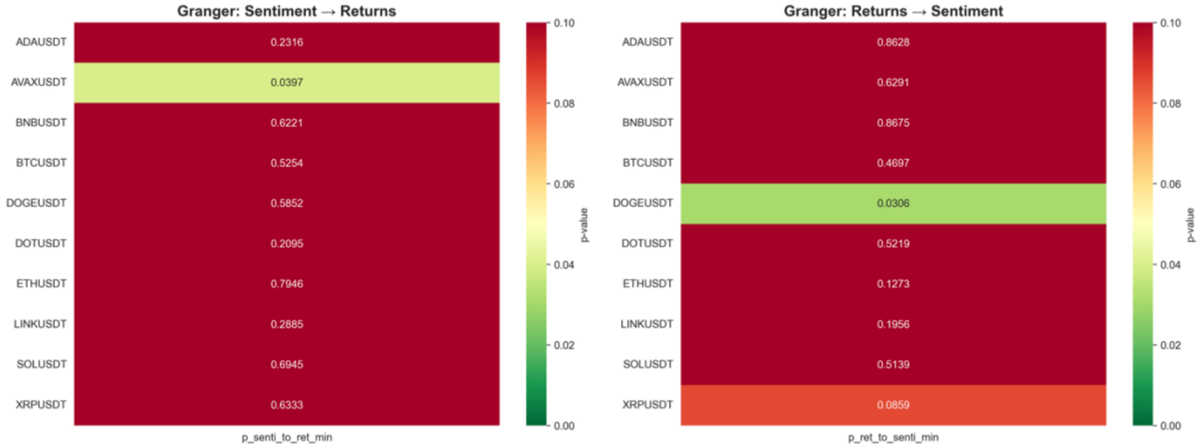


Figure 14

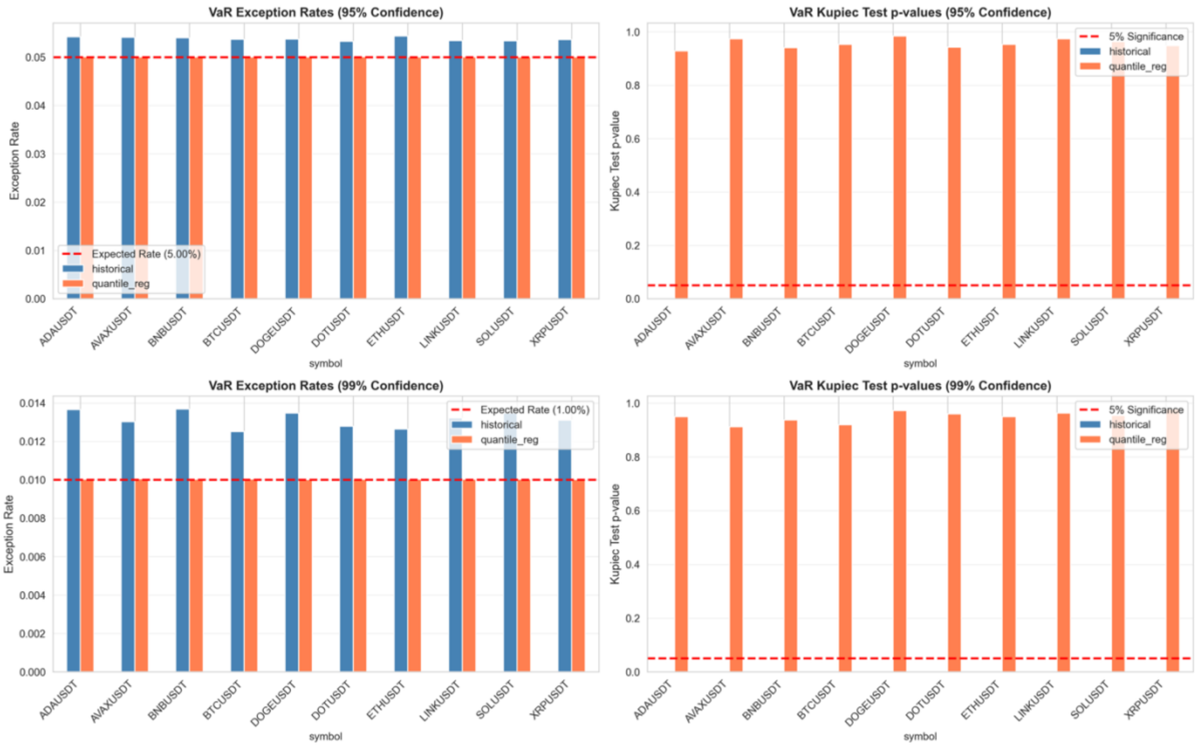


Figure 15

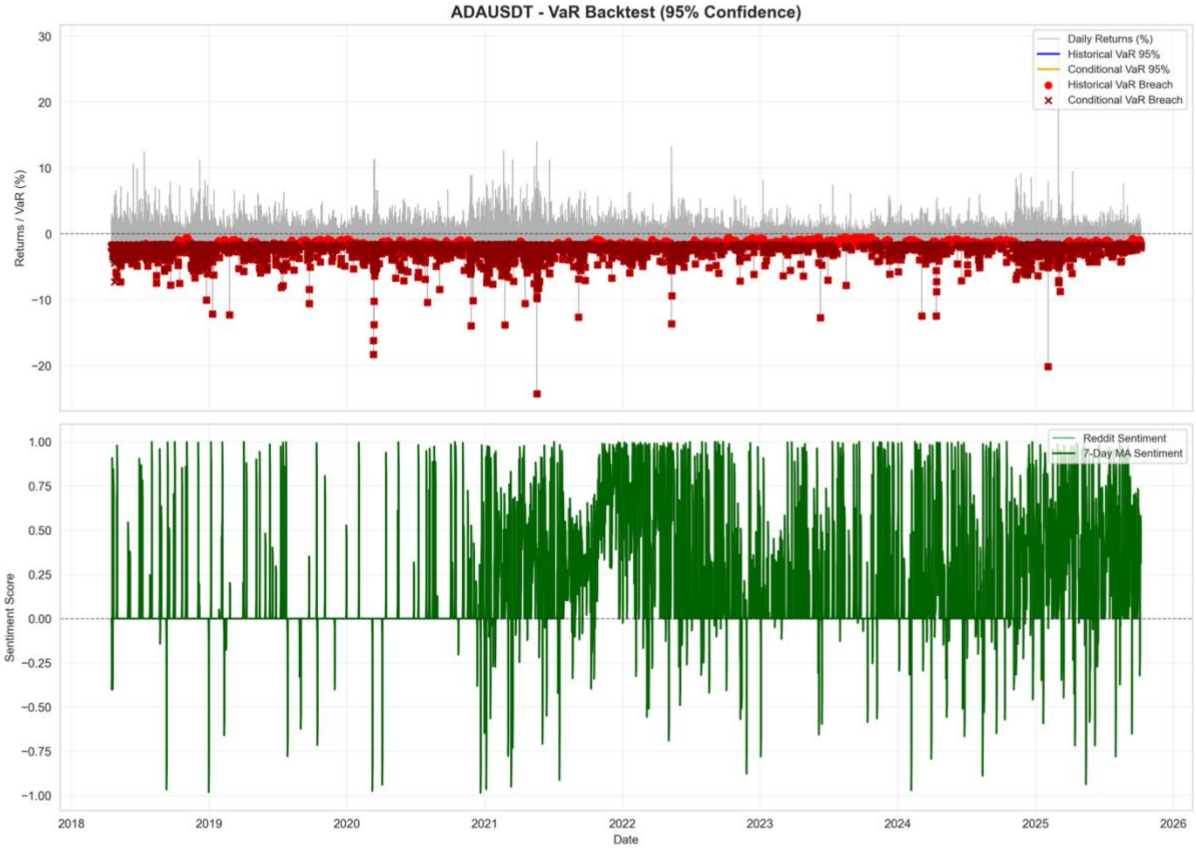


Figure 16

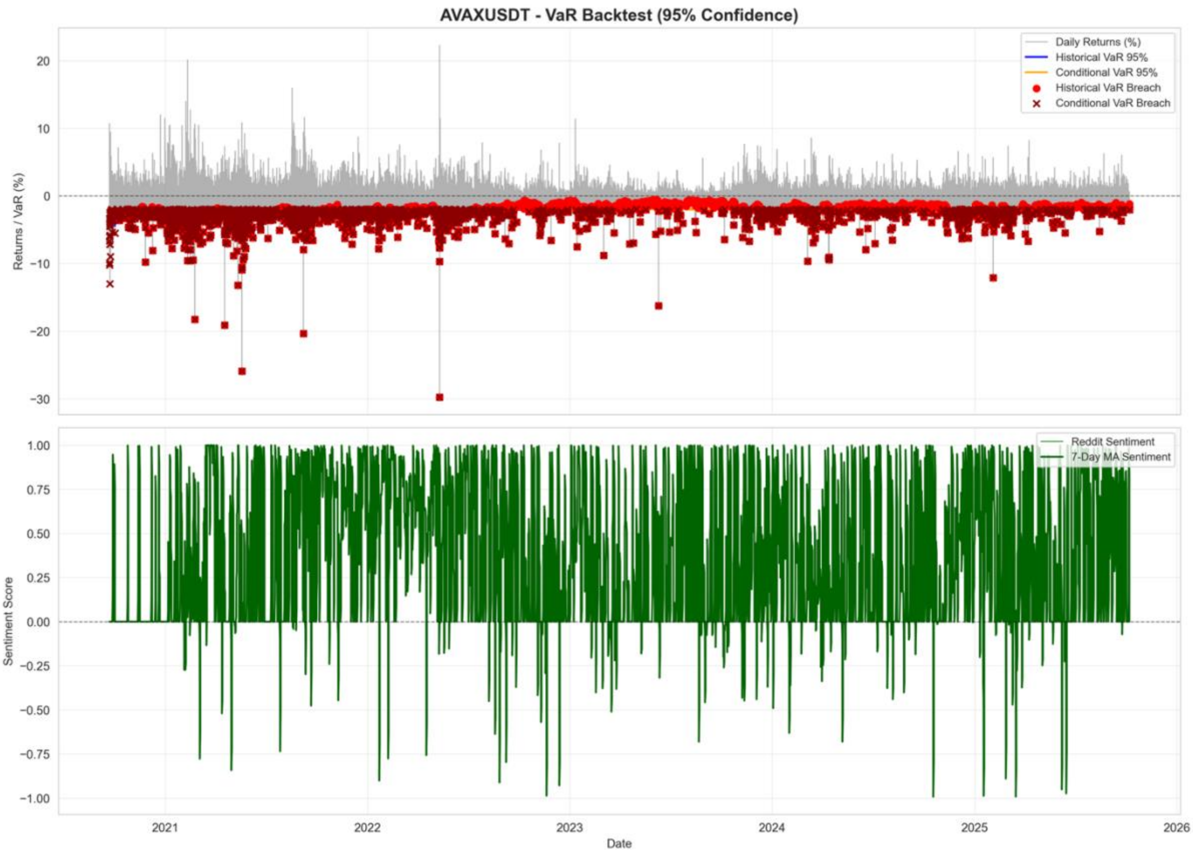


Figure 17

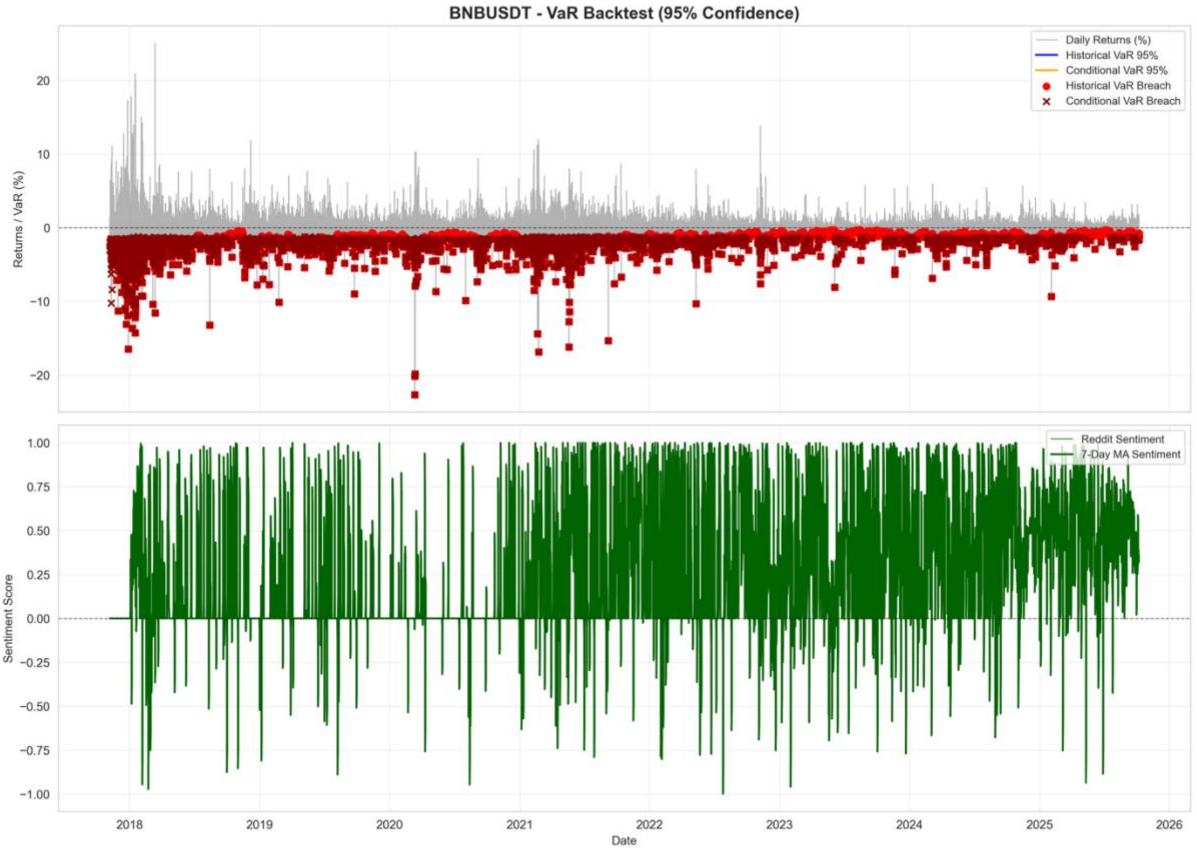


Figure 18

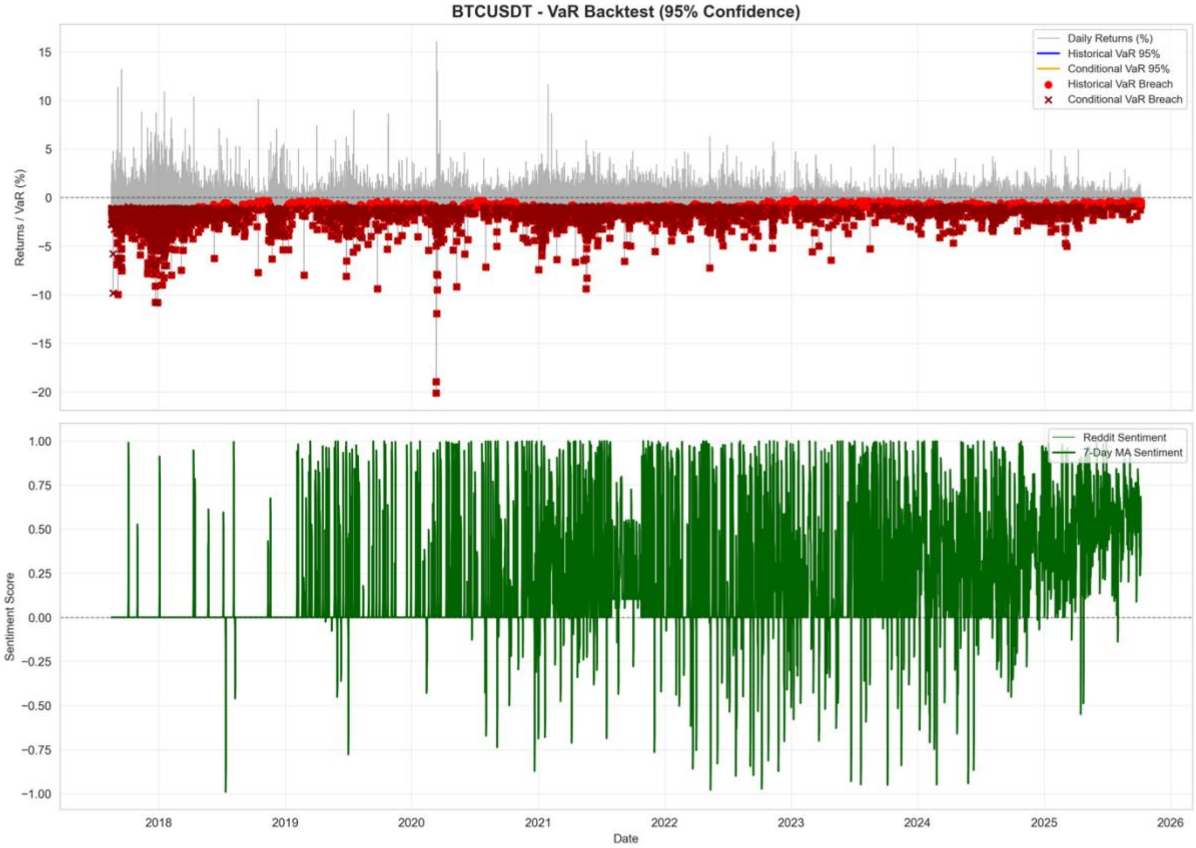


Figure 19

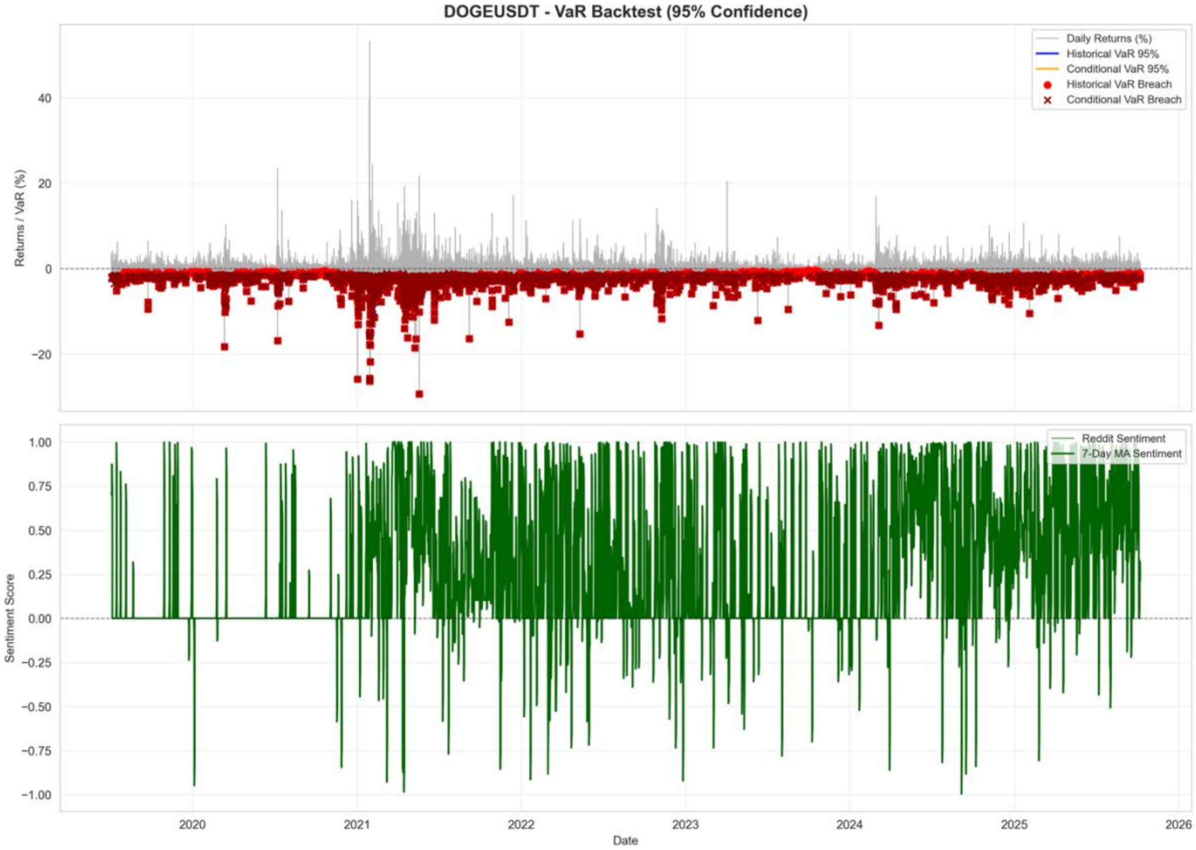


Figure 20

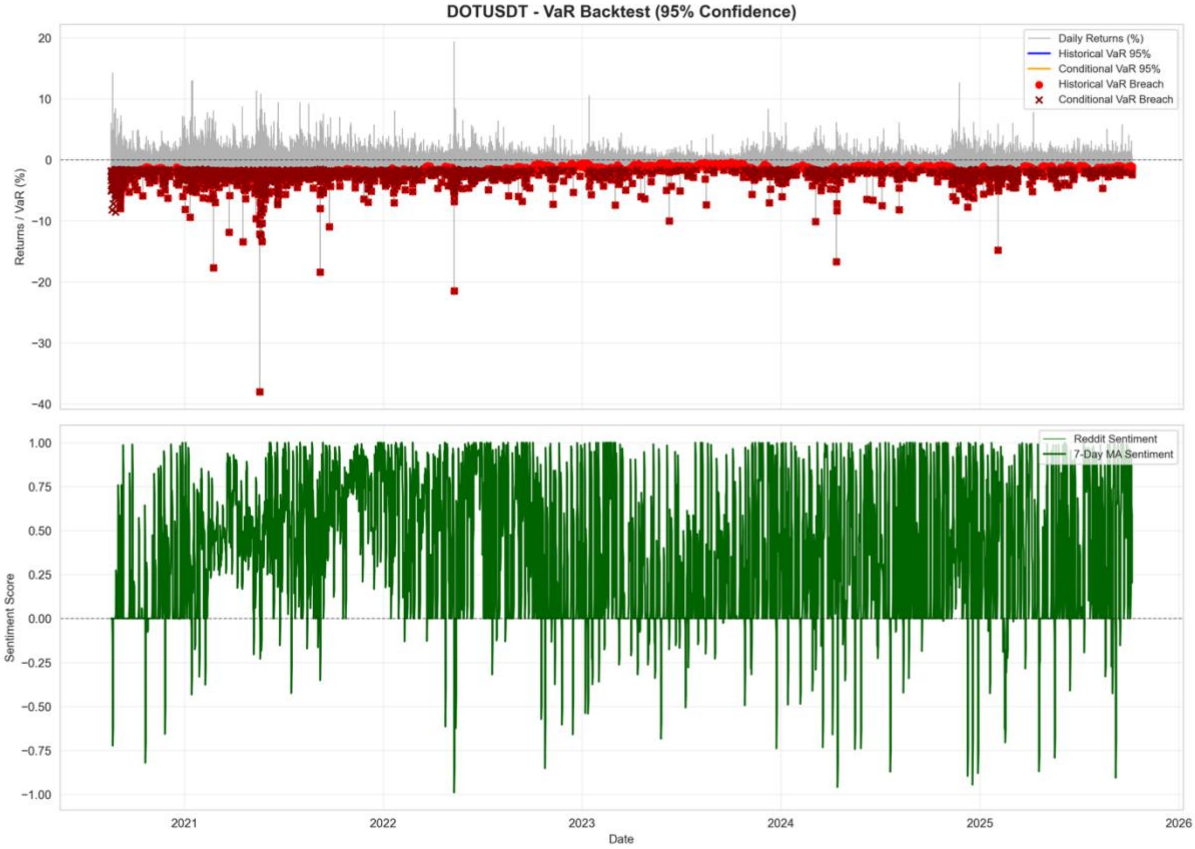


Figure 21

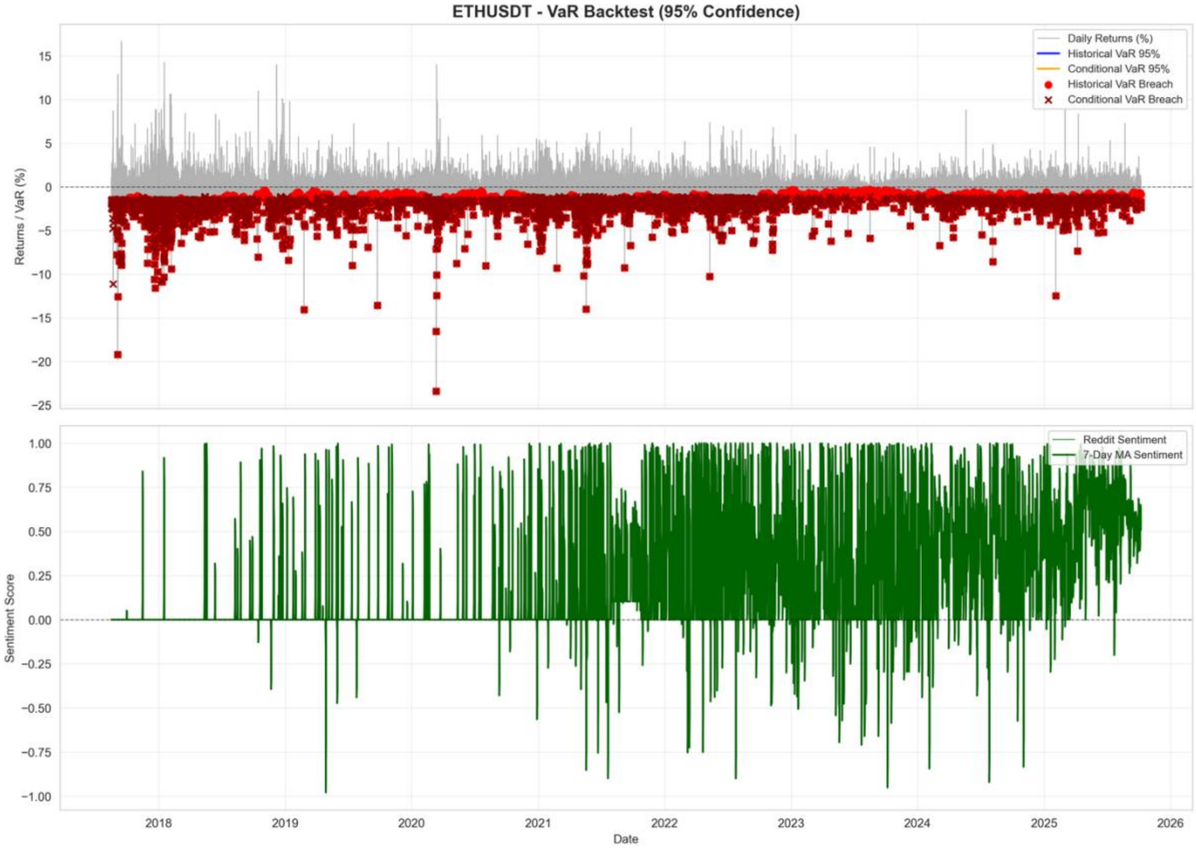


Figure 22

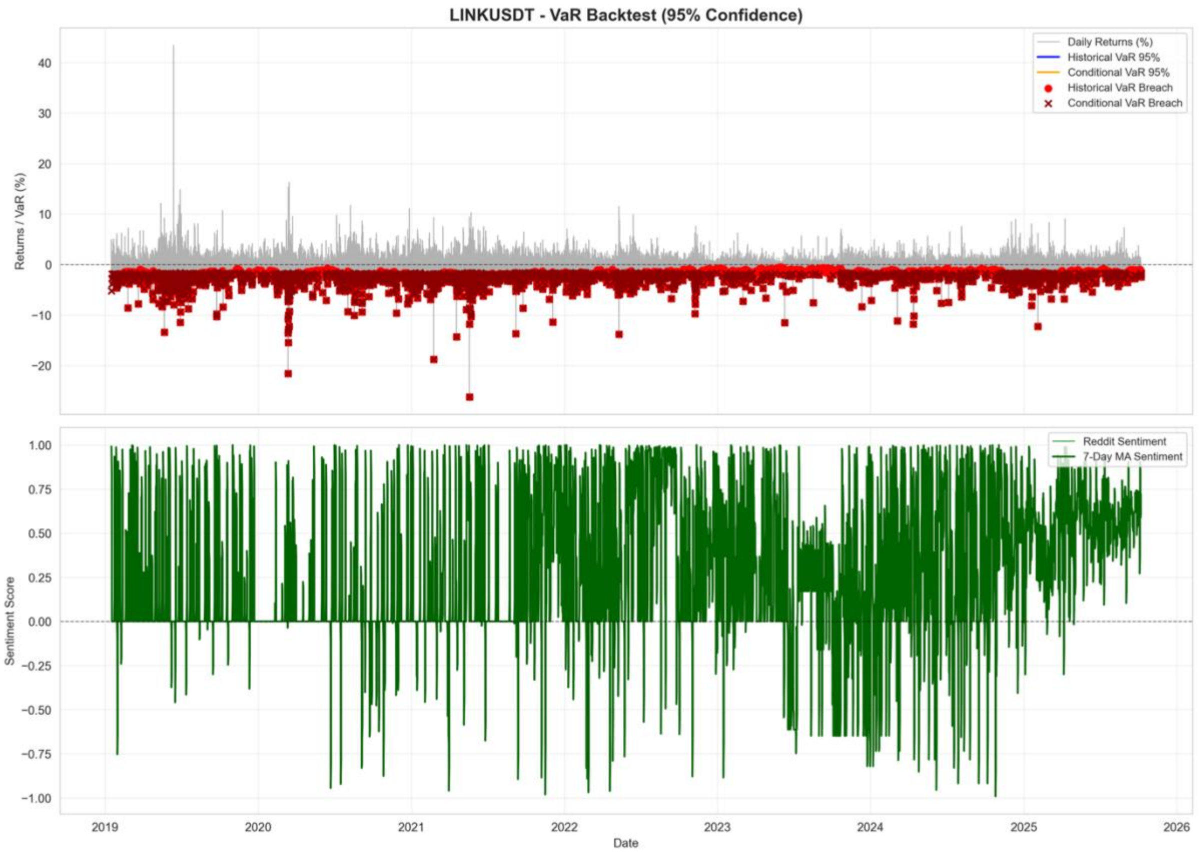


Figure 23

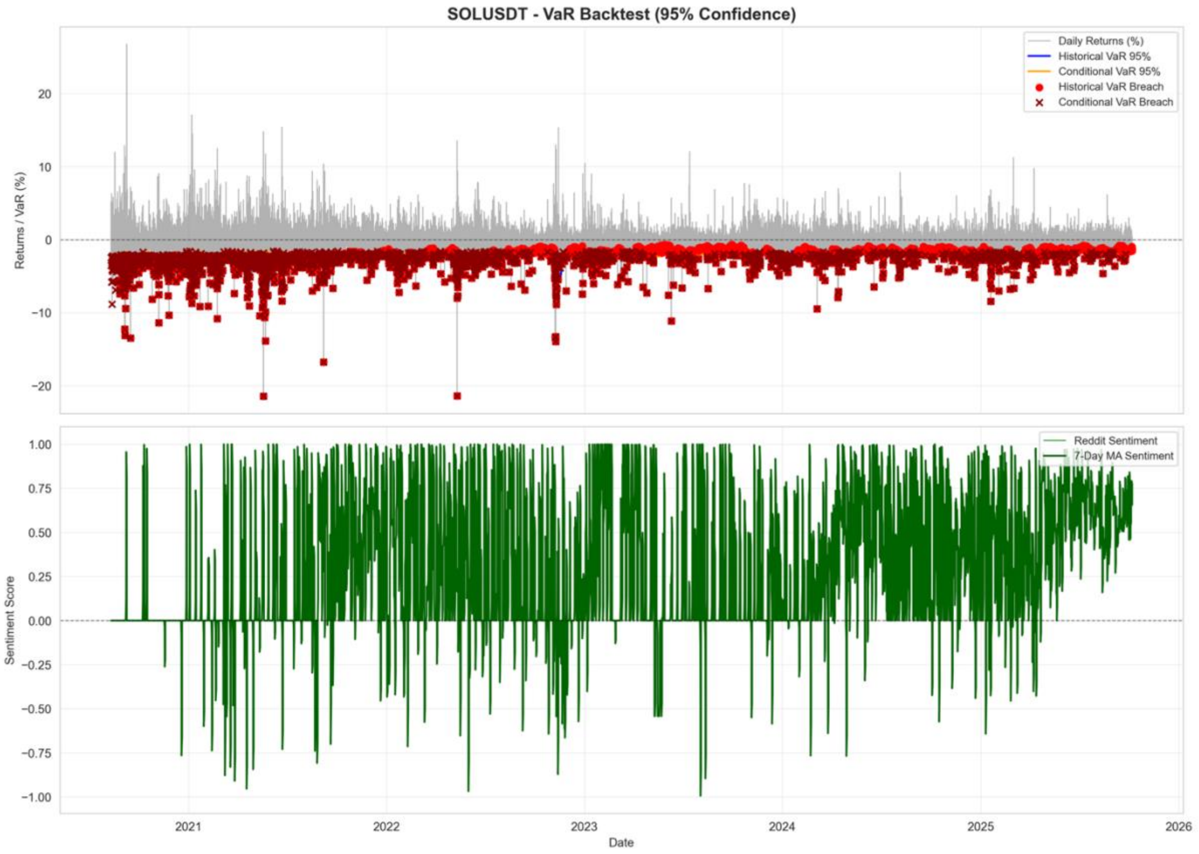
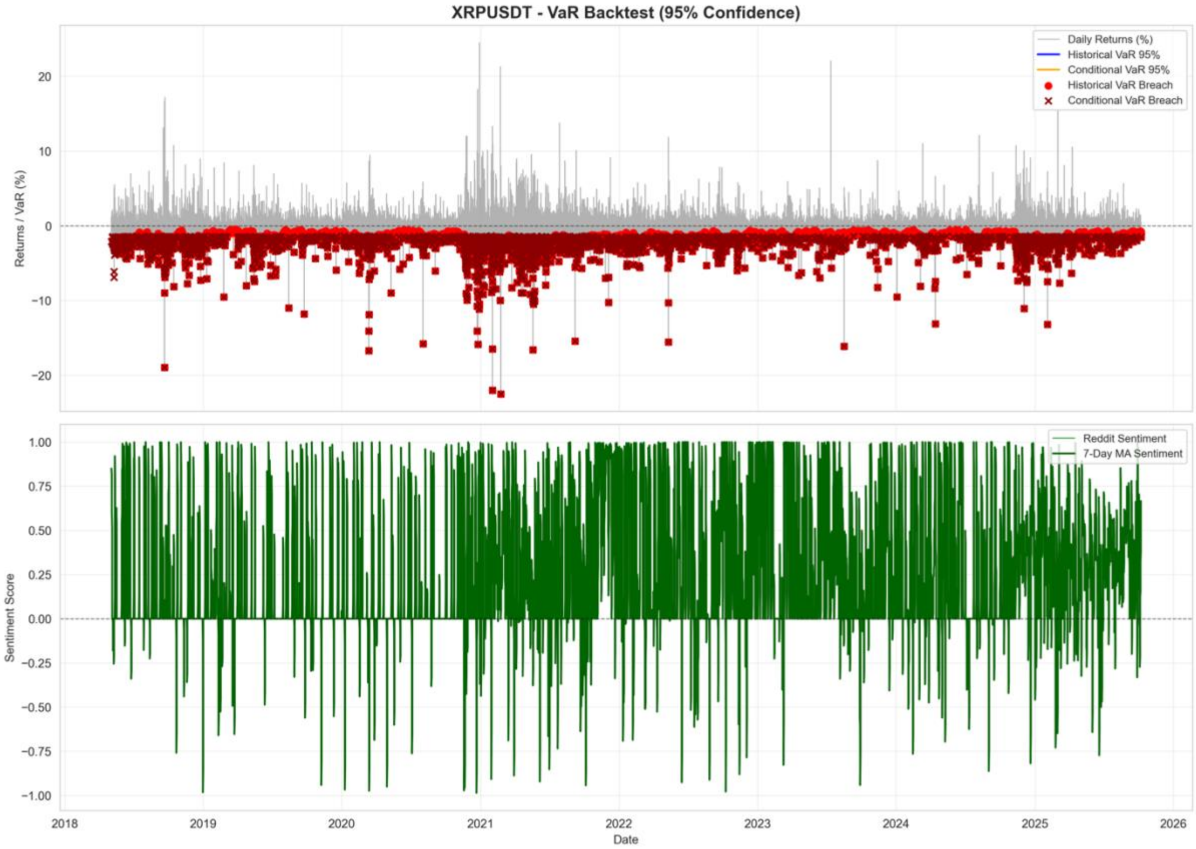


Figure 24


XRPUSDT - VaR Backtest (95% Confidence)
Daily Returns (%)
Historical VaR 95%
Conditional VaR 95%
Historical VaR Breach
Conditional VaR Breach
Returns / VaR (%)
20
10
0
−10
−20
Reddit Sentiment
7-Day MA Sentiment
Sentiment Score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
2018
2019
2020
2021
2022
2023
2024
2025
2026
Date

Citations

Academic Literature:

1) Sahoo & Sethi (2022)
2) Souza and Carvalho (2022)
3) Sakemoto and Nakagawa (2021)
4) Baranovskyi et al (2021)
5) Colianni et al (2015)
6) Wooley et al (2019)
7) Kabo et al (2025)
8) Frendo (2025)

Methodological References:

1) Risk.net explanation of VaR
2) Research paper explaining the Kupiec Test for VaR