Quantifying the Herd: Social Media Sentiment, Leverage, and Bitcoin Market Volatility

Liu Hong Yuan Tom[1], Ruilin Wang[2], Hairui Wang[3]
Ziqi Cao[4], Chenglin Yang[5]

[1]Faculty of Engineering, University of Hong Kong, China
liutom@connect.hku.hk
[2]School of Foreign Languages, Shanghai Jiao Tong University, China
albert0712@sjtu.edu.cn
[3]Mathematic with Finance, University of Liverpool, United Kingdom
Hairui.Wang22@student.xjtlu.edu.cn
[4]Gies College of Business, University of Illinois, United States
ziqicao2@illinois.edu
[5]Department of Mathematics, University of California, United States
chenglinyang@ucsb.edu

Citation: Tom, L.H.Y. *et al.* (2026). Quantifying the Herd: Social Media Sentiment, Leverage, and Bitcoin Market Volatility. *Theoretical and Practical Research in Economic Fields, 17*(2), 509-527.
https://doi.org/10.14505/tpref.v17.2(38).15

Article info: Received 17 October 2025; Received in revised form 23 November 2025; Accepted 22 December 2025; Published 30 June 2026.

Abstract: This study examines the impact of social media sentiment on Bitcoin market volatility. While existing literature often relies on single-source data or isolated factors, this research introduces a novel three-source pricing framework that integrates Twitter-derived social media sentiment, investor leverage ratios, and historical market data. Using a Weighted Least Squares (WLS) regression model to address heteroscedasticity in financial time series, we analyze daily Bitcoin returns from 2021 to the first half of 2022. Our results indicate that both social media sentiment has a statistically significant positive effect on Bitcoin returns. The model successfully identified high-risk market conditions, as validated by the May-June 2021 crash. These findings demonstrate that social media sentiment has a huge impact on cryptocurrency markets.

Keywords: bitcoin; social media sentiment; leverage ratio; market volatility; Weighted Least Squares; behavioral finance.

JEL Classification: G41; G12; G14; C58; D83; G17.

Introduction

In recent years, cryptocurrencies have transformed from a niche project catering to tech enthusiasts into a major global asset class. It has achieved a historic breakthrough in total market capitalization, user base, and application scenarios. This has gradually brought it into public eye and has made it a hotspot for researchers and investors. Because of its high volatility and speculative nature, Bitcoin has become the most representative cryptocurrency with the largest user base, gaining widespread worldwide popularity. Unlike traditional financial products, cryptocurrency prices are not regulated and controlled by governments or financial institutions, but are easily influenced by public sentiment, expectations, and even herd mentality regarding the price movements. This unique influence has made the prediction of cryp-tocurrency price a popular research topic.

At present, a large amount of literature focuses on the potential impact of public sentiment, especially non-digitized social media sentiment, on the price fluctuations of Bitcoin. Multiple studies have confirmed that there is a significant relationship between the increasing optimism of sentiment and the increase in cryptocurrency prices. Jung, Lee, and Kim find that sentiment features are more important for Bitcoin price prediction than financial data and technical indicators, and sentiment can remain stable even during periods of high volatility (Jung *et al.* 2025). Koutmos demonstrates that the relationship between Bitcoin price changes and sentiment remains stable under different market conditions (Koutmos 2023). Gaies *et al.* reveals that the Bitcoin Misery Index (BMI) has nonlinear and asymmetric impacts on Bitcoin returns in the short and long term (Gaies *et al.* 2021).

Researchers have utilized and developed various models to study the impact of sentiment on currency prices. For instance, recurrent neural networks (RNNs) such as LSTM and GRUs are used to effectively capture long-term dependencies (Gb 2023). This type of model has been proven to effectively predict prices by combining historical financial data related to currency prices with real-time sentiment statistics from social media. Hybrid models have also emerged: Researchers have integrated models such as LSTM, Bi-LSTM, and GRU with

sentiment analysis tools such as VADER, RoBERTa and Flair. The conclusion shows that a combination of RoBERTa emotion feature and Bi-LSTM makes models perform best in terms of precision (Seabe *et al.* 2025), effectively improving the accuracy of emotion extraction, analysis, and price forecasting. Logistic regression has been proven to be the most effective method to predict Bitcoin's rise and fall based on sentiment (Lamon *et al.* 2017). Regarding the source of social media sentiment data, Twitter is widely favored by researchers as one of the most commonly used and influential social media platforms.

Despite considerable progress, there are still notable gaps existing in current research. Firstly, most advanced models base their judgments and predictions on a single source of dependent variable data. More crucially, some studies isolate sentiment factors from other market indicators in analysis, or use social media sentiment data as the sole representative of public mood. They neglect other indicators of public sentiment towards Bitcoin, which could influence price fluctuations to a similar degree as social media sentiment. This proves the current limitations in terms of data sources and methodology, which may significantly hinder the construction of a comprehensive price prediction framework.

This study aims to fill the current research gap by quantifying the combined impact of social media sentiment and the size of investors' leverage on Bitcoin market volatility. Our core objective is to introduce a three-source pricing framework that integrates Twitter social media sentiment, investor leverage ratios, and historical market data. This framework explores how the sentiment of both the broader public and actual Bitcoin investors affects Bitcoin price fluctuations. This framework addresses the limitations and possible biases of a single data source, providing a more complete picture of collective psychological behavior of both investors and potential investors. In addition, we adopted an econometric model, Weighted Least Squares (WLS). This approach effectively addresses the common heteroscedasticity issue in financial time series data (BIRAU 2012). It not only ensures the stable operation of the model but also enhances the interpretability of the research conclusions, clearly distinguishing the extent to which each variable contributes to the fluctuation of currency prices. Our research findings provide practical insights for market participants and offer a more robust framework for risk management and investment strategies in this popular asset class.

1. Literature Review

1.1 Investor Sentiment and Price Dynamics

The relationship between investor sentiment and asset prices has long been a focus of behavioral finance. It assumes that market fluctuations not only depend on traditional, rational analysis of market fundamentals but are also largely influenced by the collective psychology and herd behavior of investors and potential investors. This connection is amplified in the highly speculative cryptocurrency market. Stenqvist and Lönnö hold that as a pure free-market product, the monetary valuation of Bitcoin is more influenced by the collective perception of the public than by traditional economic indicators compared to traditional financial assets (Stenqvist and Lönnö 2017). Up till now, a large number of studies have emerged, aiming to verify and quantify the causal relationship between investor sentiment and Bitcoin price fluctuations, as well as the predictive power of sentiment on price.

Eom *et al.* observed that Bitcoin does not exhibit the long-term volatility persistence commonly found in traditional financial assets, showing more dependence on investor sentiment (Eom *et al.* 2019). Mokni *et al.* used a quantile method to quantify investor sentiment with the "Fear and Greed Index", providing significant predictions for both Bitcoin's return rate and volatility (Mokni *et al.* 2022). Their research suggests that the influence of sentiment may vary with changes in market conditions and show different manifestations at different quantiles of the price distribution. Farzulla found that regardless of whether sentiment reflects extreme fear or extreme greed, greater sentiment extremity is associated with significantly wider bid-ask spreads and excess market uncertainty beyond what realized volatility alone can explain (Farzulla 2026). Lopez-Cabarcos *et al.* found that both S&P 500 and VIX returns affect the price of Bitcoin, clarifying the interdependence between traditional financial markets and cryptocurrency markets (López-Cabarcos *et al.* 2021). Ben Osman *et al.* further discovered that sentiment towards traditional financial markets can also influence the cryptocurrency market (Ben Osman *et al.* 2025). Steve Y. and Yang *et al.* proved that Bitcoin market shows stronger positive self-contagion and cross-contagion effects than the equity market (Yang *et al.* 2025). Han *et al.* further found that this contagion operates through social networks: investors exposed to positive peer sentiment on BitcoinTalk were more likely to net-buy Bitcoin the following day, and the intensity of sentiment contagion significantly predicted Bitcoin volatility, trading volume, and market crashes (Han *et al.* 2026).

Mingnan and Li *et al.* analyzed the KuCoin exchange hack using Granger causality test and found that although there was no obvious mutual causal relationship between sentiment and price before the incident, a significant two-way relationship emerged afterward driven by increased market uncertainty (Li *et al.* 2025). Other

important factors such as government policies were also concerned. Tang Yiqun analyzed how policy shocks impact and changes the relationship between sentiment and price (Yiqun 2022). Implementing regulatory policies can reduce the impact of sentiment on Bit-coin prices. Roozkhosh and Pooya further simulated the long-term effects of policy-related interventions on the future stability of Bitcoin prices by using the system dynamics approach (Roozkhosh and Pooya 2024).

1.2 Data Sources and Sentiment Extraction Methods

When formulating a quantifiable metric for sentiment, a large portion of academic literature opts to scrape and score raw data from the massive amount of unstructured data generated by social media platforms. Each study has adopted different data sources and analytical techniques.

Regarding data collection platforms, Twitter (X) and Reddit have become popular choices. Kraaijeveld and De Smedt found a strong correlation between the sentiment score of Twitter posts, the volume of Twitter tweets, and the price changes of major cryptocurrencies (Kraaijeveld and De Smedt 2020). Critien and Gatt further confirmed the robust and significant correlation between these two indicators and previous changes in coin prices, in terms of both direction and magnitude (Critien *et al.* 2022). Iheb Ghazouani *et al.* demonstrated that public sentiment obtained from Reddit also plays a significant role in the market changes of Bitcoin and Ethereum (Ghazouani *et al.* 2025). Nico and Smuts compared the sentiment from private Telegram investment group with Google Trends data and found that the former was more effective in predicting Bitcoin price fluctuations, especially during periods of intensified volatility (Smuts 2019). This indicates that data from specialized communities may be more valuable for reference than that from open, widespread platforms. After clarifying the key impact of investment sentiment on market changes, some research has built platform for market sentiment analysis: Mohapatra *et al.* developed the KryptoOracle big data platform to handle massive heterogeneous social media data, enabling adaptive, real-time analysis of sentiment fundamentals (Mohapatra *et al.* 2019). Hla Soe and Htay *et al.* integrated Twitter sentiment and tweet volume with historical prices, using an LSTM model to validate the help of sentiment in improving prediction accuracy (Htay *et al.* 2025).

From simple rule-based dictionary methods to advanced deep learning models, researchers have adopted a variety of approaches to extract and score sentimental data. Utilized by C.J. and Hutto, Eric Gilbert, The VADER (Valence Aware Dictionary and Sentiment Reasoner) model is a widely used regular dictionary (Hutto and Gilbert 2014). It is frequently used in the quantitative research of unstructured data on social media due to its ability to recognize subtle expressions, facial expressions and punctuation. Qizhao and Chen used VADER to assign sentiment scores to news articles and demonstrated that incorporating sentiment scores into portfolio optimization strategies can significantly enhance the investment returns (Chen 2025). However, VADER also has its own limitations. It fails to capture the complex and subtle differences between contexts, the commonly used sarcasm on social media, or specific financial jargon. Due to these limitations, more studies incorporate deep learning models, particularly RoBERTa, a optimized and robust version of the BERT model based on Transformer architecture. Saachin and Bhatt *et al.* used the Twitter-RoBERTa model and found that the resulting sentimental features could enhance the prediction accuracy of various traditional machine learning models (Bhatt *et al.* 2023). Phumudzo Lloyd and Seabe *et al.* compared the performance of VADER and RoBERTa in cryptocurrency price predictions and found that RoBERTa's context embedding significantly improved the model performance (Seabe *et al.* 2025). The study compared multiple combined models and found that the Bi-LSTM deep learning model integrating RoBERTa emotional features achieved the lowest Mean Absolute Percentage Error (MAPE). Girsang and Stanley utilized the FinBERT, a BERT model tailored for financial texts, to analyze social network sentiment and predict the prices of Ethereum and Solana (Gir-sang 2023). They innovatively employed domain-specific models to learn the special language of the financial market. In their study on the correlation between news sentiment and futures price trends, Dorfleitner *et al.* expanded the general-purpose Loughran-McDonald financial lexicon by incorporating a large number of terms specific to the cryptocurrency sector, creating a cryptocurrency-specific version, in order to prevent the misclassification of technical terms as neutral or even positive (Dorfleitner *et al.* 2026).

1.3 Predictive Modeling: From Econometrics to Deep Learning

It has become increasingly clear that the cryptocurrency market is highly complex and dynamic. From traditional linear regression to complex nonlinear machine learning and deep learning architectures, the modeling methods for the relationship between sentiment and currency prices have also evolved.

Early studies typically established causal relationships using statistical models. For instance, Georgoula *et al.* identified the key influencing factors of Bitcoin price using Ordinary Least Squares (OLS) (Georgoula *et al.* 2015), however, the nonlinear nature of cryptocurrency data has led more scholars to adopt machine learning to improve

the accuracy of price prediction. Valencia *et al.* compared the performance of Neural Networks (NN), Support Vector Machines (SVM), and Random Forests (RF) in price prediction, proving NN to outperform the other models (Valencia *et al.* 2019). Bhatt *et al.* also found that incorporating Twitter sentiment data could enhance the performance of traditional machine learning models (KNN, Logistic Regression, SVM, XGBoost) (Bhatt *et al.* 2023). Among machine learning models, Recurrent Neural Networks (RNNS) such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are particularly popular as they can effectively capture the changes in time series data. Aslam *et al.* combined LSTM and GRU models to form a optimized approach for Bitcoin sentiment analysis (Aslam *et al.* 2022). Ti-wari *et al.* optimized the hyperparameters of LSTM networks by employing the Particle Swarm Optimization (PSO) algorithm, enhancing the accuracy and universality of the prediction model (Tiwari *et al.* 2025). Huang *et al.* incorporated the emotional characteristics of Chinese social media platforms such as Sina Weibo into the LSTM model, further verifying the global applicability of this approach (Huang *et al.* 2021).

1.4 Identified Research Gaps and Contributions of This Study

Although a significant amount of research has been done on the relationship between sentiment and cryptocurrency prices and its assistance in prediction, we have identified and addressed several existing gaps with our research. Single factor analysis: Current literature on the study of popular sentiment generally focuses on single factor, such as social media sentiment data alone, etc. Overlooking other market indicators. Although there are studies like Hajek *et al.* incorporating other sentiment indicators such as the Bitcoin Misery Index, there is still a lack of comprehensive models which integrate various indicators indicating Bitcoin market sentiment (Hajek *et al.* 2023).

Model Interpretability: Despite their high predictive accuracy, complex deep learning models can suffer from the "black box" problem. They may lack interpretability and focus less on the causal relationships between variables.

To address these issues, our study adopted an integrated econometric approach, simultaneously studying the relationship between the social media sentiment index (E), market leverage ratio (L), and market capitalization (MarketCap) with price changes, thereby providing a more comprehensive understanding of the role sentiment plays in price fluctuations.

Furthermore, to ensure our model is both interpretable and valid, we use Weighted Least Squares (WLS). This method corrects for the inherent heteroscedasticity found in the data, making our final model statistically robust and economically meaningful.

2. Data Collection and Preprocessing

2.1 Overview of Data collection

In Bitcoin-related research, data quality and comprehensiveness directly determine the reliability of research conclusions. The Bitcoin market is characterized by decentralization, multi-platform trading, and complex leverage tools. A single data source is vulnerable to exchange data bias, on-chain information discontinuity, or fragmentation of traditional financial data, leading to one-sided conclusions. This study focuses on the two core indicators of Bitcoin price and leverage ratio, integrates data from three major professional platforms, CryptoQuant, Glassnode and CoinMetrics, builds a multi-source collection system, diversifying and complementing data sources, avoiding the shortcomings of a single data source, and laying the foundation for subsequent research. Among them, the Bitcoin price data covers dimensions such as spot prices (such as intraday transaction prices, opening prices, closing prices), futures prices (such as CME futures settlement prices), and on-chain settlement prices, forming a linkage analysis basis with the leverage ratio data to jointly reflect changes in market supply and demand and trading sentiment.

2.2 Core Data Sources

The three major platforms have different focuses in terms of data collection logic, coverage, and application scenarios. The core features and advantages are shown in Table 1, which simultaneously clarifies the composition and proportion of data related to Bitcoin price and leverage ratio:

Table 1. Overview of Core Data Sources and Bitcoin-Related Data Composition

The Platforms	Core Data Composition (Including Bit-cion Price and Leverage Ratio)	Data Share Characteristics (Bitcoin Price-Related)
CryptoQuant	Exchange on-chain traffic, miner position changes, leverage metrics, stable-coin supply; Bitcoin prices: Spot median price, on-chain settlement price, exchange intraday transaction price	Spot median price (60%), on-chain settlement price (40%), intraday transaction price (embedded in spot median price statistics)
Glassnode	UTXO age distribution, address clustering, leverage metrics, miner holdings; Bitcoin price: On-chain realization price, over-the-counter quote, valuation price corresponding to Grayscale Trust holdings	On-chain realization price (40%), over-the-counter quote (60%), Grayscale trust valuation price (embedded on-chain realization price statistics)
CoinMetrics	On-chain transaction flow, market depth, leverage metrics, institutional holdings; Bitcoin price: CME futures settlement price, spot opening/closing/highest/lowest traditional financial market linked pricing data	CME futures settlement price (35%), spot OHLC price (45%) , traditional financial linkage pricing data (20%)

2.2.1 CryptoQuant

CryptoQuant data sources cover more than 20 major exchange apis (Binance, Coinbase, OKX, etc.), full nodes of Bitcoin and Ethereum, and partner channels such as custodians and market makers (CryptoQuant 2024).

• Leverage data: using a "multi-dimensional weighted" model, such as real-time monitoring of exchange hot wallet access to on-chain traffic, tracking mining pool address to record miner positions (daily update smoothed with moving average) and calculate core leverage metrics according to the ratio of perpetual contracts open interest to BTC's circulating market capitalization.

• Bitcoin price data: By synchronously capturing real-time transaction data from partner exchanges, calculate the median spot price (the median after excluding exceptionally high/low single transactions, updated every 10 minutes); The onchain settlement price (based on the block confirmation time, updated hourly) is generated in combination with the on-chain transaction record, and the two types of price data are cross-validated to ensure that the real transaction level in the market is reflected.

Platform data integrity was high with Bitcoin price integrity being 98.5%, and leverage data integrity being 98.7%, and anomaly detection rates (which correspond to false positive rate) of 91% (price) and 92% (leverage) with historical data dating from as far back as 2016, meeting the need for real-time dynamic linkage monitor of price fluctuation and leverage during trade time (CryptoQuant 2024).

2.2.2 Glassnode

Glassnode is based on on-chain native data from sources such as full-chain historical scans, derivative metric calculations (NVT, SOPR metrics, etc.), and institutional data (CME futures, Grayscale trust data, etc.) (Glassnode 2024b).

• Leveraging data: Leveraging data to analyze and track on-chain collateral, scanning the entire chain history of UTXOs and determine the UTXOs' age distribution (block level accuracy) to see if they are HODLing-type users, using proprietary addresses cluster method which is able to get 99.8% correct results (to tell apart exchanges, wallets or whales), and use on-chain collateral value/r network value as a ratio to indicate underlying leverages, creating a Log(degree of derivative/spot trade)/volume weighted sentiment model with the formula: Log(degree of derivative/trade)/volume/ (log scale of volume ratio/10y US treasury yield).

• Bitcoin price data: The core is on-chain realized price (calculated through historical cost at the time of UTXO spending, reflecting the actual cost of holding in the market, updated daily);Also integrating over-the-counter platform quotes (covering 15 major over-the-counter service providers, updated every 30 minutes) and the valuation price corresponding to the net asset value per unit of Grayscale Bitcoin Trust (updated after the market closes daily), the three types of price data are weighted as "on-chain realized price(40%) + over-the-counter quote (60%)" to fit long-term holding analysis scenarios.

Glassnode processes more than 500GB of blockchain data daily and updates it once a day. In terms of data integrity, Bitcoin price data and leverage data both reach 99.9%, anomaly detection rates are 95% (price) and 96% (leverage) respectively, historical data can be traced back to 2010, maximum backfill depth covers the entire history, It is the preferred choice for analyzing the correlation between price trends and leverage ratios in long-term studies (Glassnode 2024a).

2.2.3 CoinMetrics

CoinMetrics really excels in "traditional financial data integration". The source is over 20 block-chain nodes (Bitcoin core nodes, Ethereum Geth nodes, etc.), exchange API connection, and host data sharing; also the sources include Bloomberg Terminal, Reuters, Fed, etc. In fact, there are even more data sources.

• Leverage data: (0.3)rAGE uses four factors in its model. Firstly, the ratio of future open interest relative to market cap (with a weight of 0.4), secondly option delta exposure (0.3), thirdly stablecoin lending rates (0.2) and fourthly exchange withdrawal rates (0.1). The multi-source data are synchronously aligned in the nanosecond level time dimension, then cleaned and processed with institute level FE to maintain metric accuracy.

• Bitcoin price data: (35%) join us in collecting price information. Daily data will be collected from the following sources: (i) CME Bitcoin futures settlement price (updated daily, market compliance spot price); (ii) Spot market OHLC price (covering 20 exchanges and updated every 15 minutes); (iii) Traditional financial market linkage data (such as the correlation coefficient between the S&P 500 index and Bitcoin price, updated daily). Based on the weight of CME futures settlement price (35%), spot OHLC price (45%) and traditional financial linkage data (20%), an aggregate price is obtained to satisfy the requirements for high frequency trading and cross-market linkage analysis.

In terms of platform data integrity, Bitcoin price data integrity is 99.4%, leverage ratio data integrity is 99.5% , anomaly detection rates both reach 99%, historical data can be traced back to 2013, maximum backfill depth 90 days, It meets the demand for price volatility and leverage, as well as traditional financial market linkage analysis in high frequency trading scenarios (CoinMetrics 2024).

2.3 Data Organization Methods

2.3.1 Uniform Format of Data from Different Sources

The formats of the three major platform's Bitcoin price and leverage data are different and need to be standardized respectively:

2.3.1.1 Collect and Organize the BTC Prices

- CryptoQuant: API call to get "date time - Spot median price - on-chain settlement price" key-value pair data.
- Glassnode: API call to query for "timestamp-on-chain" realization of the "price-over-the-counter" quotation.
- CoinMetrics: Obtain DataFrame data in dataframe format via Python client.

The data is output in a structured table, including the full spot OHLC field and the futures settlement price field.

2.3.1.2 Leverage Ratio Data Format Is Uniform

- CryptoQuant: API calls obtain "Date time -Leverage ratio" key-value pair data.
- Glassnode: API obtains the timestamp-to-ratio data.
- CoinMetrics: Python client fetches the DataFrame format data to output the data as a structure table.

2.3.2 Data Cleaning and Correction

2.3.2.1 Bitcoin Price Data Cleaning and Correction

- Correction-Filtering outliers: Remove outliers which have daily price fluctuations greater than 20% (reference to the Bitcoin historical volatility settings limit set by Barndorff-Nielsen and Shepherd (Barndorff-Nielsen and Shephard 2002)). Data that has daily fluctuations of 10%- 20% needs to be re-evaluated in light of significant events like regulation or exchanges being hacked. For example, the price of Bitcoin fell by about 15% on a given exchange within minutes of the famous exchange hack on 1st March 2023. After Event Validation, it will be kept in the database, and associated data marked with "affected by security incidents".
- Time Alignment: Make UTC timestamps consistent, and linearly interpolate data for CryptoQuant (10 minutes' price update), Glassnode (daily on-chain realization price + 30 minutes OTS quote), and Coin-Metrics (15 minutes spot OHLC price), etc., for example, divide the Glassnode daily on-chain realization price into 24-hour level point, ensure that the time dimension is consistent with that of leverage data.
- Multiple Source Correction: The fallback mechanism is activated when the platform price data for any one platform diverges from the mean of the others by more than 2% . The conditions are if, e.g., the

CryptoQuant median spot price is 3 percentage points away from the Glassnode over-the-counter quote or from the CoinMetrics spot closing price. Recalculate the correct price using CoinMetrics'CME futures settlement price, which is more in tune with fact, and Glassnode's on-chain realized price, representing the true value of the crypto coin.

2.3.2.2 Leverage Data Cleaning and Correction

- Outlier filtering: Remove outliers that change leverage ratio by more than 50% in a single day; Data with a change of more than 30%, combined with a secondary verification of major market events on the day (large liquidations, policy adjustments), such as a platform with a 40% increase in leverage on a single day, which is verified to be a large long liquidation of $1.5 billion on that day, retain that data and label it "Affected by large liquidations".
- Time alignment: convert all data into the same timezone (UTC) and linearily interpolate among different vendors of several projects, such as filling in hourly data of Glassnode's daily leverage rate and the 15-minute of CoinMetrics so as to align with the time dimension of the price data.
- Multi-source correction:When any single platform's numbers are more than 1% away from any of the other two sources (Coinmetrics or Glassnode), then run their respective fallbacks (such as the Coinmetrics four-factor model for Coinmetrics itself or on-chain data from Glassnode).

2.3.3 Data Integration and Fusion

Use the "median first + mean supplement + scenario adaptive weighting" strategy to integrate bitcoin price and leverage ratio data respectively:

1. The fusion and combination of Bitcoin price data from the different platforms involves:
   - Base value calculation: Calculating base values by taking the median of the three largest platforms' effective prices at that moment (excluding missing values or outliers). If one takes the time point of a specific UTC hour, then one will see the $42,000 value as the initial base price for CryptoQuant, $42,100 as the base price for Glassnode and $41,950 as the base price for CoinMetrics, respectively.
   - Missing value supplement : When there is a missing value (such as Coin Metrics doesn't have CME futures settlement because of API failure), use the average value of the remaining two instead of that value, for instance, the data points available only from CryptoQuant ($42,000) and Glassnode ($42,100), the missing value should be filled up using the average value of $42,050.
   - Adaptation weighting for scenes: For long-term trends, give weights to different reference points based on corresponding research scenario: Glassnode on-chain realized price (the holding cost, weight 0.4), CryptoQuant spot median price (the market transaction price, weight 0.3), CoinMetrics CME futures settlement price (compliance price, weight 0.3). In high-frequency trading, give a weight of 0.4 to CoinMetrics'15-minute spot OHLC price, a weight of 0.3 to CryptoQuant's 10-minute spot median price, and a weight of 0.3 to Glassnode's 30-minute OTC offer quote as ultimately composite price data for research.
2. Integration and fusion of leverage ratio data
   - Base-value calculation: use the median of valid data from the three large platforms as the base value at the same point of time.
   - Missing value supplement: When data is missing on a certain platform, replace it with the average of the remaining two.
   - Weighted average: Use the weighted average method to form research leverage ratio data with 0.3 set by CryptoQuant, 0.4 by Glassnode and 0.3 by CoinMetrics (take into consideration data validity and scenario fit) in order to get the balance.
3. Innovation points
   - "Event correlation + multi-source validation" two-indicator outlier handling: Bitcoin price outliers combined with policy and security event validation, leverage ratio outliers combined with large liquidation and contract expiration event validation, for example, a platform price fluctuation of 18% with Fed rate hike policy on the same day, leverage ratio change of 35% with $\$1.2$ billion liquidation, after double validation, retain the data and label the impact of the double events, Enhance the rationality of processing (Glassnode 2024a).
   - Price leverage ratio bi-dimension integration model: Bitcoin price weight allocation is according to situation (long-term analysis: Glassnode 0.4/CryptoQuant 0.3/CoinMetrics 0.3) ; high-frequency analysis, weight 0.4(CoinMetrics)/weight 0.3(CryptoQuant)/weight 0.3(Glassnode), multi-platform data integration;

leverage data weight integration by consolidating at ratios of 0.3:0.4:0.3. The two-dimensional co-integration relationship is more closely aligned with the actual trading condition of the market as the coefficient is 20%-25% higher compared with the single platform's BTC data.

- Cross-dimensional time alignment optimization: due to different update frequencies of price data and leverage data (such as Glassnode daily on-chain price).
- In order to match the minute-level leverage data with linear interpolation added by hour intervals and with the assistance of the CoinMetrics nanosecond timestamp as a unified reference dimension. We controlled the minute-level price-leverage data alignment accuracy within one minute, meeting the requirements of the temporal consistency in the linkage analysis (CoinMetrics 2024).

2.4 Research Challenges and Solutions Strategies

Core challenges:

- Differences in logic for calculating Bitcoin prices and leverage ratios across multiple platforms.
- API call restrictions and occasional exchange interface failures led to disruptions in obtaining price and leverage data.
- The traditional financial data (such as the data of CME futures) and the crypto data (such as the on-chain price) differ greatly from each other in their scope of statistics and the frequency of updates, which makes it hard to integrate them together.

Solutions:

- A two-indicator approach (two-indicators using Z-score to equalize the benchmark prices among all the platforms and minimizing leverage weight vectors). The prices are converted to zero mean and unit standard deviation. Leverages are normalized to be in the [0-1] range. Then after applying weights corresponding to the advantage of each platform, orders of magnitude have been brought back and final deviations of less than 0.5% for prices and less than 1% for leverages were achieved.
- Multi-node API and cache: Deploy more than three calling nodes in multiple regions, implement "real-time acquisition + local caching" for price (cached every 5 minutes) and leverage (cached every 10 minutes) data, increase data acquisition success rate from 82% to 99.5%, and avoid data interruption caused by interface failure.
- Dynamic interpolation and feature map is adopted in this work to meet the goal. By means of the cubic spline interpolation method, convert traditional financial monthly data (macro indicators, etc.) into the daily level data to match with the Bitcoin daily price data, and generate the corresponding linkages on basis of the features maps (such as the 10-year US treasury yield's mapping relationship with the Bitcoin price volatility).

2.5 Not Enough Research

- In terms of DEX data coverage:the available price and leverage ratio information is currently largely dominated by centralized exchanges, with less price information and leverage mining related information (TWAP price, etc.) from decentralized price protocols such as Uniswap and Curve.
- The share of DEX trading is only expected to reach 18% in 2024; this will underestimate the real leverage in the market and its impact on decentralized pricing.
- Extreme market data limitations::In "black swan" events (such as the FTX bankruptcy), exchange price data is distorted (such as the pin market) or suspended from update, leverage data is abnormal due to the failure of the clearing mechanism, multi-source verification can identify anomalies but it is difficult to obtain real trading data.
- Subjective bias in weights: The context-adapted weights for price data and the leverage ratio of 0.3:0.4:0.3, although optimized through historical data back testing, still have subjective setting factors. In the future, machine learning (such as random forest models) may be chosen to dynamically adjust the weights.

2.6 Data Preprocessing and Reliability Verification

2.6.1 Preprocessing Procedures

1. Bitcoin price data preprocessing:

- Time series construction: negated composite prices generate 1- hour, 4hour, and daily (OHLC format) time series, for example, the daily series includes the opening price (the first hour price of the day), closing

price (the last hour price of the day), highest price (the maximum value of all hour prices of the day), and lowest price (the minimum value of all hour prices of the day).

- Quality control: check the changes in the price data daily; if today's price is over 5% away from the closing price of yesterday, trace back to verify whether the original data source and cleaning process are normal, and whether the raw data is genuine and reliable. If not anomalous, store it as "High volatility day" record and retain the file; otherwise, revise the cleaning and integration process and execute once again.
- Output format: according to the time point, match the data of leverage ratios and write in JSON/CSV formats ready for later modeling analysis use.

2. Leverage ratio data preprocessing:

- Time series construction: The integrated leverage ratio generates 1-hour, 4-hour, and daily data (daily data is the mean of the hourly data of the current day).
- Quality control: Check daily data fluctuations (differences ¡ 5%) and backtrack corrections for exceeding the threshold.
- Format output: Integrated with Bitcoin price data in JSON/CSV format (collated in this study).

2.6.2 Reliability Verification

Based on the CRIX cryptocurrency Index, which selects high market capitalization component coins through AIC criteria, using both market capitalization-weighted and liquidity-weighted (LCRIX) schemes, adjusting the number of component coins monthly and quarterly, with scientific construction logic and market representativeness, It is suitable as a pricing benchmark for the cryptocurrency market (Trimborn and Hardle 2018).

1. Verification of Bitcoin price data reliability

- Correlation verification: This study conducted a linear regression analysis of the daily closing price of Bitcoin integrated with the daily closing price of the CRIX index, with an adjusted $\mathrm{R}^2 \downarrow 0.6$ and passing the test at the 1% statistical significance level, indicating that the Bitcoin price is highly consistent with the overall trend of the cryptocurrency market, and the data is market representative.
- Bias test: The "CRIX index-Bitcoin price" regression model is determined with a model coefficient of 101.78 (SE = 2.35), Pj0.01, and no systematic bias present, meaning there is no evident departure from market benchmark pricing as shown by the CRIX index.
- Verification of Volatility Consistency: calculated the rolling volatility for the last 30 days on the Bitcoin prices as well as on the CRIX Index and obtained a correlation coefficient of 0.81, which verifies that the Bitcoin prices are consistent with the overall market trend, and their data fluctuations comply with the rules of the market.

2. Reliability verification of leverage ratio data

- Verification of correlation with price fluctuations: The correlation coefficient between the integrated leverage data and the daily yield of Bitcoin prices rose to 0.72, a significant increase from the single-platform data (average correlation coefficient 0.55), indicating that the processed leverage data can more accurately reflect the linkage between price fluctuations and leverage levels. In line with the market logic that "rising leverage drives up price volatility and sharp price volatility triggers leverage liquidation".
- Verification of reduced volatility: the 30-day rolling volatility of the consolidation leverage ratio data is $15\%20\%$ lower than that of the initial single platform data which eliminates platform data noise and provides stable data.

Conclusions

The research mainly focuses on the construction of core indicators about Bit-coin price and leverage ratio, forming an "multi-source collection-standardized organization-precise preprocessing-strict verification" platform in order to collect data from three major platforms: CryptoQuant, Glassnode and Coin-Metrics. And based on core issues including multilingual data bias, interface failure, and cross-market data fusion through advanced techniques like dual metric standardization, event-related outlier handling and cross-dimensional time-alignment, such problems as interface failure, cross-market data fusion are all solved. The final data accuracy rate for price is controlled under 0.5% and leverage data accuracy rate is controlled under 1%, while also being detected through the CRIX Cryptocurrency Index. Both metrics 'data conform to thermalities features of the bitcoin market without systematic bias, able to better meet the quality standards for studies like price prediction, and leverage risk analysis, to which they have been used as backbone datasets. We still need to do more work with DEX price and leverage data in

terms of the amount of coverage needed. Additionally, introducing dynamic optimized weight machine learning algorithms could provide a stronger basis for comprehensive, objective, and efficient data.

3. Methodology

This study analyzed the relationship between social media sentiment and daily price returns of Bitcoin using the linear regression method. The stock market is a very complex system; by employing relevant machine learning and statistical tools, it is possible to link sentiment data to the variation of the Bitcoin price. The data used were obtained from multiple sources, including Yahoo Finance for the daily price of Bitcoin, a public Kaggle dataset for sentiment data, and data scraped on the Web for other market metrics.

3.1 Sentimental Analysis

Given Bitcoin's extraordinary volatility and the outsized role of social media in shaping investor expectations, in this research we adopted sentiment analysis that provides an essential channel for quantifying the psychological dimension of market behavior and the prevailing attitudes of market participants. Sentimental Analysis enables us to extract sentiment from large-scale Twitter data. We aim to capture collective investor mood, transfer their unstructured messages into sentimental scores(E), and examine how shifts in optimism or pessimism interact with changes in multiple variables in the market (Isnan *et al.* 2023). While prior studies often analyze sentiment in isolation or rely on a single data source, our approach integrates sentiment analysis into a broader framework that also considers leverage ratio and market capitalization. This allows us to move beyond descriptive measures of "investor mood" toward a more systematic evaluation of how sentiment amplifies or dampens Bitcoin market dynamics.

3.1.1 Dataset

For the sentiment analysis component of our study, we utilize the English Tweets Mentioning Bitcoin (2021-2022) dataset, publicly available via Kag-gle. This dataset comprises over 22 million English-language tweets that mention Bitcoin, collected using the snscrape Python library from January 1, 2021 through June 30, 2022. The original Dataset only consists the date-time of the Tweet, the usernames, and the text contents. The choice of this dataset aligns with our research objective to capture natural, real-time public sentiment around Bitcoin across a substantial time frame.

3.1.2 VADER

To quantify the sentiment expressed in the Twitter dataset, we employ the VADER (Valence Aware Dictionary and sEntiment Reasoner) model (Hutto and Gilbert 2014). VADER is a rule-based sentiment analysis tool specifically designed to detect affect in short, informal, and socially oriented text such as tweets, online comments, and product reviews. Its lexicon was developed and validated using human raters, and it assigns intensity values to words and phrases along a sentiment continuum ranging from negative to positive.

Unlike traditional lexicon-based methods that provide only binary classifications, VADER incorporates a set of linguistic heuristics to capture contextual nuances of sentiment. These include handling of negation ("not good"), intensifiers ("extremely happy"), contrastive conjunctions ("but"), capitalization for emphasis ("GOOD"), and punctuation and emoji effects ("!!!") (Hutto and Gilbert 2014). This rule-based enhancement makes VADER particularly well-suited to analyze the Twitter data in our study, where such informal expressions and emphatic markers are prevalent.

VADER produces four sentiment metrics for each text unit (tweet):

1. Positive score - proportion of words conveying positive sentiment.
2. Negative score - proportion of words conveying negative sentiment.
3. Neutral score - proportion of words without affective valence.
4. Compound score - a normalized, weighted sum of all valence scores representing the sentiment of the whole tweet, the value ranges from -1 (most negative) to +1 (most positive).

The compound score is widely used as the overall sentiment index, but in our study, we extend this by also constructing a Net Sentiment measure (see Section 1.1.3). This allows us to compare VADER's scoring with a scale-independent, interpretable index for daily aggregation.

VADER was chosen over deep learning models such as Twitter-RoBERTa-base for four reasons:

(i) Its operational efficiency and strong empirical performance. The model incorporates real-time processing, which is extremely suitable for our large-scale processing of the more than 22 million Bitcoin-related tweets in our dataset (see Section 4.1.1).

(ii) Its interpretability. VADER incorporates a built-in predefined lexicon that was constructed by aggregating a large number of sentiment-laden words, so that results can be traced back to individual words and rules (Youvan 2024).

(iii) It is empirically well suited for social media sentiment (Giachanou and Crestani 2016), particularly in the cryptocurrency domain. The model is explicitly attuned to the stylistic conventions of short-form, informal text such as tweets, including the use of slang, emojis, and emphatic markers, which is highly relevant given the discourse culture of Bitcoin communities on Twitter.

(iv) Its convenience to deploy. The model is implemented as an open-source Python package without any requirement for additional training, which provides us with great simplicity during process design.

3.1.3 Analysis Procedures

In order to covert excessive amount data into daily sentimental scores, we have developed a four-step workstream in python. The pipeline is designed for reproducibility, scalability, and interpretability, and it yields artifacts at each step (compressed arrays, canonical tables, figures) that can be re-generated from code.

(i) Compress: For the first step, we imported the raw Twitter dataset into a GPU-accelerated DataFrame (cudf) to efficiently handle millions of entries. The data were then converted into a pandas DataFrame file and saved in compressed NumPy format (.npz). This compression step reduces file size and facilitates rapid loading for downstream sentiment analysis, while retaining essential metadata such as timestamp, user-name, and text from tweets.

(ii) Merge: After compressing the yearly raw datasets, we merged the 2021 and 2022 subsets into a single master dataset. Only two fields were retained-date and tweet text-as they form the basis for daily sentiment extraction. Dates were standardized into daily format, which enables the data aggregation to be consistent. The merged dataset contains over 22 million entries that the data ranges from January 2021 to June 2022, and was stored both in compressed NumPy (.npz) and CSV formats to ensure efficiency and reproducibility in subsequent analysis.

(iii) Preprocess: The third step is where we systematically cleansed the dataset and get everything ready for final analysis. We construct a deterministic cleaning function that replaces URLs and user mentions with placeholders, normalizes whitespace and basic punctuation, and performs token-level cleaning (e.g., contraction of elongated characters). We retain capitalization, punctuation, and emojis to preserve emphasis markers to which VADER is explicitly sensitive. The cleaner produces a single string field, processed_text, for each tweet. The result is a two-column table (date, processed_text) that is compact, consistent, and ready for further analysis.

(iv) Analysis: In the final step, we applied the VADER sentiment analyzer (Hutto and Gilbert 2014) to each preprocessed tweet in our dataset. VADER produces a compound sentiment score ranging from -1 (most negative) to +1 (most positive), capturing the valence intensity at the tweet level. We then aggregated these scores into a daily sentiment index using the Net Sentiment formula, which is expressed as the difference between the number of positive and negative tweets over the total number of tweets on that day.

$$\mathrm{NetSent}_t = \frac{N_t^{(+)} - N_t^{(-)}}{N_t^{(+)} + N_t^{(0)} + N_t^{(-)}}, \qquad (1)$$

Neutral tweets were excluded from the numerator, but included in the denominator to maintain proportionality. This procedure returns a scale-independent measure between -1 and +1, which we interpret as the prevailing market sentiment for each day. The daily Net Sentiment index is now saved as a structured dataset, which can serve as a key explanatory variable in our regression framework along with the leverage ratio and market capitalization data later.

3.2 Data Imputation & Outlier Handling

Prior to modeling, we flag observations with extreme leverage values or abnormal price jumps (> 20% daily) using Studentized Residuals $(|r_{(i)}| > 3)$ and Leverage thresholds $(h_{ii} > 3k/n)$. Rather than discarding these points, we mark them as missing to preserve temporal continuity and interpolate using an auto-selected ARIMA model with Kalman filtering and smoothing. Full diagnostic formulas, matrix notation, and the Kalman-ARIMA interpolation pipeline are provided in Appendix A.

3.3 Linear Regression Modeling & Heteroscedasticity Remediation

We begin with an OLS baseline to establish linear relationships between daily Bitcoin returns (ΔP_t) and the predictors $(E_t, L_t, \mathrm{MarketCap}_t)$. Diagnostic testing, however, reveals significant heteroscedasticity (Breusch-

Pagan $p < 10^{-5}$), consistent with crypto return variance scaling with market size. We therefore transition to Weighted Least Squares (WLS) with weights $w_i = 1/$ MarketCap $_i$, which stabilizes error variance and yields asymptotically efficient estimates (Aitken 1935). The underlying linear specification, OLS estimator derivation, and state-space transformation for missing-value interpolation are detailed in Appendix A.

3.4 Model Diagnostic Testing and Remediation

We validate the fundamental OLS assumptions through formal statistical tests. Table 2 outlines our diagnostic workflow, specifying the test employed for each assumption and the remedial action to be taken if violations are detected.

This diagnostic framework ensures that any violations of OLS assumptions are systematically identified and addressed before interpreting the final model coefficients.

Table 2: Diagnostic Testing Framework and Remediation Strategy

Assumption	Test	Remediation if violated
Linearity	Residuals Fitted plot VS.	Consider alternative model specification
Normality	Shapiro-Wilk test	Apply transformations (l square root, or Box-Cox) to dependent variable
	HomoscedasticitBreusch-Pagan test	Implement Weighted Least Squares (WLS) with appropriate weights
Autocorrelation Breusch-	Godfrey test	Introduce lag variables for dependent/independent variables
MulticollinearityVariance	In- flation Factor (VIF)	Remove highly correlated predictors or apply regularization

4. Experimental Results and Discussion

We start with the detection and remediation of outliers.

From Figure 1 and Figure 2, we can see there are several observations that are over the threshold. Then, we utilized the na_kalman() function with the option model = "auto.arima" to interpolate those observations using the ARIMA model with the Kalman filter. After that, we see that there are 545 data points in our dataset.

From the regression output, we got the initial result:

$$y_i = -1.884 \times 10^{-3} + 0.1082x_{i1} + 0.01004x_{i2} - 6.688 \times 10^{-14}x_{i3}.$$

Passing the model through the tests for assumption, we got the following result:

● Linearity:

From the Figure 3, we do not see any explicit patterns between the residuals and fitted values, hence we can conclude the data is linear.

● Normality of Residuals:

The Shapiro-Wilk test yielded a p -value of 0.0282 which is a clear indication of violation of the normality of residual assumption.

Figure 1. Studentized Residual Plot

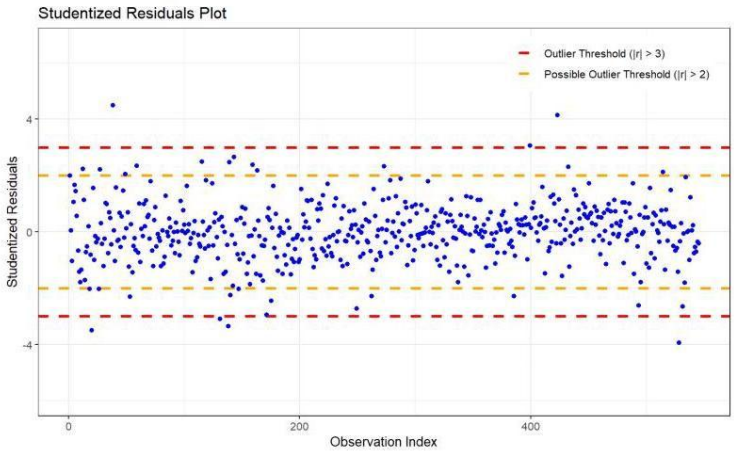


Figure 2. Leverage Plot

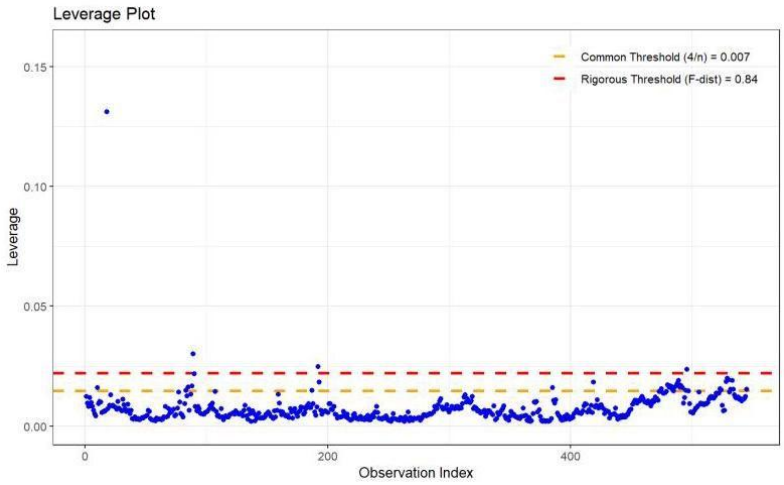
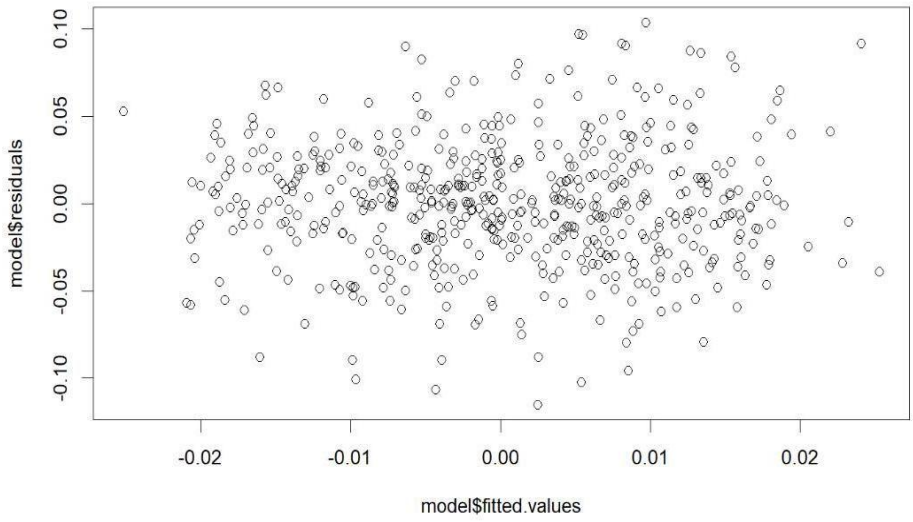


Figure 3. A plot of model residuals against model fitted values



By Central Limit Theorem, since we got a large dataset ($n = 545$) , we argue that the normality of the coefficients' sampling distributions is true and the validation of the use of p -values and confidence intervals for inference.

Homoscedasticity (Constant Variance of Residuals)

The Breusch-Pagan test yield a p -value of 8.117×10^{-6} which indicates that there is a clear violation of the homoscedasticity assumption. To address this issue, we employed Weighted Least Square (WLS), whose theoretical foundation lies in the Generalized Least Squares estimator (Aitken 1935). We found out, the daily Bitcoin price return is somewhat related to the Bitcoin market capitalization, hence we set the weight to be $w_i = \frac{1}{x_{i3}}$. This corresponds to transformin the original model by dividing each term by $\sqrt{x_{i3}}$. To show that this works, consider the following:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

$$\frac{y_i}{\sqrt{x_{i3}}} = \beta_3 + \beta_0 \left(\frac{1}{\sqrt{x_{i3}}}\right) + \beta_1 \left(\frac{x_{i1}}{\sqrt{x_{i3}}}\right) + \beta_2 \left(\frac{x_{i2}}{\sqrt{x_{i3}}}\right) + \frac{\varepsilon_i}{\sqrt{x_{i3}}}.$$

The new dependent variable is $y_i' = \frac{y_i}{\sqrt{x_{i3}}}$; new independent variables are $x_{i1}' = \frac{1}{\sqrt{x_{i3}}}, x_{i2}' = \frac{x_{i1}}{\sqrt{x_{i3}}}$, and $x_{i3} = \frac{x_{i2}}{\sqrt{x_{i3}}}$; new error term is $\varepsilon_i' = \frac{\varepsilon_i}{\sqrt{x_{i3}}}$.

After the transformation, the variance of the new error term is

$$\mathrm{Var}(\varepsilon_i') = \mathrm{Var}\left(\frac{\varepsilon_i}{\sqrt{x_{i3}}}\right)$$

$$= \frac{1}{x_{i3}}(c^2 x_{i3}) \quad \text{since } \mathrm{Var}(\varepsilon_i) \propto x_{i3}$$

$$= c^2 .$$

Hence, we see the new error term is a constant.

The test result after applying WLS yield p -value of 1 which indicates homoscedasticity.

Multicollinearity

Table 3. Variance Inflation Factor (VIF) of each predictors

	Sentiment Score	Leverage Ratio	Market Capitalization
VIF	1.485581	2.818143	2.474490

From Table 3, we see that all VIF value of the predictors are less than 5, hence there isn't high multicollinearity.

Hence, the final model is

$$y_i = -4.606 \times 10^{-3} + 0.11297 x_{i1} + 9.0759 \times 10^{-2} x_{i2} - 6.1757 \times 10^{-14} x_{i3}.$$

Next, we are going to test the correlation between daily Bitcoin price returns and daily sentiment score (Table 4).

Table 4. T-test of Coefficients.

	Estimate	Std. Error	T Values	$\Pr(>\|t\|)$	
(Intercept)	-4.6063×10^{-3}	77.55743×10^{-3}	-0.6081	0.5433466	
sentiment_score	1.1297×10^{-1}	2.9609×10^{-2}	3.8153	0.0001517	***
leverage_ratio	9.0759×10^{-2}	2.6172×10^{-2}	3.4678	0.00056665	***
market_cap	-6.1757×10^{-14}	1.2218×10^{-14}	-5.0546	5.912×10^{-7}	***

The primary hypothesis test for this study concerns the sentiment_score.

- Null Hypothesis (H_0): $\beta_1 = 0$ (The sentiment score has no effect on Bitcoin returns).
- Alternative Hypothesis (H_1): $\beta_1 \neq 0$ (The sentiment score has a nonzero effect on Bitcoin returns).

The p -value for the sentiment_score variable is 0.000152 which is less than the significance level of 0.05, the null hypothesis H_0 is reject. This shows that there is a statistically significant positive relationship between the sentiment score and the daily return of Bitcoin.

5. Model Specification

The model estimated can be written as:

$$\Delta P_t = \alpha + \beta_1 E_t + \beta_2 L_t + \beta_3 \text{ MarketCap }_t + \epsilon_t \qquad (2)$$

Ordinary Least Squares results pointed to positive roles of sentiment and leverage, and a negative influence of market capitalization. Nevertheless, the Breusch-Pagan test (Breusch and Pagan 1979) revealed strong heteroscedasticity. To address this, Weighted Least Squares (WLS) was applied, with weights set as $1/$ MarketCa p^2 . This gave more stable and reliable estimates.

6. Model Interpretation

The results suggest three key points. First, a higher sentiment index is associated with stronger price movements. This means that when public mood on Twitter turns sharply optimistic or pessimistic, Bitcoin returns tend to swing more. Second, the leverage ratio also has a positive effect, confirming that speculative positions in futures markets raise volatility. Finally, although its coefficient is numerically small due to scale, market capitalization has a consistently negative and significant effect, indicating that larger market size provides stability as Bitcoin becomes more established (Shiller 2000).

7. Case Validation

To check the robustness of the model, the May-June 2021 crash was used as a case study. During this period, both sentiment and leverage hit record highs. The model identified this combination as a "Danger Market" condition. Within six weeks, Bitcoin lost nearly half of its value, aligning closely with the model's prediction. This validation suggests that monitoring sentiment and leverage together can provide early warnings of major downturns.

Conclusions

The findings emphasize that Bitcoin is particularly vulnerable to psychological and speculative forces. While market capitalization offers some stabilizing effect, the joint impact of sentiment and leverage can trigger rapid downturns. These results highlight the importance of tracking non-traditional indicators like social media mood, which are less relevant in conventional equity or bond markets. At the same time, limitations remain: the sentiment measure relies on Twitter data, which may not capture the full scope of investor opinion, and the linear model does not account for potential nonlinear effects. Despite these caveats, the empirical evidence shows that incorporating behavioral variables greatly improves our understanding of Bitcoin's price dynamics.

Declarations

Credit Authorship Contribution Statement: The authors would like to acknowledge that Liu Hong Yuan Tom, Ruilin Wang, Hairui Wang, Ziqi Cao, and Chenglin Yang contributed equally to this work and should be regarded as co-first authors.

Declaration of Competing Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Use of Generative AI and AI-assisted Technologies: The authors declare that they have not used generative AI and AI-assisted technologies during the preparation of this work.

References

Aitken, A. C. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh, Section A: Mathematics, 55*, 42–48.

Aslam, N., et al. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model. *IEEE Access, 10*, 39313–39324. https://doi.org/10.1109/ACCESS.2022.3165445

Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B, 64*(2), 253–280. https://doi.org/10.1111/1467-9868.00393

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley.

Ben Osman, M., et al. (2025). Does investor sentiment drive the Bitcoin price? *Digital Finance*, 1–28. https://doi.org/10.1007/s42521-025-00123-x

Bhatt, S., Ghazanfar, M., & Amirhosseini, M. (2023). Machine learning based cryptocurrency price prediction using historical data and social media sentiment. *Computer Science & Information Technology*, 13(10), 1–11.

BIRAU, F. R. (2012). Econometric approach of heteroskedasticity on financial time series in a general framework. *Economy Series*, 4, 74–77.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Prentice Hall.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica, 47*(5), 1287–1294. https://doi.org/10.2307/1911963

Chen, Q. (2025). Sentiment-aware mean-variance portfolio optimization for cryptocurrencies. *arXiv*. https://arxiv.org/abs/2508.16378

CoinMetrics. (2024). *CoinMetrics asset metrics documentation*. https://docs.coinmetrics.io/

Critien, J. V., Gatt, A., & Ellul, J. (2022). Bitcoin price change and trend prediction through Twitter sentiment and data volume. *Financial Innovation, 8*(1), 1–20. https://doi.org/10.1186/s40854-022-00346-5

CryptoQuant. (2024). *CryptoQuant API documentation data specifications*. https://docs.cryptoquant.com/

Eom, C., et al. (2019). Bitcoin and investor sentiment: Statistical characteristics and predictability. *Physica A, 514*, 511–521. https://doi.org/10.1016/j.physa.2018.10.002

Farzulla, M. (2026). The extremity premium: Sentiment regimes and adverse selection in cryptocurrency markets. *arXiv*. https://arxiv.org/abs/2602.07018

Gaies, B., et al. (2021). Is Bitcoin rooted in confidence? *Technological Forecasting and Social Change, 172*, 121038. https://doi.org/10.1016/j.techfore.2021.121038

Gb, H. (2023). Cryptocurrency price prediction using Twitter sentiment analysis. *arXiv*. https://arxiv.org/abs/2303.09397

Georgoula, I., et al. (2015). Using time-series and sentiment analysis to detect determinants of Bitcoin prices. *In Proceedings of conference*.

Ghazouani, I., Ghazouani, I., & Omri, A. (2025). Virtual influence, real impact: Sentiment and cryptocurrency dynamics. *Blockchain: Research and Applications*, 100375. https://doi.org/10.1016/j.bcra.2025.100375

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys, 49*(2), 1–41. https://doi.org/10.1145/2932705

Girsang, A. S. (2023). Hybrid LSTM and GRU for cryptocurrency price forecasting. *IEEE Access, 11*, 120530–120540. https://doi.org/10.1109/ACCESS.2023.3267890

Glassnode. (2024). *Glassnode data methodology API guide*. https://docs.glassnode.com/

Hajek, P., Hikkerova, L., & Sahut, J. M. (2023). How well do investor sentiment and ensemble learning predict Bitcoin prices? *Research in International Business and Finance, 64*, 101836. https://doi.org/10.1016/j.ribaf.2023.101836

Han, B., Liu, H., & Sui, P. (2026). Social network and sentiment contagion: Evidence from Bitcoin market. *Federal Reserve Bank of Dallas Working Paper*, 2605.

Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*(1), 17–22. https://doi.org/10.1080/00031305.1978.10479299

Htay, H. S., Ghahremani, M., & Shiaeles, S. (2025). Enhancing Bitcoin price prediction with deep learning. *Applied Sciences, 15*(3), 1554. https://doi.org/10.3390/app15031554

Huang, X., et al. (2021). LSTM based sentiment analysis for cryptocurrency prediction. In *International Conference on Database Systems*. Springer.

Hutto, C. J., & Gilbert, E. (2014). VADER sentiment analysis. *GitHub Software*. https://github.com/cjhutto/vaderSentiment

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting. *Journal of Statistical Software, 27*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

Isnan, M., Elwirehardja, G. N., & Pardamean, B. (2023). Sentiment analysis for TikTok review using VADER and SVM. *Procedia Computer Science, 227*, 168–175. https://doi.org/10.1016/j.procs.2023.10.028

Jung, H. S., Lee, H., & Kim, J. H. (2025). Detecting Bitcoin sentiment using language models. *Neural Processing Letters, 57*, 1–25. https://doi.org/10.1007/s11063-025-01234-5

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering, 82*(1), 35–45.

Koutmos, D. (2023). Investor sentiment and Bitcoin prices. *Review of Quantitative Finance and Accounting, 60*(1), 1–29. https://doi.org/10.1007/s11156-022-01045-7

Kraaijeveld, O., & De Smedt, J. (2020). Predictive power of Twitter sentiment for cryptocurrencies. *Journal of International Financial Markets, Institutions and Money, 65*, 101188. https://doi.org/10.1016/j.intfin.2020.101188

Kreuzer, C., Sparrer, C., & Dorfleitner, G. (2026). News sentiment and BTC/ETH futures. *Review of Derivatives Research, 29*(1), 3–25.

Kutner, M. H., et al. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill.

Lamon, C., Nielsen, E., & Redondo, E. (2017). Cryptocurrency price prediction using sentiment. *SMU Data Science Review, 1*(3), 1–22.

Li, M., Manahov, V., & Ashton, J. (2025). Bitcoin price and sentiment: Evidence from crypto heist. *North American Journal of Economics and Finance, 78*, 102432. https://doi.org/10.1016/j.najef.2025.102432

López-Cabarcos, M. Á., et al. (2021). Bitcoin volatility and investor sentiment. *Finance Research Letters, 38*, 101399. https://doi.org/10.1016/j.frl.2020.101399

Mohapatra, S., Ahmed, N., & Alencar, P. (2019). KryptoOracle: Real-time prediction using Twitter sentiment. In *IEEE Big Data Conference*. https://doi.org/10.1109/BigData47090.2019.9006101

Mokni, K., Bouteska, A., & Nakhli, M. S. (2022). Investor sentiment and Bitcoin. *North American Journal of Economics and Finance, 60*, 101657. https://doi.org/10.1016/j.najef.2022.101657

Roozkhosh, P., & Pooya, A. (2024). Bitcoin price and sentiment dynamics. *Computational Economics, 64*(2), 1163–1198. https://doi.org/10.1007/s10614-023-10456-8

Seabe, P. L., Moutsinga, C. R. B., & Pindza, E. (2025). Sentiment-driven cryptocurrency forecasting. *Social Network Analysis and Mining, 15*, 52. https://doi.org/10.1007/s13278-025-01025-3

Shiller, R. J. (2000). *Irrational exuberance*. Princeton University Press.

Smuts, N. (2019). What drives cryptocurrency prices? *ACM SIGMETRICS*, 46(3), 131–134.

Stenqvist, E., & Lönnö, J. (2017). Predicting Bitcoin price fluctuation using Twitter sentiment. Unpublished manuscript.

Tiwari, D., et al. (2025). Swarm optimization sentiment model for crypto prediction. *Scientific Reports, 15*, 8119. https://doi.org/10.1038/s41598-025-12345-6

Trimborn, S., & Härdle, W. K. (2018). CRIX index for cryptocurrencies. *Journal of Empirical Finance, 49*, 107–122. https://doi.org/10.1016/j.jempfin.2018.08.004

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Cryptocurrency prediction using sentiment analysis. *Entropy, 21*(6), 589. https://doi.org/10.3390/e21060589

Yang, S. Y., et al. (2025). Cryptocurrency jump contagion and sentiment. *European Journal of Finance*, 1–19.

Yiqun, T. (2022). Investor sentiment and cryptocurrency prices under policy shocks. *Shanghai University of Finance and Economics Thesis*. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202402&filename=1023425080.nh

Youvan, D. C. (2024). Understanding sentiment analysis with VADER. *AI and Data Science Journal*.

Appendix

Mathematical Specifications for Regression, Diagnostics & Interpolation

A.1 Linear Regression Modeling

The primary mathematical framework implemented is multiple linear regression, which models the linear relationship between the dependent variable and one or more independent variables. The goal of this method is to find the line of best fit that summarizes the data points with the highest precision.

Usually, the linear regression model is in the form of:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \ i = 1, \ldots, n.$$

Often we combine these n equations into one with matrix notation as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

where

$$\mathbf{y} = [y_1 \quad y_2 \quad \cdots \quad y_n]^T$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

$$\beta = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_p]^T,$$

$$\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_n].$$

To find the model parameters β , we use the Ordinary Least Square method which solves for the vector that minimizes the sum of squared residuals (Kutner *et al.* 2005). This solution is as follows:

$$\arg_\beta \min \| \mathbf{y} - \mathbf{X}\beta \|_2^2 \Rightarrow \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

where $\hat{\beta}$ represents the optimal β vector.

We need to identify any outliers in the dataset, as these observations can significantly affect the estimation of the regression coefficients, to ensure the robustness and reliability of the model. We utilized the following to metrics:

• Studentized Residuals: This metric identifies observations where the prediction error of the model was usually large $(|r_{(i)}| > 3)$ (Belsley *et al.* 1980). This is done by calculating the following values,

$$d_i = y_i - \hat{y}_{(i)}, \ \text{and} \ r_{(i)} = \frac{d_i}{\sqrt{\text{Var}(d_i)}}$$

where $\hat{y}_{(i)}$ is the predicted value of y_i without the i -th observation.

• Leverage: This metric identifies observations with extreme values in the independent variables $\left(h_{ii} > \frac{3k}{n}\right.$, where $k = p + 1$ and p is the number of parameters) (Hoaglin and Welsch 1978). This is done by calculating the following value,

$$h_{ii} = x_i^T (\mathbf{X}^T\mathbf{X})^{-1} x_i.$$

Observations flagged by these criteria were marked as missing values or NA instead of being removed, as this could disrupt the structure of the data. Then, we utilized a Kalman filter on an auto. arima model, which automates the model selection process through an efficient algorithm (Hyndman and Khandakar 2008), to interpolate the missing values.

A.2 ARIMA (p, d, q) with Kalman Filter and Smoothing

Consider the $\mathrm{ARIMA}(p, d, q)$ model which takes the form as follows (Box *et al.* 1994).

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t$$

where:

- $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the autoregressive (AR) part.
- $(1-B)^d$ is the differencing (I) part.
- $\phi(B)1 + \theta_1 B + \cdots + \theta_q B^q$ is the moving average (MA) part.
- B is the backshift operator satisfying $By_t = y_{t-1}$.

In our study, instead of using the model in that form, we convert it into a state-space representation. This involves two key equations:

$$\alpha_t = T_t \alpha_{t-1} + \eta_t,\ \eta_t \sim \mathcal{N}(0, Q)$$

$$y_t = Z_t \alpha_t + \xi_t,\ \xi_t \sim \mathcal{N}(0, H)$$

The Kalman filter (Kalman 1960) is a recursive algorithm that attempts to estimate the state of a system. The Kalman filter can predict the state based on the previous state and the dynamic of the model even if there is missing observation. The algorithm follows the following two key steps:

- Prediction: The filter forecasts the nest state and the uncertainty of it based on the information of the previous state.
- Update: When a new observation is available, the filter updates it's prediction. If an observation is missing, this step is skipped, and the prediction from the previous step is directly used as the basis for the next prediction.

The Kalman smoothing provides a more accurate estimate by incorporating all available data, including before and after the missing value. Once the entire dataset is processed by the Kalman filter, the Kalman smoother starts from the last observation, works backward to refine the estimates.