

RESEARCH

Open Access



News sentiment analysis using ChatGPT for Bitcoin price dynamics

Piotr Fiszeder^{1,2*} , Witold Orzeszko^{1,4} and Radosław Pietrzyk³

*Correspondence:

Piotr Fiszeder

piotr.fiszeder@umk.pl

¹Faculty of Economic Sciences and Management, Nicolaus Copernicus University in Torun, Gagarina 13a, 87-100 Torun, Poland

²Faculty of Finance and Accounting, Prague University of Economics and Business, Nám. W. Churchilla 1938/4, 13067 Prague 3, Czech Republic

³Faculty of Economics and Finance, Wrocław University of Economics and Business, Komandorska 118-120, 53-345 Wrocław, Poland

⁴Institute of Advanced Studies, Nicolaus Copernicus University, Torun Gagarina 11, 87-100 Torun, Poland

Abstract

This study investigates the use of ChatGPT as an automated tool for extracting and labeling Bitcoin-related news sentiment and examines how the resulting sentiment indicators affect Bitcoin returns and volatility. A large dataset of news headlines is processed via an API-based workflow, and the ChatGPT-derived sentiment indicators are subsequently incorporated as explanatory variables into selected statistical and machine learning models, including autoregressive (AR), heterogeneous autoregressive (HAR), Bayesian model averaging (BMA), least absolute shrinkage and selection operator (LASSO), and support vector regression (SVR). We find that while the sentiment indicators significantly improve in-sample estimation accuracy for returns and volatility, they do not lead to statistically significant gains in out-of-sample forecasting performance. This result suggests that ChatGPT-based sentiment measures primarily capture contemporaneous market-relevant information rather than persistent predictive signals, consistent with semi-strong market efficiency.

Keywords Large language models, LASSO, SVR, HAR, Machine learning, Cryptocurrencies, Bitcoin returns, Volatility

Introduction

The rapid development of artificial intelligence (AI) has sparked significant interest in applying AI tools across various areas of life. One of the key domains where AI methods have garnered considerable attention is finance, primarily due to their ability to process and analyze large volumes of data at remarkable speed. Large language models (LLMs) play a pivotal role in this context. These models are trained to understand and generate human-like responses based on vast amounts of text.

The use of LLMs has emerged as a major topic of interest within the scientific community. Dong et al. [12] and Du et al. [14] identified several hundred papers focusing on LLMs in the fields of accounting, finance, and economics. Comparative analyses suggest that LLMs often outperform traditional methods in financial applications across diverse markets, though their relative superiority remains strictly dependent on the specific nature of the application. In particular, LLMs perform very well in tasks involving high-dimensional unstructured data processing and natural language processing (NLP), such as sentiment analysis, financial report summarization, and news information extraction.

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

However, for time series modelling and forecasting, traditional econometric approaches often remain more reliable and computationally efficient.

Among its many applications, sentiment analysis stands out as a valuable tool for evaluating textual data's tone and emotional underpinnings. Research has demonstrated the potential of AI models to perform sentiment analysis for descriptive purposes and predictive insights (see e.g. Ahamad and Mishra [1], Drašković and Milanović [13], Gao et al. [23], Qin et al. [51]). Beyond standard categorizations such as "positive," "negative," and "neutral," these systems could be refined to generate more nuanced sentiment classifications, expanding their utility across various domains.

Among the most well-known LLMs is ChatGPT. One of the key applications of ChatGPT in finance is its use in investing in financial markets. ChatGPT's role in investment decisions encompasses a range of activities, including forecasting earnings and returns, sentiment analysis, asset valuation, asset selection, portfolio management, risk analysis, trading strategies, and financial advisory. Research studies on these topics are reviewed in Sect. "Literature review". The adoption of ChatGPT typically results in increased efficiency of the strategies analyzed.

An important capability of LLM models is their ability to review large sets of information and extract valuable insights. The studies presented in Sect. "Literature review" highlight ChatGPT's efficiency in extracting sentiment from financial texts, which can be utilized to predict market trends or enhance existing financial models. In this study, we analyze ChatGPT's ability to synthesize financial news and evaluate its effectiveness in identifying actionable investment insights. We construct a comprehensive dataset of news headlines relevant to Bitcoin (BTC) and use ChatGPT to assess the information in each headline. These responses are converted into numerical indicators, which are then utilized in selected statistical and machine learning (ML) models to analyze the returns and variances of BTC. The primary objective of the study is to construct ChatGPT-derived sentiment indicators and to assess their usefulness in modelling and forecasting Bitcoin returns and volatility, using both in-sample estimation and out-of-sample prediction.

Despite the rapidly growing literature on the use of large language models in finance, several important gaps remain regarding how LLM-based sentiment can be systematically integrated into quantitative asset pricing and volatility frameworks, particularly in cryptocurrency markets. This study addresses these gaps through the following contributions.

First, we propose and empirically evaluate a structured hybrid research design that combines ChatGPT-based news sentiment extraction with a broad set of established modelling frameworks. Rather than treating ChatGPT as a forecasting device, we explicitly use it as an automated, scalable sentiment-labeling tool whose outputs are embedded into autoregressive (AR), heterogeneous autoregressive (HAR), Bayesian model averaging (BMA), least absolute shrinkage and selection operator (LASSO), and support vector regression (SVR) models. This design advances prior work by moving beyond single-model or purely regression-based settings and by systematically comparing how LLM-derived sentiment interacts with different modelling paradigms.

Second, the study introduces a multi-dimensional sentiment construction tailored to short-term cryptocurrency market dynamics. Instead of focusing exclusively on return direction, we distinguish between sentiment related to price changes, volatility changes,

and volatility levels. This separation allows us to disentangle contemporaneous informational content from predictive effects and to examine how different sentiment dimensions are absorbed by returns and realized volatility, which has been largely overlooked in existing ChatGPT-based sentiment studies.

Third, we contribute to the still limited literature on large language models in cryptocurrency markets. While most existing studies focus on equity markets or social-media sentiment and rely on relatively simple linear specifications, empirical evidence for cryptocurrencies – especially beyond return predictability – remains scarce. By focusing on Bitcoin and jointly analyzing returns and volatility across multiple modelling frameworks, our study extends the scope of LLM-based sentiment research to a market characterized by high volatility, rapid regime shifts, and rapid information assimilation.

Finally, our results provide new economic insights into the role of LLM-derived news sentiment in forecasting. We find that although ChatGPT-based sentiment indicators significantly improve in-sample explanatory power, they do not translate into robust out-of-sample forecasting gains. Rather than viewing this as a methodological shortcoming, we interpret it as evidence that such sentiment measures primarily capture contemporaneous or quickly assimilated information. This finding helps reconcile mixed evidence in the literature and highlights the differing roles of LLM-based sentiment in in-sample modelling versus out-of-sample forecasting.

The remainder of the paper is organized as follows. Sect. "[Literature review](#)" reviews studies on the application of ChatGPT in financial market investments. In Sect. "[Data and methods](#)", we describe the data, the construction of the ChatGPT-derived sentiment indicators, and the modelling framework, including the research procedures and methods applied. Sect. "[Empirical results](#)" reports the results of the analysis, followed by a discussion and an overview of the methodological limitations. The final section concludes the paper.

Literature review

Although ChatGPT is not trained to predict the prices of financial assets, most research findings suggest that using ChatGPT has a beneficial effect on investment outcomes. Research on the use of ChatGPT in investing can be divided into two main groups. The first group consists of descriptive or conceptual studies suggesting possible advantages of adopting ChatGPT, typically supported by small-scale experiments using data generated by researchers (e.g. Niszczoła and Abbas [46], Oehler and Horn [47]). The second group concentrates on historical studies, exploring real-life situations in which researchers have implemented ChatGPT in their workflows. These analyses aim to identify differences between ChatGPT users and non-users, providing concrete insights into the effects of its adoption in financial markets. Such studies are usually conducted on large research samples using statistical models, with most of the analyses presented below falling into this category.

Table 1 summarizes key empirical studies using ChatGPT in financial markets, highlighting differences in data sources, asset classes, modelling approaches, and evaluation settings. This overview facilitates a structured comparison of existing contributions and helps identify common methodological patterns and open research questions.

Most studies highlight significant gains from the application of ChatGPT in investment strategies but also stress the need for caution and thoughtful consideration of the

Table 1 Research papers examining the application of ChatGPT in investment strategies within financial markets

Reference	Assets	Applied methods	Main findings
Bond et al. [3]	Stocks	Panel regression model, SR	Negative correlation between news sentiment indicator from ChatGPT and short-term market returns
Chen et al. [4]	Stocks	Panel regression model, long-short portfolio, ridge regression, LASSO, random forest, neural networks, SR	Predictions from LLMs, including i.a. ChatGPT, improve over leading technical signals or simpler NLP methods
Chen et al. [5]	Stocks	OLS regression model, SR	News classified by ChatGPT as good can predict the stock market and are related to macroeconomic conditions
Chen et al. [6]	Stocks	Panel regression model	The sentiment scores generated by ChatGPT from corporate financial statements can predict firms' risk-management capabilities and stock return performance
Cheng et al. [7]	Stocks	Long-short portfolio, OLS regression model, SR	ChatGPT-generated financial factors based on financial statements effectively predict company performance across various industries
Chiu et al. [8]	Stocks	VAR, Granger causality, Shannon and effective transfer entropy	Stocks receiving higher buying scores from ChatGPT tend to outperform those with lower scores. The strategy does not outperform the market
Chou et al. [9]	Stocks	Pearson correlation, risk return analysis, ARMA-GARCH models, SR	ChatGPT performance is comparable to that of traditional statistical software, such as Stata
Cohen and Aiche [10]	Bitcoin	Trading algorithm based on technical analysis, macroeconomic indicators, market sentiment, SR	Chat-GPT-driven algorithm outperforms both ML-based and buy and hold strategy
Fatouros et al. [17]	Stocks	Bootstrapping, hit ratio, SR, Sortino ratio, win rate, standard deviation	An investment service based on ChatGPT delivers an excess alpha of 10–30% and achieves a cumulative return of up to 72% over a 15-month period while maintaining a risk profile comparable to that of the broader market
Fatouros et al. [18]	Currencies	Pearson correlation, directional accuracy	ChatGPT exhibited approximately 35% enhanced performance in sentiment classification and 36% higher correlation with market returns compared to FinBERT
Fieberg et al. [19]	Financial products	Six factor regression model, SR	ChatGPT can provide proper financial advice and suggests specific investment portfolios that reflect an investor's individual characteristics such as risk capacity, preference and risk tolerance
Han et al. [25]	Currencies	RNN, LSTM, LSTM-Attention, GRU, LEST (LLM-EMD-SG-TL)	More accurate predictions using sentiment indices derived from ChatGPT
Jeong et al. [30]	Lithium	Impulse response function, Diebold and Yilmaz spillover index, artificial neural network, OLS regression model	Impulse response analysis reveals a significant impact of ChatGPT scores from news on the lithium market. Bidirectional spillover transmission exists between the ChatGPT sentiment indicator and the lithium market, ChatGPT scores acts as net receivers
Kang et al. [31]	Cryptocurrencies	VAR, OLS regression model, factors and quintile portfolios, trading strategies with long and short positions	Sentiment is strongly correlated with contemporaneous cryptocurrency market returns. Cryptocurrency prices significantly influence ChatGPT-derived news sentiment, not vice versa
Kang et al. [32]	Bitcoin	GARCH, KNN regression, VaR	Inclusion in KNN regression sentiment indicator from Chat GPT improves VaR forecasts

Table 1 (continued)

Reference	Assets	Applied methods	Main findings
Kim [33]	Exchange-traded funds	Mean–variance framework, SR, VaR, CVaR	Asset class recommendations based on ChatGPT's economic indicators can improve portfolio efficiency
Kirtac and Germano [34]	Stocks	Panel regression model, SR	LLMs identify sentiment, enhance forecasting and portfolio management. ChatGPT model yields the most profitable investments
Ko and Lee [35]	Stocks, cryptocurrencies, commodities, currencies, bonds	Simpson and Shannon diversity indices, average correlation, mean–variance framework, SR	Portfolios based on ChatGPT's selections outperform portfolios built on randomly selected assets. ChatGPT's selections are significantly better in diversity index than randomly selected assets
Li et al. [37]	Stocks	Panel regression model	ChatGPT can correct the optimistic biases of human analysts
LoGrasso [41]	Stocks	Panel regression model, CAPM model, 3- and 5-factors models	ChatGPT model was able to average alphas of approximately 1% per month for two-year holding periods every year from 1985 to 2021, outperforming traditional methods
Lopez-Lira et al. [42]	Stocks	Panel regression model, CAPM model, 5-factor model, SR	Significant correlation between ChatGPT scores (based on news headlines) and subsequent daily stock returns, outperforming traditional methods
Ma et al. [43]	Stocks	OLS regression model	Predictive scores from ChatGPT positively and significantly forecast equity premiums
Nguyen et al. [45]	Bitcoin	OLS regression models	The sentiment indicator derived by ChatGPT from the X platform has a significant impact on BTC returns
Niszczoła and Abbas [46]	Financial products	Financial Literacy Baseline Survey, OLS and beta regression models	ChatGPT achieves a near-perfect 99% score on the financial literacy test and can successfully serve as a financial robo-advisor for the masses. People with lower subjective financial knowledge rely more heavily on advice from ChatGPT
Oehler and Horn [47]	Financial products	Survey for three investor profiles with different risk attitudes	ChatGPT provides better financial advice for one-time investments than robo-advisors. ChatGPT's recommendations align with the benchmark, only three out of the 17 robo-advisors come close to it
Pele et al. [48]	Indices for cryptocurrencies, stocks, clean energy, bonds, and commodities	GARCH, EWMA with Dynamic Conditional Score, zero-shot forecasting approach	LLMs (GPT-3.5, GPT-4 and GPT-4o) perform well in VaR and ES forecasting when short rolling windows are used. GPT-3.5 frequently outperforms or matches the performance of newer models. Econometric models prove more reliable for longer windows
Pelster and Val [49]	Stocks	Cross-sectional OLS regressions, SR	ChatGPT's earnings forecast significantly correlate with actual earnings. ChatGPT's ratings significantly correlate with future stock returns
Romanko et al. [53]	Stocks	Mean–variance framework, VaR, SR	ChatGPT is effective in stock selection but may not perform as well in assigning optimal weights to stocks within the portfolio. The best results are achieved when stock selection by ChatGPT is combined with portfolio optimization models
Yang et al. [57]	Currencies	Interval multilayer perceptron, interval extreme learning machine, autoregressive conditional interval, threshold autoregressive interval	Topic importance scores generated by ChatGPT enhance the forecasting accuracy of interval models for exchange rates. The superiority of ChatGPT over LDA and BERT in both topic extraction and importance assessment is demonstrated

Notes: The table summarizes selected studies that examine the use of ChatGPT in financial market applications. It reports the main research focus, the market or asset class considered, the role of ChatGPT, and the key methodological approaches and findings. Abbreviations refer to the methods used in the cited studies: ARMA: autoregressive moving average; BERT: Bidirectional Encoder Representations from Transformers; CAPM: capital asset pricing model; CVaR: conditional value at risk; EWMA: exponentially weighted moving average; FinBERT: Financial BERT; GARCH: generalized autoregressive conditional heteroskedasticity; GRU: gated recurrent unit; KNN: k-nearest neighbors; LDA: latent Dirichlet allocation; LSTM: long short-term memory; NLP: natural language processing; OLS: ordinary least squares; RNN: recurrent neural networks; SR: Sharpe ratio; VaR: value at risk; VAR: vector autoregression

risks associated with adopting these new technologies. These may include privacy and security concerns, legal considerations, ambiguity in decision-making processes, labor displacement, and information inaccuracy (see more in Dong et al. [12]). Romanko et al. [53] emphasize that, due to the mistakes or hallucinations made by ChatGPT, its recommendations should be subjected to robust validation checks. Pelster and Val [49] point out that the performance of ChatGPT varies over time. Niszczoła and Abbas [46] highlight the risk of overreliance on LLMs, emphasizing that people may conflate the confidence of these models with their actual competence as advisors. According to Chen et al. [6], ChatGPT is responsive to variations in the prompt structure, which makes it important to use effective strategies to address its sensitivity.

The reviewed literature demonstrates rapidly growing interest in the application of ChatGPT and other LLMs in financial markets. Existing studies consistently show that ChatGPT-derived textual indicators contain economically meaningful information and can improve in-sample model fit, portfolio selection, or risk assessment, particularly in equity markets.

Importantly, while out-of-sample evaluations are present in the literature, they are predominantly conducted in the context of trading strategies or portfolio performance rather than within a systematic econometric forecasting framework for returns and volatility. Moreover, a substantial share of existing studies relies on relatively simple empirical tools, such as linear models estimated by OLS or panel regression models, with fewer contributions exploring a broader set of statistical and predictive modelling approaches.

Notably, a large share of prior ChatGPT-based sentiment studies relies on headline-level news text, particularly in large-scale empirical settings, reflecting a trade-off between scalability and informational depth. Applications to cryptocurrency markets remain relatively scarce and are largely limited to return predictability, with volatility dynamics receiving considerably less attention.

Against this background, the present study contributes to the literature by examining ChatGPT-extracted news sentiment in the context of Bitcoin returns and volatility, explicitly distinguishing between in-sample explanatory power and out-of-sample forecasting performance across a broad set of predictive models. Our study differs significantly from traditional sentiment analysis (see, e.g., Birjali et al. [2]), since we do not limit ourselves to returns but also examine the impact on volatility.

Data and methods

The paper introduces novel sentiment indicators derived from ChatGPT. The news headlines relevant to Bitcoin are analyzed by ChatGPT, which assesses their sentiment based on three predefined criteria. The predictive relevance of the proposed indicators is subsequently evaluated in the context of modeling and forecasting Bitcoin returns and volatility using both in-sample and out-of-sample approaches. This general empirical research design is illustrated in the flowchart shown in Fig. 1.

Data and prompt design

We use two types of data: the first consists of BTC prices obtained from the cryptocurrency exchange Kraken (<https://kraken.com>), while the second includes news headlines from the Crypto News API database (<https://cryptonews-api.com>) relevant to BTC. Both datasets cover the period from January 1, 2021, to March 31, 2024.

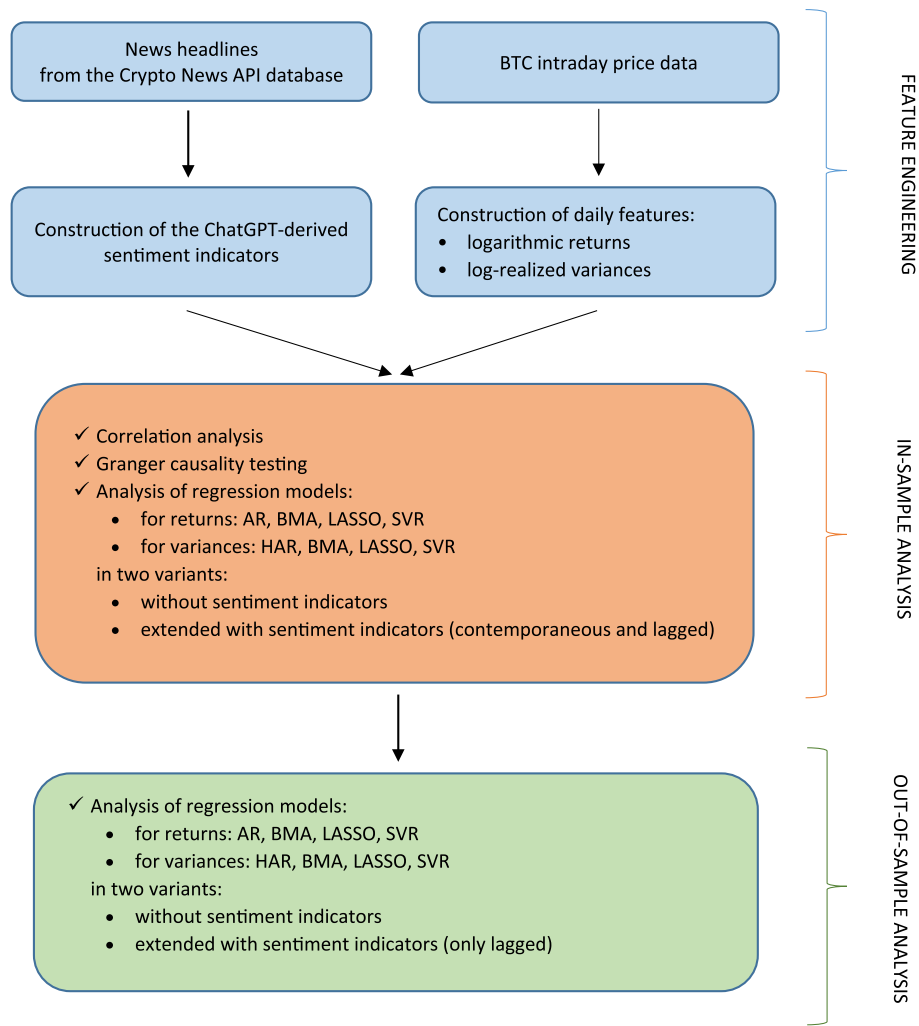


Fig. 1 Flowchart of the empirical research framework

The selected sample period encompasses both bull and bear market phases in the Bitcoin market, as illustrated in Fig. 2. This allows the analysis to capture heterogeneous market conditions and enhances the generalizability of the empirical findings.

Using BTC prices, we calculate daily logarithmic returns $r_{d,t} = \ln(P_t/P_{t-1})$, where P_t is the closing price on date t . In order to measure the variance of returns, we use intraday data and calculate realized variance:

$$RV_{d,t} = \sum_{k=1}^K r_{k,t}^2, \quad (1)$$

where $r_{k,t}$ is the within-day return, K is the number of observations during a day. For calculations of realized variance, we utilize 15-min price intervals.

Additionally, we apply weekly and monthly average realized variances defined as:

$$RV_{w,t} = \frac{RV_{d,t-6} + RV_{d,t-5} + \dots + RV_{d,t}}{7}, \quad (2)$$

$$RV_{m,t} = \frac{RV_{d,t-29} + RV_{d,t-28} + \dots + RV_{d,t}}{30}. \quad (3)$$

Table 2 presents descriptive statistics for the variables described above.



Fig. 2 Bitcoin price dynamics over the sample period, with bull and bear market phases separated by vertical red lines

Table 2 Summary statistics of the investigated variables

Variable	Mean	Minimum	Maximum	Standard deviation	Skewness	Excess kurtosis
r_d	0.748E-03	−0.164	0.147	0.034	−0.265	2.938
RV_d	0.130E-02	0.727E-05	0.045	0.002	8.832	118.031
RV_w	0.130E-02	0.691E-04	0.016	0.002	4.637	28.234
RV_m	0.131E-02	0.120E-03	0.007	0.001	2.455	6.234

Notes: The table presents summary statistics for daily logarithmic returns r_d and realized variances of Bitcoin. RV_d denotes the daily realized variance, while RV_w and RV_m denote the weekly and monthly average realized variances, respectively. The analyzed period is January 1, 2021—March 31, 2024

The obtained statistics are consistent with those of other studies (see, e.g. Dudek et al. [15]; Fiszeder et al. [20]) and confirm the presence of very high volatility. The findings also suggest that forecasting BTC is a challenging task. The Crypto News API database is used to construct ChatGPT-derived sentiment indicators. It comprises 113,781 headlines related to Bitcoin and aggregates articles and video headlines from more than 30 crypto-focused news sources such as CoinTelegraph, Coindesk, NewsBTC, Decrypt, and CoinMarketCap as well as leading global financial outlets, including Reuters, Bloomberg, and Forbes. These sources include both mainstream and specialist outlets covering global cryptocurrency developments, contributing to a broad and representative news sample. While more traditional financial news feeds (e.g., Reuters, Bloomberg) provide valuable context, they often contain limited crypto-specific content. In contrast, the Crypto News API’s pool of sources ensures both volume and relevance for Bitcoin-focused sentiment analysis.

We apply preprocessing steps to remove exact duplicate headlines to avoid overcounting repeated items. Low-information items (e.g., headlines with minimal text) are retained in the raw dataset but contributed proportionally less to the aggregated sentiment indicators, as aggregation methods mitigate the impact of low-content headlines. Conflicting signals across queries are resolved at the indicator construction stage by categorizing each headline independently for each query objective. The headlines from the Crypto News are subsequently analyzed using ChatGPT 4.0. For this purpose, we construct three queries to identify information that indicates:

- 1. a potential short-term price increase or decrease,
- 2. a potential short-term increase or decrease in volatility,

3. high or low short-term volatility.

To evaluate individual headlines, we formulate the following queries:

Query 1: *"Pretend you are a financial expert with cryptocurrency recommendation experience. Elaborate with one short and concise sentence. Can the headline (column 'title' in the attached file) cause a 'price increase' or 'price decrease' for Bitcoin in the short term? Answer 'YES' if it is 'price increase' news, 'NO' if it is 'price decrease' news, or 'UNKNOWN' if the news is uncertain."*

Query 2: *"Pretend you are a financial expert with cryptocurrency recommendation experience. Elaborate with one short and concise sentence in the next column. Can the headline (column 'title' in the attached file) cause a 'volatility increase' or 'volatility decrease' for Bitcoin in the short term? Answer 'YES' if it is 'volatility increase' news, 'NO' if it is 'volatility decrease' news, or 'UNKNOWN' if the news is uncertain."*

Query 3: *"Pretend you are a financial expert with cryptocurrency recommendation experience. Elaborate with one short and concise sentence in the next column. Does the headline (column 'title' in the attached file) mean 'high volatility' or 'low volatility' for Bitcoin in the short term? Answer 'YES' if it is 'high volatility' news, 'NO' if it is 'low volatility' news, or 'UNKNOWN' if the news is uncertain."*

The first query follows the design proposed by López-Lira and Tang [42], who demonstrate that LLM-based classification of news into expected short-term price increases or decreases contains predictive information for asset returns. Queries 2 and 3 are designed to capture two conceptually distinct dimensions of volatility-related information: (i) *directional news about changes in volatility* and (ii) *news describing the expected level of volatility*, regardless of its recent dynamics.

While an expected increase in volatility may lead to high volatility, these two concepts are not equivalent. For example, volatility may increase from a very low base and still remain moderate, or decrease while remaining high. The use of two separate volatility-related queries is intentional and serves as an internal sensitivity check to assess whether LLM-based sentiment extraction is robust to alternative semantic formulations of volatility-related information.

We do not interpret ChatGPT as a forecasting model, but as a high-dimensional text encoder that maps unstructured news headlines into structured indicators based on pre-defined semantic questions. Although prompt design inevitably introduces subjectivity, this subjectivity is fixed ex ante and applied consistently across the entire sample, ensuring internal coherence and reproducibility. Moreover, all prompts are explicitly reported in the paper, and the sentiment extraction procedure is fully deterministic conditional on the LLM version, ensuring transparency of the analysis.

To empirically assess whether Queries 2 and 3 capture distinct informational constructs, we examine the correlation structure and descriptive statistics of the resulting sentiment indicators (see Table 5), defined in Sect. "[Construction of the ChatGPT-derived sentiment indicators](#)". The correlation between the indicators capturing volatility decrease and low volatility is only 0.117, indicating very limited overlap. While the correlation between the 'volatility increase' and 'high volatility' categories is relatively high (0.734), they differ substantially in terms of average daily headline counts (18.704 versus 8.155, respectively; see Table 3. This suggests that ChatGPT more frequently identifies news signaling a change in volatility than news describing a high-volatility state. Overall, these results support the view that Query 2 captures dynamic,

Table 3 Summary statistics for the daily total number of news headlines and ChatGPT query-based counts

Headlines	Mean	Minimum	Maximum	Standard deviation
Total	95.937	19	386	47.886
Price decrease	7.080	0	74	7.714
Price increase	9.016	0	77	9.135
Price direction – neutral	79.841	16	346	39.122
Volatility decrease	7.262	0	24	4.392
Volatility increase	18.704	0	147	13.591
Volatility direction – neutral	69.970	13	323	34.808
Volatility low	5.799	0	26	4.304
Volatility high	8.155	0	66	7.592
Volatility level – neutral	81.982	17	345	39.888

Notes: The table reports summary statistics for the daily number of news headlines and the categorical classifications provided by ChatGPT. The "Total" variable represents the aggregate daily volume of Bitcoin-related news headlines retrieved via the Crypto News API. The remaining rows provide daily counts of headlines assigned by ChatGPT to specific categories based on three distinct queries: price direction (increase, decrease, neutral), volatility direction (increase, decrease, neutral), and volatility level (high, low, neutral). The analyzed period is January 1, 2021—March 31, 2024

shock-related information, whereas Query 3 reflects a more static volatility regime, and that the two prompts are not empirically redundant.

Regarding the operational process, the data is fed into the LLM in batches rather than headline-by-headline or as a single chunk. This batching approach is adopted to remain within the model's context window limits and to optimize processing efficiency, ensuring consistent analysis across the entire dataset while managing API throughput.

Table 3 presents basic information on the daily total number of news headlines and the daily counts of headlines identified by ChatGPT as significant for individual categories.

On average, 95.937 BTC-related news items were posted daily in the Crypto News database. The lowest number of items posted in a single day was 19, while the highest was 386. Considering that our analysis includes Saturdays and Sundays, these figures represent a substantial volume of information, indicating that our database should be a reliable source of BTC-related data.

When asked about BTC price decreases or increases in the short term, ChatGPT identifies an average of 7.080 and 9.016 items of information per day, respectively. The highest average daily number of significant items identified by ChatGPT concerned, the expected increase in BTC volatility in the short term, amounting to 18.704. Conversely, the lowest average daily number of significant items was related to the low expected volatility of BTC in the short term, amounting to 5.799.

Figure 3 presents the daily total counts of all news headlines devoted to BTC and their respective categories throughout the entire period under investigation.

Figure 3 illustrates a pronounced weekly pattern in the volume of BTC-related news. The average daily number of headlines on weekdays (Monday–Friday) equals 115.43, compared to 47.43 on weekends (Saturday–Sunday), indicating substantially lower news activity during weekends. Moreover, a noticeable increase in the total daily volume of information is observed at the end of the period under study, i.e., at the end of 2023 and the beginning of 2024. This rise is also apparent in the daily counts of headlines for individual categories related to the expected increase in BTC price, increase in volatility, and high volatility.

As a robustness check, we examine the consistency between the ChatGPT-derived sentiment indicators and the sentiment labels provided by the Crypto News database

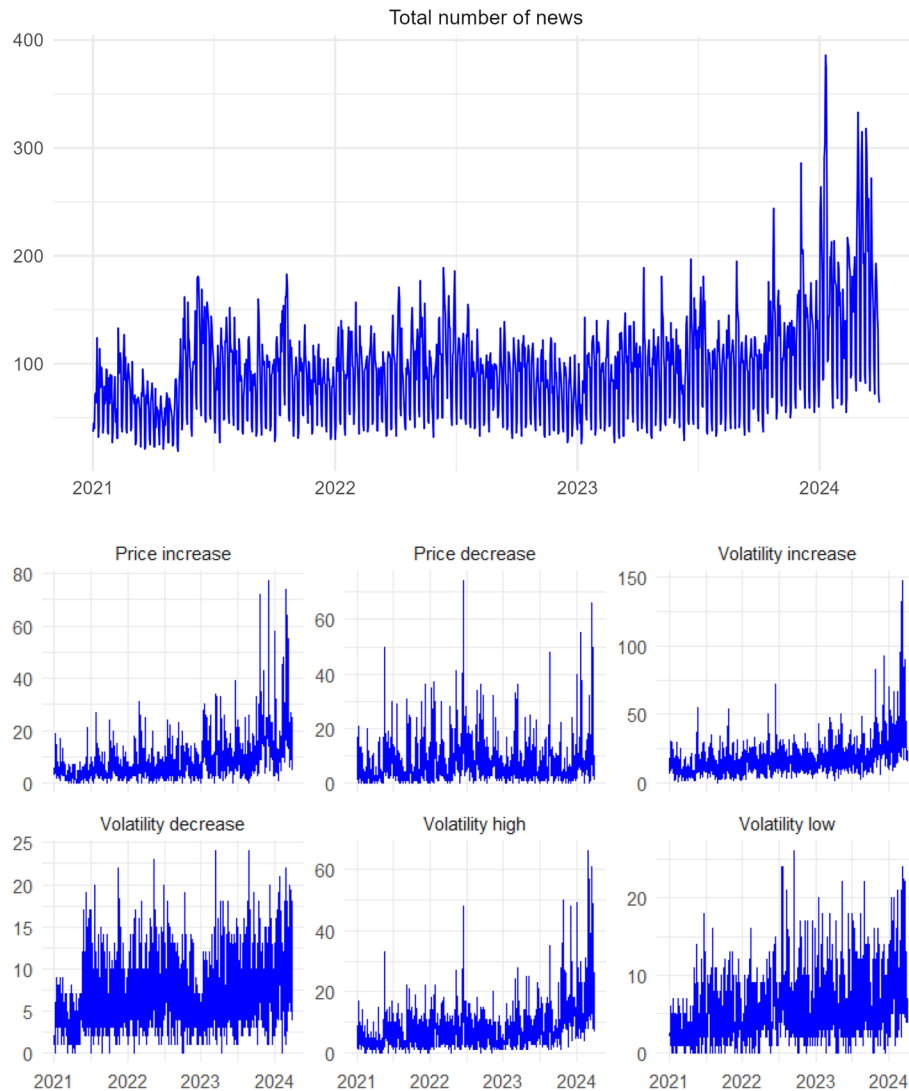


Fig. 3 Daily total number of the BTC-related news headlines (above) and the daily counts of the headlines identified by ChatGPT as significant for individual categories (below)

using a chi-square test of independence based on a three-category classification ($-1, 0, +1$). The null hypothesis of independence is strongly rejected for all three queries (Query 1: $\chi^2 = 2507.6$; Query 2: $\chi^2 = 1135.3$; Query 3: $\chi^2 = 14088.5$; p-values ≈ 0), indicating a statistically significant association between the two sentiment measures.

Inspection of the contingency tables (reported in Table 20 in Appendix A) shows that the highest degree of agreement occurs for news items classified as neutral by both approaches, while non-neutral classifications also display systematic co-movement. These results suggest that the ChatGPT-based sentiment indicators are meaningfully related to an independent sentiment proxy, while not being mechanically identical to it.

Construction of the ChatGPT-derived sentiment indicators

In the next stage, we aggregate the ChatGPT-generated responses and construct the final sentiment indicators. Specifically, for Query 1 (price change), the indicator P^{chg} is defined, for Query 2 (volatility change) the indicator V^{chg} , and for Query 3 (volatility level) the indicator V^{lvl} , according to the following formulas:

$$P^{chg} = \frac{\text{number of 'price increase' news} - \text{number of 'price decrease' news}}{\text{total number of news}},$$

$$V^{chg} = \frac{\text{number of 'volatility increase' news} - \text{number of 'volatility decrease' news}}{\text{total number of news}},$$

$$V^{lvl} = \frac{\text{number of 'high volatility' news} - \text{number of 'low volatility' news}}{\text{total number of news}}.$$

The proposed indicators represent the relative difference between the number of headlines tagged as ‘Yes’ and those tagged as ‘No’, normalized by the total number of news headlines.

Additionally, we disaggregate above indicators by creating unidirectional indicators:

$$P_{inc}^{chg} = \frac{\text{number of 'price increase' news}}{\text{total number of news}},$$

$$P_{dec}^{chg} = -\frac{\text{number of 'price decrease' news}}{\text{total number of news}},$$

$$V_{inc}^{chg} = \frac{\text{number of 'volatility increase' news}}{\text{total number of news}},$$

$$V_{dec}^{chg} = -\frac{\text{number of 'volatility decrease' news}}{\text{total number of news}},$$

$$V_{high}^{lvl} = \frac{\text{number of 'high volatility' news}}{\text{total number of news}},$$

$$V_{low}^{lvl} = -\frac{\text{number of 'low volatility' news}}{\text{total number of news}}.$$

As a result, we obtain nine sentiment indicators in total, calculating their values on a daily basis.

As a robustness check, we construct alternative sentiment indicators in which individual headlines are weighted by their length. The Spearman rank correlations between the proposed baseline and length-weighted sentiment indicators range from 0.98 to 0.99, indicating that the aggregation scheme has a negligible impact on the resulting sentiment dynamics.

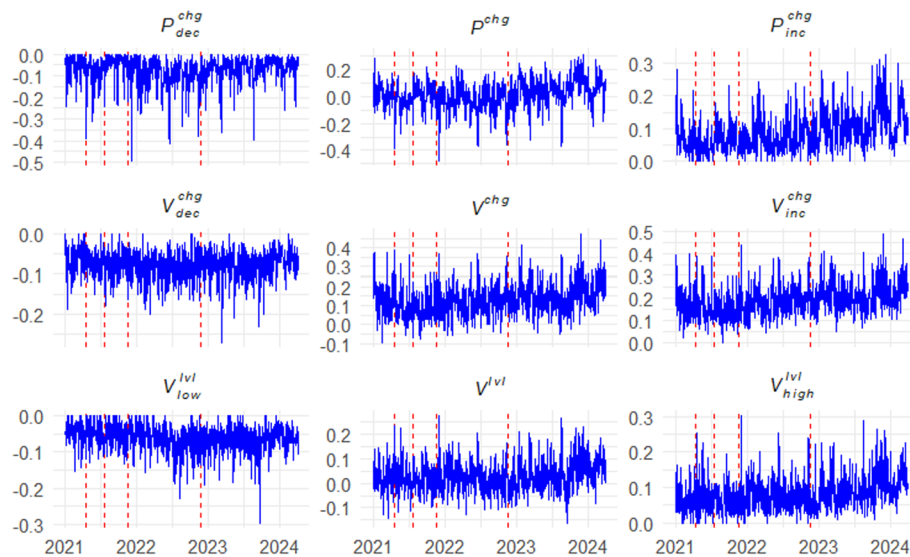
Table 4 presents the descriptive statistics for the nine sentiment indicators derived from ChatGPT.

The distribution of the variable P^{chg} , representing the sentiment indicator for expected prices, differs significantly from those of the variables V^{chg} and V^{lvl} , which represent indicators for expected volatility. The distribution of P^{chg} exhibits strong left-skewness and substantially higher kurtosis. The distributions of P_{dec}^{chg} , V_{dec}^{chg} , and V_{low}^{lvl} show strong left-sided asymmetry in contrast to P_{inc}^{chg} , V_{inc}^{chg} , and V_{high}^{lvl} , which exhibit strong right-sided asymmetry. The sentiment variables exhibit moderate skewness and kurtosis, comparable to those commonly observed in financial time series. Therefore, no transformations are applied in order to preserve interpretability. The greatest variability in the indicators is observed for P^{chg} and V^{chg} , and the smallest for V_{dec}^{chg} and V_{low}^{lvl} .

Table 4 Summary statistics of the sentiment indicators

Variable	Mean	Minimum	Maximum	Standard deviation	Skewness	Excess kurtosis
P^{chg}	0.020	-0.479	0.316	0.101	-0.505	1.443
V^{chg}	0.129	-0.095	0.471	0.082	0.522	0.562
V^{lvl}	0.025	-0.162	0.274	0.060	0.554	0.751
P_{dec}^{chg}	-0.073	-0.493	0.000	0.064	-1.852	5.161
P_{inc}^{chg}	0.089	0.000	0.329	0.058	1.043	1.030
V_{dec}^{chg}	-0.077	-0.273	0.000	0.035	-0.653	1.180
V_{inc}^{chg}	0.192	0.000	0.485	0.070	0.556	0.521
V_{low}^{lvl}	-0.060	-0.297	0.000	0.035	-1.009	2.663
V_{high}^{lvl}	0.081	0.000	0.301	0.047	1.001	1.410

Notes: The table reports summary statistics for the daily sentiment indicators derived from ChatGPT. The analyzed period is January 1, 2021—March 31, 2024

**Fig. 4** Sentiment indicator values over time

To assess the stability of sentiment indicators across market regimes, we compute descriptive statistics separately for bull and bear market subperiods. Overall, the distributions are broadly comparable across regimes, suggesting that most sentiment measures are not driven by specific market phases. One notable exception is the price-change sentiment indicator P^{chg} , whose average value switches sign between bull and bear periods, reflecting a higher prevalence of price-increase (price-decrease) news during market upturns (downturns), which is consistent with economic intuition (see Table 21 in Appendix A). The dynamics of the constructed sentiment indicators are presented in Fig. 4.

Notes: Red dashed vertical lines indicate the starting points of bull and bear market periods; their exact duration for BTC is shown in Fig. 2.

The figure shows very high variability across all the sentiment indicators, with a significant number of outlier values also visible. The average values of the variables remains stable, except at the end of 2023 when an increase is observed in P^{chg} , V^{chg} , V^{lvl} , P_{inc}^{chg} , V_{inc}^{chg} , and V_{high}^{lvl} .

Table 5 presents the estimates of correlation coefficients between the indicators.

Table 5 Correlation coefficients between the sentiment indicators

Variable	P^{chg}	V^{chg}	V^{lvl}	P^{chg}_{inc}	P^{chg}_{dec}	V^{chg}_{inc}	V^{chg}_{dec}	V^{lvl}_{high}	V^{lvl}_{low}
P^{chg}	1								
V^{chg}	0.078 (0.007)	1							
V^{lvl}	0.041 (0.156)	0.582 (0.000)	1						
P^{chg}_{inc}	0.818 (0.000)	0.460 (0.000)	0.377 (0.000)	1					
P^{chg}_{dec}	0.825 (0.000)	-0.318 (0.000)	-0.290 (0.000)	0.359 (0.000)	1				
V^{chg}_{inc}	0.072 (0.013)	0.922 (0.000)	0.613 (0.000)	0.500 (0.000)	-0.365 (0.000)	1			
V^{chg}_{dec}	0.033 (0.251)	0.386 (0.000)	0.063 (0.030)	0.013 (0.658)	0.041 (0.156)	0.034 (0.243)	1		
V^{lvl}_{high}	0.040 (0.170)	0.666 (0.000)	0.841 (0.000)	0.456 (0.000)	-0.368 (0.000)	0.734 (0.000)	0.000 (0.992)	1	
V^{lvl}_{low}	0.009 (0.768)	0.064 (0.027)	0.555 (0.000)	-0.003 (0.927)	0.021 (0.464)	0.020 (0.500)	0.117 (0.000)	0.035 (0.233)	1

Notes: The table reports the Pearson correlation coefficients between the daily sentiment indicators. The p-values for the null hypothesis that the correlation coefficient is zero are provided in parentheses. Correlation coefficients that are statistically significant at the 10% significance level are highlighted in bold. The analyzed period is January 1, 2021—March 31, 2024

The variable P^{chg} is very weakly correlated with V^{chg} and V^{lvl} . However, when considering disaggregated indicators, it becomes evident that P^{chg}_{inc} is strongly correlated with V^{chg}_{inc} and V^{lvl}_{high} , representing indicators related to expected increases in volatility and high volatility, respectively. The correlation coefficient between V^{chg} and V^{lvl} is relatively high and equal to 0.582. The highest correlation among unidirectional variables is between V^{chg}_{inc} and V^{lvl}_{high} with an estimate of 0.734. These results support the use of methods insensitive to strong correlation among explanatory variables, such as LASSO.

Applied models

In this paper, we apply selected statistical and machine learning methods to model the mean and variance of daily logarithmic returns.

Autoregressive model – AR

To describe the mean, we apply the AR model. A model of order p is denoted as $AR(p)$ and defined as:

$$x_t = \alpha_0 + \sum_{i=1}^p \alpha_i x_{t-i} + \varepsilon_t. \quad (4)$$

We consider the AR model estimated using ordinary least squares (OLS) as a benchmark for the mean of returns.

Heterogeneous autoregressive model – HAR

The model that combines the three variance measures, namely daily, weekly and monthly average realized variances given by Eqs. 1–3 is the HAR model introduced by Corsi [11]. We apply a logarithmic transformation to ensure positive values for the variance estimates and to mitigate the impact of outliers:

$$\ln RV_{d,t} = \gamma_0 + \gamma_1 \ln RV_{d,t-1} + \gamma_2 \ln RV_{w,t-1} + \gamma_3 \ln RV_{m,t-1} + \varepsilon_t. \quad (5)$$

We regard the HAR model estimated by OLS as a benchmark for modelling the variance of returns.

Bayesian model averaging – BMA

Standard linear regression models typically involve selecting explanatory variables while excluding those with minimal influence on the dependent variable. However, this process introduces uncertainty regarding the validity of the chosen variables and model structure. The BMA approach, which Leamer [40] initially introduced and was further developed by Mitchell and Beauchamp [44], Raftery et al. [52], and Hoeting et al. [28] resolves this uncertainty. BMA evaluates all possible combinations of explanatory variables, assigns each a probability of inclusion, and generates a final model as a weighted average of these possible models. The weights reflect the individual models' posterior probabilities, producing more comprehensive and reliable predictions.

By exploring a wide range of model specifications and assigning probabilities based on their posterior distributions, BMA mitigates the risks of variable selection, such as excluding significant factors or overemphasizing irrelevant ones. This approach proves particularly valuable in fields where identifying the most relevant variables poses challenges or involves subjective decisions.

Consider a linear regression model of the form:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \dots + \beta_k x_{k,t} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2 \mathbf{I}), \quad (6)$$

where $\beta_0, \beta_1, \dots, \beta_k$ – model coefficients, k – total number of candidate explanatory variables $x_{1,t}, x_{2,t}, \dots, x_{k,t}$, ε_t – normal IID error term with variance σ^2 , $\mathbf{I}$ – unit matrix.

Given k explanatory variables, there are 2^k possible subsets of variables, denoted as M_j , for $j = 1, 2, \dots, 2^k$.

The vast number of potential models highlights the inherent uncertainty in choosing a single "best" model. Traditional regression methods focus on one model and often ignore other viable alternatives, whereas BMA integrates information from the entire model space.

The BMA framework assigns weights to each model M_j using its posterior probability $P(M_j | D)$, where D represents the observed dataset. The posterior model probability follows the definition:

$$P(M_j | D) = \frac{P(D | M_j) P(M_j)}{P(D)} = \frac{P(D | M_j) P(M_j)}{\sum_{i=1}^{2^k} P(D | M_i) P(M_i)}, \quad (7)$$

where $P(D | M_j)$ is the marginal likelihood of a model M_j , and $P(M_j)$ represents its prior probability. The denominator sums the marginal likelihoods of all possible models weighted by their prior probabilities. The marginal likelihood $P(D | M_j)$ can be expressed as:

$$P(D | M_j) = \int P(D | \theta_j, M_j) P(\theta_j | M_j) d\theta_j, \quad (8)$$

where θ_j denotes the vector of parameters for a model M_j , $P(\theta_j | M_j)$ represents the prior distribution of the parameters, and $P(D | \theta_j, M_j)$ is the probability.

To implement BMA, the model requires a specified prior structure. The g-prior proposed by Zellner [58] is a common choice. Under the g-prior, the vector β of the model's coefficients follows a normal distribution with mean 0 and covariance proportional to $\sigma^2 (g \mathbf{X}_j' \mathbf{X}_j)^{-1}$, where g is a proportionality factor. This relationship can be expressed as:

$$P(\beta | \sigma^2, M_j) \sim N\left(0, \sigma^2 (g \mathbf{X}_j' \mathbf{X}_j)^{-1}\right). \quad (9)$$

BMA generates posterior inclusion probabilities (PIP) for each variable, quantifying its importance in the model. PIP equals the sum of the posterior probabilities for all models that include the variable. These probabilities effectively rank the variables and reflect their contributions to the predictive power of the final model. The total sum of PIPs also provides insight into the average model size, revealing the overall sparsity or complexity of the final model space.

BMA offers a rigorous framework for addressing model uncertainty. By combining evidence across multiple models and weighting them by their posterior probabilities, BMA delivers more balanced and accurate predictions while objectively assessing the importance of explanatory variables. By construction, BMA accounts for model uncertainty and guards against overfitting by assigning lower posterior probabilities to over-parameterized models.

Least absolute shrinkage and selection operator – LASSO

LASSO (Tibshirani, [54]) is a popular regression method that combines variable selection with regularization to improve the interpretability and predictive power of statistical models. The technique is mainly useful for multivariate data and also for collinearity among explanatory variables, which often occurs in financial time series forecasting. Like the least squares method (OLS), LASSO minimizes the sum of squares of the residuals but with an additional penalty that causes some coefficients to take a value of zero, which excludes these variables from the model. The objective function in LASSO has the form:

$$\min_{\beta} \left(\sum_{t=1}^N (y_t - \hat{y}_t)^2 + \lambda \sum_{j=1}^k |\beta_j| \right), \quad (10)$$

where N is the number of observations, k is the number of the explanatory variables, y_t and $\hat{y}_t$ represent, respectively, actual and predicted observations of the response variable, and $\lambda \geq 0$ is the tuning parameter which controls the amount of shrinkage. The parameter λ is crucial for LASSO's properties. When λ is large, the penalty term $\sum_{j=1}^k |\beta_j|$ increases, leading more coefficients to be estimated as zero, therefore, LASSO performs variable selection (e.g., Kuhn and Johnson [36]; James et al. [29]). As λ approaches zero, LASSO converges to OLS, incorporating all variables in the model.

An important advantage of LASSO is its ability to effectively mitigate overfitting through ℓ_1 -regularization. By shrinking coefficient estimates toward zero and setting a subset of them exactly to zero, LASSO reduces model complexity and variance, which improves out-of-sample performance. The tuning parameter λ governs the strength of regularization and thus controls the bias-variance trade-off: larger values of λ yield sparser models with lower variance, while smaller values allow for more flexible

specifications. This property is particularly valuable in high-dimensional settings and in applications involving strongly correlated predictors, such as financial time series forecasting.

Support vector regression – SVR

SVR (Vapnik et al. [56]) is a supervised learning algorithm derived from the support vector machine (SVM, Vapnik [55]) methodology, designed specifically for regression tasks. Advantages of support vector regression include: efficient handling of nonlinear relationships, robustness to outliers, control of over-fitting, performance in multidimensional spaces (i.e., when the number of variables is large compared to the number of data points). In SVR, the input data $\mathbf{x}$ is first mapped to a high-dimensional feature space using some fixed (non-linear) mapping φ_j . The goal of SVR is to find the function:

$$f(\mathbf{x}) = \sum_{j=1}^d \omega_j \varphi_j(\mathbf{x}) + b \quad (11)$$

which approximates the relationship between the input vector $\mathbf{x}$ and the response variable while balancing two main goals: minimizing prediction error and avoiding over-fitting.

The optimization problem in SVR can be formulated as follows:

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{t=1}^n (\xi_t + \xi_t^*), \quad (12)$$

subject to

$$y_t - f(\mathbf{x}_t) \leq \varepsilon + \xi_t^*,$$

$$f(\mathbf{x}_t) - y_t \leq \varepsilon + \xi_t,$$

$$\xi_t, \xi_t^* \geq 0,$$

where ω represents the weight vector, b is the bias term, ξ_t and ξ_t^* are slack variables, ε defines the margin of tolerance within which errors are ignored. An important feature of SVR is its inherent ability to mitigate overfitting. This is achieved through the principle of structural risk minimization, which balances model complexity and empirical risk. In particular, the regularization term $\frac{1}{2} \|\omega\|^2$ penalizes overly complex models, while the regularization parameter C explicitly controls the trade-off between the flatness of the regression function and the tolerance for deviations exceeding the ε -insensitive loss. As a result, SVR tends to exhibit strong generalization performance, especially in high-dimensional settings and in cases with limited sample sizes.

The optimization problem described above can be transformed into a dual problem (e.g. Fiszeder and Orzeszko [21]), for which the solution is given by:

$$f(\mathbf{x}) = \sum_{t=1}^{T_{SV}} (\alpha_t - \alpha_t^*) K(\mathbf{x}_t, \mathbf{x}), \text{ s.t. } 0 \leq \alpha_t \leq C, 0 \leq \alpha_t^* \leq C, \quad (13)$$

where α_t and α_t^* are the Lagrange multipliers, T_{SV} is the number of support vectors and K is the kernel function of the form:

$$K(\mathbf{x}_t, \mathbf{x}) = \sum_{i=1}^d \varphi_i(\mathbf{x}) \varphi_i(\mathbf{x}_t). \quad (14)$$

Any function satisfying the Mercer's condition can be used as the kernel (Vapnik [55]). In our study we applied two most popular kernel functions:

- linear kernel (dot product)

$$K(\mathbf{x}_t, \mathbf{x}) = \mathbf{x}_t^T \mathbf{x}, \quad (15)$$

- radial basis function (RBF, Gaussian)

$$K(\mathbf{x}_t, \mathbf{x}) = \exp\left(-\gamma \|\mathbf{x}_t - \mathbf{x}\|^2\right). \quad (16)$$

Research procedure

In our research, we assess the usefulness of ChatGPT-derived sentiment indicators in modelling and forecasting BTC returns and volatility. In the case of return analysis, the subject of the study is the daily logarithmic return, and for volatility, the log-realized variance $\ln RV_{d,t}$ (see Eq. 1).

The study considers nine sentiment indicators: P^{chg} , V^{chg} , V^{lvl} , P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} and V_{low}^{lvl} , presented in Sect. "Construction of the ChatGPT-derived sentiment indicators". The analysis examines the impact of both contemporaneous indicators and those with a lag of 1. A single-period lag is adopted to capture short-term dependence typical of financial time series while maintaining model parsimony and limiting the risk of overfitting.

The research is divided into two parts: in-sample analysis and out-of-sample analysis.

In-sample analysis

For the in-sample analysis, data from the entire study period, i.e., from January 1, 2021, to March 31, 2024, is analyzed. Three research procedures are performed:

1) Correlation Analysis: The correlation between the sentiment indicators and both the logarithmic return $r_{d,t}$ and the realized variance $\ln RV_{d,t}$ is examined. Pearson's linear correlation coefficient and its significance test are applied to pairs of variables, including both contemporaneous and lagged indicators.

2) Granger Causality Testing: The presence of Granger causality (Granger, [24]) is verified for pairs of variables using the Wald test. First, the relationship between $r_{d,t}$ and each sentiment indicator is examined. Then, the relationship between $\ln RV_{d,t}$ and each of these indicators is analyzed.

3) Analysis of Regression Models: The in-sample performance of models with the sentiment indicators and models without these variables is compared. The following models for daily logarithmic returns and for volatility are constructed:

a) For daily logarithmic returns:

a1) Autoregressive models of order 1 without the sentiment indicators (the order of an autoregression is determined based on the partial autocorrelation function and the Ljung–Box test), i.e. the models of the form:

$$r_{d,t} = f(r_{d,t-1}). \quad (17)$$

Two linear models of this form are considered, estimated using OLS (denoted as AR hereafter) and LASSO, as well as two nonlinear models, specifically SVR with linear and Gaussian kernels (denoted as SVR-L and SVR-G, respectively).

a2) Autoregressive models extended with the sentiment indicators. To this end, the set of all nine indicators is divided into two subsets:

- $X1 = \{P^{chg}, V^{chg}, V^{lvl}\},$
- $X2 = \{P_{inc}^{chg}, P_{dec}^{chg}, V_{inc}^{chg}, V_{dec}^{chg}, V_{high}^{lvl}, V_{low}^{lvl}\}.$

For both subsets, we consider models with both contemporaneous and lagged indicators (lag=1). This results in four types of models:

- For the X1 subset:

$$r_{d,t} = f(r_{d,t-1}, P_t^{chg}, V_t^{chg}, V_t^{lvl}), \quad (18)$$

$$r_{d,t} = f(r_{d,t-1}, P_{t-1}^{chg}, V_{t-1}^{chg}, V_{t-1}^{lvl}), \quad (19)$$

- For the X2 subset:

$$r_{d,t} = f(r_{d,t-1}, P_{inc,t}^{chg}, P_{dec,t}^{chg}, V_{inc,t}^{chg}, V_{dec,t}^{chg}, V_{high,t}^{lvl}, V_{low,t}^{lvl}), \quad (20)$$

$$r_{d,t} = f(r_{d,t-1}, P_{inc,t-1}^{chg}, P_{dec,t-1}^{chg}, V_{inc,t-1}^{chg}, V_{dec,t-1}^{chg}, V_{high,t-1}^{lvl}, V_{low,t-1}^{lvl}). \quad (21)$$

For all above general types of models, we consider linear models estimated using three methods: OLS, BMA, and LASSO (denoted as AR-X, BMA-X, and LASSO-X, respectively), as well as nonlinear SVR models with linear and Gaussian kernels (denoted as SVR-L-X and SVR-G-X). The hyperparameters of the LASSO model (λ) and SVR models (ε , C and γ in case of the SVR-G model) are optimized using a grid-search procedure with tenfold cross-validation.

b) For volatility:

b1) Models without the sentiment indicators, of the form:

$$\ln RV_{d,t} = f(\ln RV_{d,t-1}, \ln RV_{w,t-1}, \ln RV_{m,t-1}). \quad (22)$$

We consider five models of this form. Three of them are linear, following the HAR specification (see Eq. 5), with parameters estimated using OLS, BMA, and LASSO. The remaining two are nonlinear models, namely SVR with linear and Gaussian kernels.

b2) Models extended with the sentiment indicators. Similar to the daily logarithmic returns, four types of models are considered:

- For the X1 subset:

$$\ln RV_{d,t} = f(\ln RV_{d,t-1}, \ln RV_{w,t-1}, \ln RV_{m,t-1}, P_t^{chg}, V_t^{chg}, V_t^{lvl}), \quad (23)$$

$$\ln RV_{d,t} = f(\ln RV_{d,t-1}, \ln RV_{w,t-1}, \ln RV_{m,t-1}, P_{t-1}^{chg}, V_{t-1}^{chg}, V_{t-1}^{lvl}), \quad (24)$$

- For the X2 subset:

$$\ln RV_{d,t} = f\left(\ln RV_{d,t-1}, \ln RV_{w,t-1}, \ln RV_{m,t-1}, P_{inc,t}^{chg}, P_{dec,t}^{chg}, V_{inc,t}^{chg}, V_{dec,t}^{chg}, V_{high,t}^{lvol}, V_{low,t}^{lvol}\right), \quad (25)$$

$$\ln RV_{d,t} = f\left(\ln RV_{d,t-1}, \ln RV_{w,t-1}, \ln RV_{m,t-1}, P_{inc,t-1}^{chg}, P_{dec,t-1}^{chg}, V_{inc,t-1}^{chg}, V_{dec,t-1}^{chg}, V_{high,t-1}^{lvol}, V_{low,t-1}^{lvol}\right). \quad (26)$$

As in the case of daily logarithmic returns, linear models are estimated using OLS, BMA, and LASSO (denoted as HAR-X, BMA-X and LASSO-X, respectively), and non-linear SVR models with linear and Gaussian kernels (denoted as SVR-L-X and SVR-G-X) are constructed.

We evaluate and compare the constructed models in terms of in-sample accuracy using selected metrics: the mean squared error (MSE), the mean absolute error (MAE), the coefficient of determination (R^2), and the model confidence set (MCS, Hansen et al. [27]) procedure, as well as the Superior Predictive Ability (SPA, Hansen [26]) test.

The research procedures for the in-sample analysis of returns and volatility are presented in Fig. 5 and Fig. 6, respectively.

Out-of-sample analysis

In the out-of-sample analysis, the predictive power of the models described above is compared. Similar to the in-sample analysis, separate models are constructed for returns and volatility. The accuracy of forecasts obtained from models without sentiment indicators is compared with those from models incorporating these variables, which are divided into two alternative sets: X1 and X2. However, unlike the in-sample analysis, we only consider models with lagged indicators (with a lag of 1) since models with contemporaneous indicators are not suitable for forecasting purposes due to the lack of knowledge about predictor values during the forecasted period.

The constructed models are employed to generate one-step-ahead forecasts for observations spanning the period from August 1, 2022, to March 31, 2024. Initially, to produce a forecast for the first day (i.e., August 1, 2022), the models are trained using data from the preceding period (January 1, 2021, to July 31, 2022). Subsequently, to forecast August 2, 2022, the training sample is updated following a rolling window approach. Specifically, the oldest observation is removed and replaced with the observed value from August 1, 2022. With each update of the training sample, the models are re-estimated.

The hyperparameters of the LASSO model (λ) and SVR models (ε , C and γ in case of the SVR-G model) are optimized every 28 observations (on Sundays) using a grid-search procedure with tenfold cross-validation. The 28-observation updating frequency is chosen as a compromise between computational efficiency and the need to adapt model complexity to evolving market conditions. Re-estimating hyperparameters on a rolling basis helps mitigate overfitting and improves out-of-sample stability, particularly in financial time series characterized by time-varying dynamics. In the SVR-L and SVR-G models, the data are standardized.

As a result, 609 forecasts are generated for each type of model. In order to compare their accuracy, we apply the same metrics and tests as in the in-sample analysis.

The research procedures for generating forecasts of daily logarithmic returns and volatility are detailed in Procedure 1 and Procedure 2, respectively.

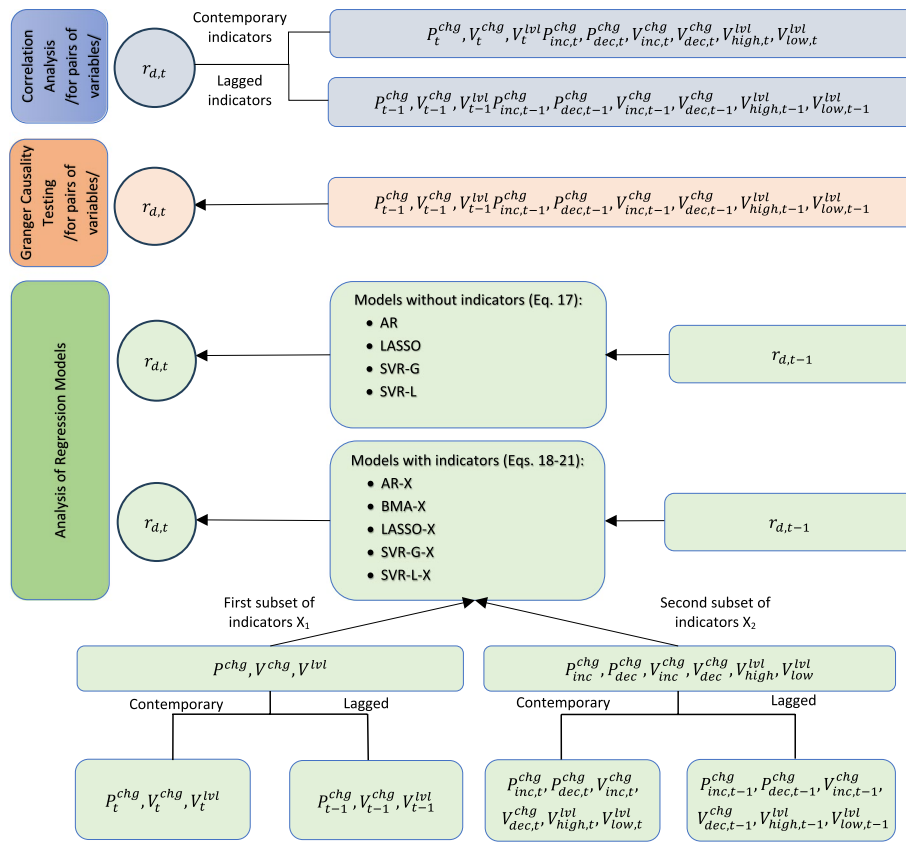


Fig. 5 In-sample analysis for returns

Procedure 1 Procedure for generating forecasts of daily logarithmic returns

- Input:** daily logarithmic returns $r_{d,t}$ and nine sentiment indicators grouped in two subsets $X_1 = \{P_t^{chg}, V_t^{chg}, V_t^{lvl}\}$ and $X_2 = \{P_{inc,t}^{chg}, P_{dec,t}^{chg}, V_{inc,t}^{chg}, V_{dec,t}^{chg}, V_{high,t}^{lvl}, V_{low,t}^{lvl}\}$, from January 1, 2021, to March 31, 2024.
- Setting training samples:
 - $\{[r_{d,t-1}], r_{d,t}; t = 2, 3, \dots, T\}$ for model (17)
 - $\{[r_{d,t-1}, P_{t-1}^{chg}, V_{t-1}^{chg}, V_{t-1}^{lvl}], r_{d,t}; t = 2, 3, \dots, T\}$ for model (19)
 - $\{[r_{d,t-1}, P_{inc,t-1}^{chg}, P_{dec,t-1}^{chg}, V_{inc,t-1}^{chg}, V_{dec,t-1}^{chg}, V_{high,t-1}^{lvl}, V_{low,t-1}^{lvl}], r_{d,t}; t = 2, 3, \dots, T\}$ for model (21)

In the first iteration the training sample spans January 1, 2021 to July 31, 2022.
- Training predictive models (17), (19) and (21).
The hyperparameters of LASSO and SVR models are optimized every 28 observations (on Sundays).
- Generation of the forecast $\hat{r}_{d,T+1}$.
- Updating a training sample according to a rolling window procedure, i.e. adding one new observation while removing the oldest one:
 - $\{[r_{d,T}], r_{d,T+1}\}$ for model (17)
 - $\{[r_{d,T}, P_T^{chg}, V_T^{chg}, V_T^{lvl}], r_{d,T+1}\}$ for model (19)
 - $\{[r_{d,T}, P_{inc,T}^{chg}, P_{dec,T}^{chg}, V_{inc,T}^{chg}, V_{dec,T}^{chg}, V_{high,T}^{lvl}, V_{low,T}^{lvl}], r_{d,T+1}\}$ for model (21)
- Setting $T = T + 1$.
- Repeating steps 2-6 until forecasts for the period from August 1, 2022, to March 31, 2024 are generated.
- Output:** Forecasts of daily logarithmic returns $\hat{r}_{d,t}$ for the period from August 1, 2022, to March 31, 2024.

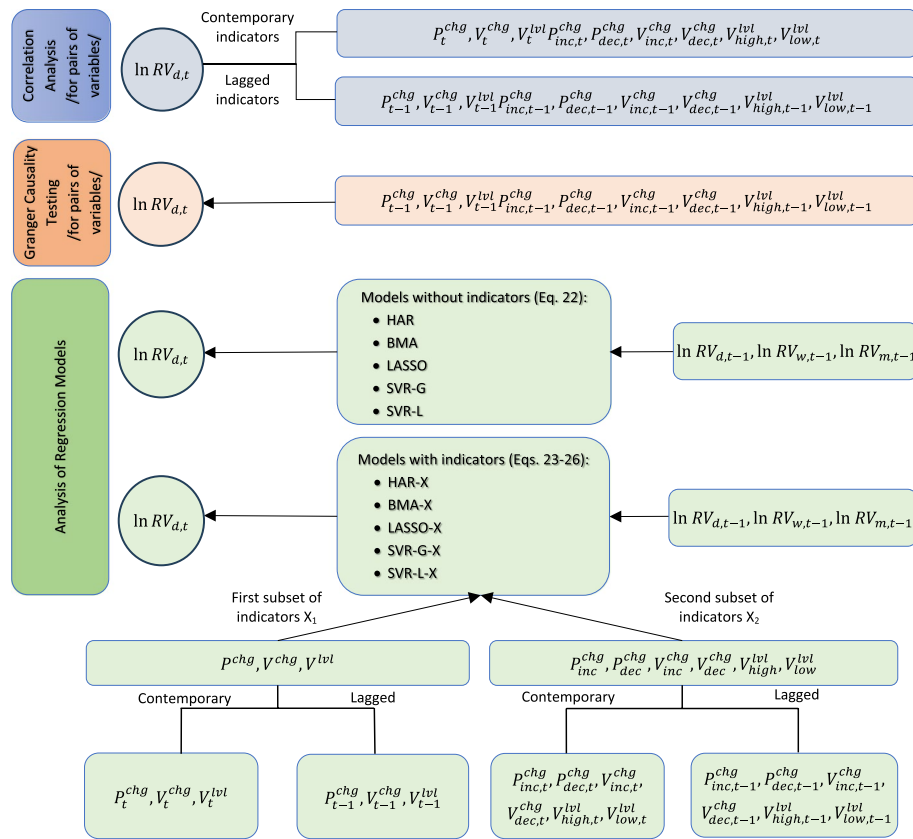


Fig. 6 In-sample analysis for volatility

Procedure 2 Procedure for generating forecasts of volatility

- Input:** daily realized variances $RV_{d,t}$ and nine sentiment indicators grouped in two subsets $X_1 = \{p_t^{chg}, V_t^{chg}, V_t^{lvl}\}$ and $X_2 = \{p_{inc,t}^{chg}, p_{dec,t}^{chg}, V_{inc,t}^{chg}, V_{dec,t}^{chg}, V_{high,t}^{lvl}, V_{low,t}^{lvl}\}$, from January 1, 2021, to March 31, 2024.
- Calculation of weekly and monthly realized variances $RV_{w,t}$ and $RV_{m,t}$ (see (3) and (4)).
- Logarithmic transformation: $\ln RV_{d,t}, \ln RV_{w,t}, \ln RV_{m,t}$.
- Setting training samples:
 - $\{\ln RV_{d,t-1}, \ln RV_{w,t-1}, \ln RV_{m,t-1}, \ln RV_{d,t}; t = 2, 3, \dots, T\}$ for model (22)
 - $\{\ln RV_{d,t-1}, \ln RV_{w,t-1}, \ln RV_{m,t-1}, p_{t-1}^{chg}, V_{t-1}^{chg}, V_{t-1}^{lvl}, \ln RV_{d,t}; t = 2, 3, \dots, T\}$ for model (24)
 - $\{\ln RV_{d,t-1}, \ln RV_{w,t-1}, \ln RV_{m,t-1}, p_{inc,t-1}^{chg}, p_{dec,t-1}^{chg}, V_{inc,t-1}^{chg}, V_{dec,t-1}^{chg}, V_{high,t-1}^{lvl}, V_{low,t-1}^{lvl}, \ln RV_{d,t}; t = 2, 3, \dots, T\}$ for model (26)
- In the first iteration the training sample spans January 1, 2021 to July 31, 2022.
- Training predictive models (22), (24) and (26). The hyperparameters of LASSO and SVR models are optimized every 28 observations (on Sundays).
- Generation of the forecast $\ln \widehat{RV}_{d,T+1}$.
- Calculation of the forecast $\widehat{RV}_{d,T+1} = \exp(\ln \widehat{RV}_{d,T+1})$.
- Updating a training sample according to a rolling window procedure, i.e. adding one new observation while removing the oldest one:
 - $\{\ln RV_{d,T}, \ln RV_{w,T}, \ln RV_{m,T}, \ln RV_{d,T+1}\}$ for model (22)
 - $\{\ln RV_{d,T}, \ln RV_{w,T}, \ln RV_{m,T}, p_T^{chg}, V_T^{chg}, V_T^{lvl}, \ln RV_{d,T+1}\}$ for model (24)
 - $\{\ln RV_{d,T}, \ln RV_{w,T}, \ln RV_{m,T}, p_{inc,T}^{chg}, p_{dec,T}^{chg}, V_{inc,T}^{chg}, V_{dec,T}^{chg}, V_{high,T}^{lvl}, V_{low,T}^{lvl}, \ln RV_{d,T+1}\}$ for model (26)
- Setting $T = T + 1$.
- Repeating steps 4-9 until forecasts for the period from August 1, 2022, to March 31, 2024 are generated.
- Output:** Forecasts of daily realized variances $\widehat{RV}_{d,t}$ for the period from August 1, 2022, to March 31, 2024.

Empirical results

In-sample results

First, we examine whether the news sentiment indicators derived from ChatGPT significantly impact the returns (Sect. "Analysis of returns") and the volatility (Sect. "Analysis of volatility") of BTC in an in-sample analysis.

Analysis of returns

In this section, we examine the influence of the sentiment indicators on BTC returns. As a preliminary stage of the research, we conduct a correlation analysis between the logarithmic return $r_{d,t}$ and the indicators. We consider both contemporaneous and lagged relationships. The results are presented in Table 6.

For contemporaneous relationships, statistically significant (at the 0.1 significance level) positive correlations are observed between BTC returns and the variables P_t^{chg} , $P_{inc,t}^{chg}$ and $P_{dec,t}^{chg}$, while a negative correlation is found with $V_{high,t}^{lvl}$. On the other hand, significant positive lagged relationships are observed between $r_{d,t}$ and the indicators V_{t-1}^{chg} , $P_{inc,t-1}^{chg}$ and $V_{inc,t-1}^{chg}$ along with a negative correlation for $V_{low,t-1}^{lvl}$. The highest values of the correlation coefficients are observed between BTC returns and the contemporaneous variables P_t^{chg} , $P_{inc,t}^{chg}$ and $P_{dec,t}^{chg}$. These findings indicate a strong, statistically significant relationship between the sentiment indicators and BTC returns.

In the second step, we test for Granger causality. We examine whether each indicator Granger-causes BTC returns. The results of this investigation are displayed in Table 7.

Four indicators: P^{chg} , V^{chg} , P_{inc}^{chg} and V_{inc}^{chg} Granger-cause BTC returns, implying that past values of these variables contain information that helps predict BTC returns. The results of the Granger causality tests are reinforced by the impulse response analysis presented in Fig. 7 in Appendix B, indicating that the dynamic effects of sentiment shocks on BTC returns are consistent with the results implied by the causality findings.

In the next step, we build models in which the dependent variable is BTC returns $r_{d,t}$ and the explanatory variables are the indicators derived from ChatGPT. We group the explanatory variables into two sets: the first set X1 includes three indicators: P^{chg} , V^{chg} and V^{lvl} , the second set X2 includes six disaggregated indicators: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} and V_{low}^{lvl} . Due to the autocorrelation of BTC returns, the lagged return $r_{d,t-1}$ is also included as an explanatory variable. We present the estimation results for the AR-X, BMA-X and LASSO-X models in Table 22 in Appendix A.

For contemporaneous relationships, all models identify P_t^{chg} , V_t^{chg} , $P_{inc,t}^{chg}$ and $P_{dec,t}^{chg}$ as significant regressors of BTC returns. In contrast, the influence of lagged indicators is much weaker. According to the AR-X model, only P_{t-1}^{chg} and V_{t-1}^{chg} exhibit a statistically significant impact, while for the BMA-X and LASSO-X models contributions of additional indicators are marginal.

We use the constructed models to estimate BTC returns for the whole in-sample period January 1, 2021—March 31, 2024. Specifically, we apply the AR-X, BMA-X, LASSO-X, SVR-G-X, and SVR-L-X models and compare the results with models that do not include sentiment indicators, namely AR, LASSO, SVR-G, and SVR-L. Two sets of the indicators, referred to as X1 and X2, are incorporated into the analysis. We assess whether the inclusion of information derived from ChatGPT improves the accuracy of BTC return estimates in-sample. To evaluate the performance of our models, we

implement the MCS procedure and the SPA test. The outcomes of the MCS test are displayed in Table 8 and Table 9 for models with contemporaneous and lagged relationships, respectively. The SPA test results are summarized in Table 10.

First, we test all models jointly using the MCS procedure. Among models with contemporaneous relationships, SVR-G-X2 best estimates BTC returns and is also the only one included in the set of best models. This result remains regardless of whether the MSE or MAE measure is used for evaluation. Additionally, according to the R^2 coefficient, the best model is also SVR-G-X2. Across all considered evaluation criteria, the second-best model is SVR-G-X1. In general, models based on the variables from the X2 set rank higher than those based on the X1 set.

All models including the sentiment indicators exhibit lower MSE and MAE errors compared to models that do not incorporate these variables. Substantial differences between these models are observed in the R^2 measure. The values of the coefficient of determination for models that do not contain the indicators are 0.003 or less, while for those that do, they are 0.24 or higher.

The results of the MCS test for models that incorporate the lagged sentiment indicators are notably different. For the MSE measure, all models belong to the set of best models, while for MAE, all models except AR-X2 are included in this set. This indicates that it is impossible to identify models that are significantly better than the others. The R^2 values for models using the indicators are slightly higher than those for models without these variables; however, these values remain relatively low.

To assess the influence of the sentiment indicators on individual models' performance, we compare pairs of models using the SPA test. In each case, we compare a model without the indicators with a model that includes these variables. The results for contemporaneous relationships indicate that, for all models and both the MSE and MAE measures, incorporating the indicators improves the accuracy of BTC returns estimation. This conclusion holds true for both sets of variables, X1 and X2. For lagged relationships, the influence of the sentiment indicators is significant (at the 0.1 significance level) for the MSE metric in the following models: AR-X1, AR-X2, LASSO-X1, LASSO-X2, SVR-G-X1, and SVR-L-X1, as well as for the MAE criterion in the SVR-L-X1 model.

Table 6 Pearson's correlation coefficients between the BTC logarithmic return $r_{d,t}$ and the sentiment indicators

Contemporaneous indicators								
P_t^{chg}	V_t^{chg}	V_t^{lvl}	$P_{inc,t}^{chg}$	$P_{dec,t}^{chg}$	$V_{inc,t}^{chg}$	$V_{dec,t}^{chg}$	$V_{high,t}^{lvl}$	$V_{low,t}^{lvl}$
0.429	-0.024	-0.021	0.299	0.412	-0.029	-0.006	-0.051	0.034
(0.000)	(0.404)	(0.471)	(0.000)	(0.000)	(0.315)	(0.828)	(0.079)	(0.236)
Lagged indicators								
P_{t-1}^{chg}	V_{t-1}^{chg}	V_{t-1}^{lvl}	$P_{inc,t-1}^{chg}$	$P_{dec,t-1}^{chg}$	$V_{inc,t-1}^{chg}$	$V_{dec,t-1}^{chg}$	$V_{high,t-1}^{lvl}$	$V_{low,t-1}^{lvl}$
0.029	0.063	0.033	0.057	-0.005	0.062	-0.002	0.030	-0.051
(0.313)	(0.030)	(0.262)	(0.051)	(0.861)	(0.033)	(0.945)	(0.305)	(0.078)

Notes: The table contains correlation coefficient values with the corresponding p-values in parentheses. Correlation coefficients that are statistically significant at the 10% significance level are highlighted in bold. The analyzed period is January 1, 2021—March 31, 2024

Table 7 Results of the Granger causality test for the sentiment indicators and the BTC returns $r_{d,t}$

Variable	P-value	Lags in VAR
P^{chg}	0.050	1
V^{chg}	0.033	1
V^{lvl}	0.279	1
P_{dec}^{chg}	0.544	1
P_{inc}^{chg}	0.009	1
V_{dec}^{chg}	0.934	1
V_{inc}^{chg}	0.037	1
V_{low}^{lvl}	0.475	1
V_{high}^{lvl}	0.350	1

Notes: The table reports the results of the Granger causality test between Bitcoin daily returns $r_{d,t}$ and the sentiment indicators derived from ChatGPT. The p-value corresponds to the null hypothesis that the sentiment indicator does not Granger-cause $r_{d,t}$. P-values smaller than 0.1, indicating rejection of the null hypothesis, are highlighted in bold. The number of lags in the VAR models is determined by the Schwarz information criterion (AIC and HQC gave the same results; see Polyzos and Siriopoulos [50] for more advanced lag selection methods). The analyzed period is January 1, 2021—March 31, 2024

Table 8 Evaluation of in-sample BTC returns estimation using the MCS test for models without and with contemporaneous sentiment indicators

Method	MSE	Rank	P-value	MAE	Rank	P-value	R^2	Rank
AR	1.149	12	0.000	2.332	12	0.000	0.003	12
LASSO	1.152	13	0.000	2.335	13	0.000	0.000	14
SVR-G	1.149	11	0.000	2.332	11	0.000	0.003	11
SVR-L	1.152	14	0.000	2.339	14	0.000	0.003	13
AR-X1	0.874	7	0.026	2.196	10	0.000	0.242	7
BMA-X1	0.874	9	0.024	2.193	8	0.000	0.241	9
LASSO-X1	0.874	8	0.030	2.193	9	0.000	0.242	8
SVR-G-X1	0.857	2	0.039	2.129	2	0.084	0.261	2
SVR-L-X1	0.887	10	0.018	2.163	3	0.001	0.240	10
AR-X2	0.857	3	0.071	2.173	6	0.000	0.256	3
BMA-X2	0.860	5	0.039	2.176	7	0.000	0.254	5
LASSO-X2	0.857	4	0.039	2.170	5	0.000	0.256	4
SVR-G-X2	0.830	1	1.000*	2.107	1	1.000*	0.287	1
SVR-L-X2	0.870	6	0.034	2.164	4	0.001	0.250	6

Notes: The table reports the results of the Model Confidence Set (MCS) test for comparing the in-sample estimation accuracy of Bitcoin daily returns. The analysis includes the benchmark models – which do not incorporate sentiment indicators and the models with contemporaneous indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} , V_{low}^{lvl} , the values of MSE are multiplied by 10^3 , the values of MAE are multiplied by 10^2 , the lowest values of MSE and MAE and the highest for R^2 are in bold, p-value is for the MCS test, * indicates that models belong to MCS with a confidence level of 0.90. The analyzed period is January 1, 2021—March 31, 2024

Table 9 Evaluation of in-sample BTC returns estimation using the MCS test for models without and with lagged sentiment indicators

Method	MSE	Rank	P-value	MAE	Rank	P-value	R^2	Rank
AR	1.149	8	0.455*	2.332	5	0.450*	0.003	12
LASSO	1.152	12	0.194*	2.335	10	0.245*	0.000	14
SVR-G	1.149	7	0.400*	2.332	4	0.450*	0.003	11
SVR-L	1.153	13	0.284*	2.339	13	0.175*	0.003	13
AR-X1	1.142	2	0.761*	2.336	12	0.245*	0.009	3
BMA-X1	1.151	9	0.560*	2.334	7	0.342*	0.007	7
LASSO-X1	1.142	3	0.761*	2.333	6	0.342*	0.009	4
SVR-G-X1	1.144	4	0.761*	2.326	1	1.000*	0.009	2
SVR-L-X1	1.146	6	0.761*	2.328	2	0.456*	0.006	8
AR-X2	1.140	1	1.000*	2.341	14	0.095	0.011	1
BMA-X2	1.152	10	0.362*	2.334	9	0.342*	0.007	6
LASSO-X2	1.145	5	0.761*	2.331	3	0.450*	0.009	5
SVR-G-X2	1.153	14	0.105*	2.335	11	0.245*	0.004	10
SVR-L-X2	1.152	11	0.374*	2.334	8	0.342*	0.005	9

Notes: The table reports the results of the Model Confidence Set (MCS) test for comparing the in-sample estimation accuracy of Bitcoin daily returns. The analysis includes the benchmark models – which do not incorporate sentiment indicators and the models with lagged indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} , V_{low}^{lvl} , the values of MSE are multiplied by 10^3 , the values of MAE are multiplied by 10^2 , the lowest values of MSE and MAE, and the highest for R^2 are in bold, p-value is for the MCS test, * indicates that models belong to MCS with a confidence level of 0.90. The analyzed period is January 1, 2021—March 31, 2024

Table 10 Evaluation of in-sample BTC returns estimation using the SPA test: models without additional variables vs. models with contemporaneous or lagged sentiment indicators

Compared models	Contemporaneous		Lagged	
	MSE	MAE	MSE	MAE
	P-value	P-value	P-value	P-value
AR vs. AR-X1	0.000	0.001	0.092	0.696
AR vs. AR-X2	0.000	0.000	0.086	0.863
AR vs. BMA-X1	0.000	0.001	0.665	0.669
AR vs. BMA-X2	0.000	0.000	0.676	0.672
LASSO vs. LASSO-X1	0.000	0.001	0.077	0.413
LASSO vs. LASSO-X2	0.000	0.000	0.034	0.194
SVR-G vs. SVR-G-X1	0.000	0.000	0.084	0.155
SVR-G vs. SVR-G-X2	0.000	0.000	0.748	0.743
SVR-L vs. SVR-L-X1	0.000	0.000	0.024	0.017
SVR-L vs. SVR-L-X2	0.000	0.000	0.468	0.277

Notes: The table reports the results of the Superior Predictive Ability (SPA) test for the in-sample estimation of Bitcoin daily returns. The test compares the benchmark models – without sentiment indicators against models incorporating contemporaneous or lagged indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} , V_{low}^{lvl} , the table shows the p-values of the SPA test for pairs of models (listed on the left). A p-value below the significance level, highlighted in bold, indicates that the estimates from the second model are more accurate than those from the first model, which serves as the benchmark. The analyzed period is January 1, 2021—March 31, 2024

Analysis of volatility

In this subsection, we analyze the influence of the sentiment indicators on BTC volatility. In the first stage, we calculate the correlation coefficient between $\ln RV_{d,t}$ and both the contemporaneous and lagged values of these indicators. The results are presented in Table 11.

For contemporaneous relations, a statistically significant negative correlation exists between $\ln RV_{d,t}$ and P_t^{chg} , $P_{inc,t}^{chg}$ and $P_{dec,t}^{chg}$, while positive correlations are observed between $\ln RV_{d,t}$ and V_t^{lvl} , $V_{high,t}^{lvl}$ and $V_{low,t}^{lvl}$. The correlations for the same indicators

Table 11 Pearson's correlation coefficients between the logarithm of realized BTC variance $\ln RV_{d,t}$ and the sentiment indicators

Contemporaneous indicators								
P_t^{chg}	V_t^{chg}	V_t^{lvl}	$P_{inc,t}^{chg}$	$P_{dec,t}^{chg}$	$V_{inc,t}^{chg}$	$V_{dec,t}^{chg}$	$V_{high,t}^{lvl}$	$V_{low,t}^{lvl}$
-0.245	0.039	0.178	-0.140	-0.254	0.016	0.029	0.112	0.167
(0.000)	(0.180)	(0.000)	(0.000)	(0.000)	(0.577)	(0.318)	(0.000)	(0.000)
Lagged indicators								
P_{t-1}^{chg}	V_{t-1}^{chg}	V_{t-1}^{lvl}	$P_{inc,t-1}^{chg}$	$P_{dec,t-1}^{chg}$	$V_{inc,t-1}^{chg}$	$V_{dec,t-1}^{chg}$	$V_{high,t-1}^{lvl}$	$V_{low,t-1}^{lvl}$
-0.195	0.007	0.105	-0.159	-0.155	-0.017	0.013	0.039	0.146
(0.000)	(0.805)	(0.000)	(0.000)	(0.000)	(0.558)	(0.655)	(0.176)	(0.000)

Notes: The table contains correlation coefficient values and the corresponding p-values in parentheses. Correlation coefficients that are statistically significant at the 10% significance level are highlighted in bold. The analyzed period is January 1, 2021—March 31, 2024

are also significant for lagged relations, except for variable $V_{high,t-1}^{lvl}$. The highest absolute values of the correlation coefficients are between BTC volatility and the contemporaneous variables P_t^{chg} and $P_{dec,t}^{chg}$.

Afterwards, we evaluate Granger causality by testing whether each indicator does not Granger-cause BTC volatility. The results of this investigation are depicted in Table 12.

All indicators, except P_{inc}^{chg} and V_{dec}^{chg} , Granger-cause BTC volatility. This suggests that these variables are effective for predicting BTC volatility. The results of the Granger causality tests are reinforced by the impulse response analysis presented in Fig. 8 in Appendix B.

In the next step, we construct models to describe BTC variance, incorporating the sentiment indicators. The dependent variable is the logarithm of realized BTC variance, $\ln RV_{d,t}$, while the explanatory variables are the indicators grouped into two sets, X1 and X2. Additionally, we include weekly and monthly average realized variances as regressors, as described in Eqs. (2) and (3). The estimation results for the HAR, BMA, and LASSO models are displayed in Table 23 in Appendix A.

For contemporaneous relationships, all models identify the following sentiment indicators as significant explanatory factors for BTC variance: V_t^{lvl} , $P_{inc,t}^{chg}$, $V_{dec,t}^{chg}$ and $V_{high,t}^{lvl}$ (positive influence), as well as P_t^{chg} , $P_{dec,t}^{chg}$, and $V_{inc,t}^{chg}$ (negative influence). In contrast, in the case of lagged relationships, none of the indicators show a significant impact on BTC volatility.

We use the estimated models to evaluate the in-sample variance of BTC. Specifically, we apply the following models: HAR-X, BMA-X, LASSO-X, SVR-G-X, and SVR-L-X, incorporating two sets of indicators, X1 and X2. To assess whether the information from ChatGPT improves the accuracy of variance estimation, we compare these results with models that do not include such information, namely HAR, BMA, LASSO, SVR-G, and SVR-L. The statistical significance of the findings is verified using the MCS and SPA tests. The MCS outcomes are summarized in Table 13 and Table 14 for contemporaneous and lagged relationships, respectively, while Table 15 presents the SPA test results.

According to the MCS test for the MSE measure and contemporaneous relationships, all models incorporating the sentiment indicators, except BMA-X1, belong to the set of best models. In contrast, the SVR-G-X1 model, which achieves the lowest MAE value, is the only model included in the set of best models for this criterion. In addition, it is worth mentioning that according to all three criteria, MSE, MAE, and the R^2 coefficient, the two best models are the SVR-X variants.

Table 12 Results of the Granger causality test for the sentiment indicators and the logarithm of realized BTC variance, $\ln RV_{d,t}$

Variable	P-value	Lags in VAR
P^{chg}	0.095	14
V^{chg}	0.000	14
V^{lvl}	0.000	14
P_{dec}^{chg}	0.000	14
P_{inc}^{chg}	0.178	14
V_{dec}^{chg}	0.271	21
V_{inc}^{chg}	0.000	14
V_{low}^{lvl}	0.029	14
V_{high}^{lvl}	0.000	14

Notes: The table reports the results of the Granger causality test between the logarithm of realized BTC variance, $\ln RV_{d,t}$ and the sentiment indicators derived from ChatGPT. The p-values correspond to the null hypothesis that the sentiment indicators do not Granger-cause $\ln RV_{d,t}$. P-values smaller than 0.1, indicating rejection of the null hypothesis, are highlighted in bold. The number of lags in the VAR models is determined using the Schwarz information criterion (alternative lag selection criteria, including AIC and HQC, suggest different lag lengths; however, the main Granger causality results remain qualitatively unchanged). The analyzed period is January 1, 2021—March 31, 2024

Across all evaluation measures, models incorporating information from ChatGPT demonstrate greater accuracy than those without additional variables. Specifically, the MSE and MAE values are significantly lower, while the R^2 values are notably higher for models with this information.

The results for lagged relationships are significantly different from those for contemporaneous relationships. For the MSE measure, all models, except SVR-L, SVR-G-X1, SVR-G-X2, and SVR-L-X2, are included in the set of best models. In the case of the MAE criterion, all models constitute the set of best models. This indicates that it is not possible to differentiate among most models based on these evaluation measures. Similarly, the R^2 coefficient values for all models do not differ significantly from each other.

Additionally, we test whether the sentiment indicators improve the estimation of BTC variance using the SPA test for pairs of models. For contemporaneous relationships evaluated using both MSE and MAE criteria and both variable sets, X1 and X2, all models augmented with indicators show better performance. In contrast, for lagged relationships, statistically significant improvements are observed only for HAR-X1 and SVR-L-X1 according to the MSE metric and for SVR-G-X1 and SVR-G-X2 under the MAE criterion.

Out-of-sample forecasting

In this section, we employ the sentiment indicators to predict BTC time series. The analysis is conducted separately for returns (Sect. "Forecasts of returns") and for volatility (Sect. "Forecasts of volatility"). Daily forecasts are formulated for the period from August 1, 2022, to March 31, 2024.

Forecasts of returns

We use the models presented in Sect. "Out-of-sample forecasting" to forecast out-of-sample returns. Similarly to the in-sample analysis, we aim to examine whether information derived from ChatGPT can improve the accuracy of BTC return forecasts compared to models without such information. The results of this comparison are presented for all

Table 13 Evaluation of the in-sample BTC variance estimation using the MCS test for models without and with contemporaneous sentiment indicators

Model	MSE	Rank	P-value	MAE	Rank	P-value	R^2	Rank
HAR	0.430	13	0.057	0.678	12	0.006	0.342	13
BMA	0.429	12	0.054	0.678	14	0.003	0.342	12
LASSO	0.430	14	0.057	0.678	13	0.004	0.342	14
SVR-G	0.427	11	0.066	0.683	15	0.002	0.338	15
SVR-L	0.447	15	0.051	0.673	11	0.008	0.343	11
HAR-X1	0.380	7	0.207*	0.652	8	0.018	0.421	8
BMA-X1	0.383	9	0.057	0.654	10	0.013	0.416	10
LASSO-X1	0.380	8	0.152*	0.652	9	0.015	0.421	9
SVR-G-X1	0.358	2	0.432*	0.615	1	1.000*	0.465	3
SVR-L-X1	0.385	10	0.251*	0.647	7	0.021	0.428	7
HAR-X2	0.360	3	0.412*	0.644	5	0.021	0.462	4
BMA-X2	0.363	6	0.304*	0.642	3	0.024	0.456	6
LASSO-X2	0.360	4	0.340*	0.644	4	0.021	0.462	5
SVR-G-X2	0.360	5	0.412*	0.641	2	0.021	0.469	1
SVR-L-X2	0.347	1	1.000*	0.647	6	0.021	0.466	2

Notes: The table reports the results of the Model Confidence Set (MCS) test for comparing the in-sample estimation accuracy of Bitcoin variance. The analysis includes the benchmark models – which do not incorporate sentiment indicators and the models with contemporaneous indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P^{inc} , P^{dec} , V^{inc} , V^{dec} , V^{high} , V^{low} , the values of MSE are multiplied by 10^5 , the values of MAE are multiplied by 10^3 , the lowest values of MSE and MAE, and the highest for R^2 are in bold, p-value is for the MCS test, * indicates that models belong to MCS with a confidence level of 0.90. The analyzed period is January 1, 2021—March 31, 2024

Table 14 Evaluation of the in-sample BTC variance estimation using the MCS test for models without and with lagged sentiment indicators

Model	MSE	Rank	P-value	MAE	Rank	P-value	R^2	Rank
HAR	0.430	8	0.503*	0.678	8	0.273*	0.342	11
BMA	0.429	7	0.503*	0.678	10	0.273*	0.342	8
LASSO	0.430	9	0.466*	0.678	9	0.273*	0.342	12
SVR-G	0.427	2	0.503*	0.683	14	0.252*	0.338	14
SVR-L	0.447	14	0.033	0.673	3	0.273*	0.343	7
HAR-X1	0.429	6	0.503*	0.678	7	0.273*	0.343	6
BMA-X1	0.429	5	0.503*	0.679	11	0.273*	0.343	4
LASSO-X1	0.433	10	0.219*	0.677	6	0.273*	0.342	10
SVR-G-X1	0.441	12	0.094	0.673	2	0.293*	0.341	13
SVR-L-X1	0.417	1	1.000*	0.696	15	0.252*	0.336	15
HAR-X2	0.429	4	0.503*	0.679	12	0.273*	0.343	5
BMA-X2	0.428	3	0.503*	0.679	13	0.273*	0.344	3
LASSO-X2	0.436	11	0.167*	0.676	5	0.273*	0.342	9
SVR-G-X2	0.445	13	0.033	0.671	1	1.000*	0.347	1
SVR-L-X2	0.450	15	0.033	0.674	4	0.273*	0.344	2

Notes: The table reports the results of the Model Confidence Set (MCS) test for comparing the in-sample estimation accuracy of Bitcoin variance. The analysis includes the benchmark models – which do not incorporate sentiment indicators and the models with lagged indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P^{inc} , P^{dec} , V^{inc} , V^{dec} , V^{high} , V^{low} , the values of MSE are multiplied by 10^5 , the values of MAE are multiplied by 10^3 , the lowest values of MSE and MAE, and the highest for R^2 are in bold, p-value is for the MCS test, * indicates that models belong to MCS with a confidence level of 0.90. The analyzed period is January 1, 2021—March 31, 2024

models jointly (using the MCS test in Table 16) and for pairs of models (using the SPA test in Table 17).

The results of the MCS test indicate that almost all models belong to the set of best models, with the exceptions of SVR-G, SVR-L, SVR-L-X2 for MSE, and SVR-L, SVR-L-X1, AR-X2, SVR-L-X2 for MAE. This means that it is not possible to identify a unique

Table 15 Evaluation of in-sample BTC variance estimation using the SPA test: models without additional variables vs. models with contemporaneous or lagged sentiment indicators

Compared models	Contemporaneous		Lagged	
	MSE	MAE	MSE	MAE
	P-value	P-value	P-value	P-value
HAR vs. HAR-X1	0.033	0.003	0.067	0.167
HAR vs. HAR-X2	0.027	0.002	0.240	0.665
BMA vs. BMA-X1	0.034	0.004	0.114	0.803
BMA vs. BMA-X2	0.025	0.001	0.121	0.823
LASSO vs. LASSO-X1	0.033	0.003	0.952	0.130
LASSO vs. LASSO-X2	0.027	0.002	0.956	0.155
SVR-G vs. SVR-G-X1	0.024	0.000	0.946	0.046
SVR-G vs. SVR-G-X2	0.026	0.001	0.956	0.056
SVR-L vs. SVR-L-X1	0.023	0.011	0.076	0.922
SVR-L vs. SVR-L-X2	0.027	0.061	0.917	0.586

Notes: The table reports the results of the Superior Predictive Ability (SPA) test for the in-sample estimation of Bitcoin variance. The test compares the benchmark models – without sentiment indicators against models incorporating contemporaneous or lagged indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} , V_{low}^{lvl} , the table shows the p-values of the SPA test for pairs of models (listed on the left). A p-value below the significance level, highlighted in bold, indicates that the estimates from the second model are more accurate than those from the first model, which serves as the benchmark. The analyzed period is January 1, 2021—March 31, 2024

best-performing model. The R^2 values are low for all models, suggesting a generally limited predictive quality of forecasts. The outcomes of the SPA test demonstrate that incorporating the sentiment indicators does not significantly increase the accuracy of BTC returns forecasts for any of the models.

Forecasts of volatility

In this section, we use the models introduced in Sect. "Out-of-sample forecasting" to produce out-of-sample forecasts of BTC variance. The same news sentiment indicators as in the earlier sections serve as explanatory variables. The results of this analysis are depicted in Table 18 and Table 19 for the MCS and SPA tests, respectively.

According to the MCS test, all models belong to the set of best models. This result does not depend on whether the MSE or MAE measure is considered. The highest values of R^2 are observed for the BMA-X1 and LASSO-X1 models; however, their value of 0.189 is only slightly higher than the values of 0.178 and 0.177 for the models without news sentiment indicators, BMA and LASSO. The SPA test indicates that models incorporating these variables do not produce significantly more accurate forecasts of volatility compared to models without them.

Discussion

Cryptocurrencies exhibit specific properties, such as very high volatility, a large share of noise in price fluctuations, a significant number of outliers, chaotic trends, a lack of distinctive patterns in time series, and the presence of bubbles (see, e.g., Kyriazis [39]; Dudek et al. [15]). For these reasons, modelling and forecasting cryptocurrencies is a formidable task. In this study, we investigate whether sentiment indicators derived from ChatGPT based on a comprehensive dataset of news headlines related to BTC influence

Table 16 Evaluation of out-of-sample forecasts for BTC returns using the MCS test for models without and with lagged sentiment indicators

Model	MSE	Rank	P-value	MAE	Rank	P-value	R^2	Rank
AR	6.616	6	0.940*	1.713	1	1.000*	0.002	10
LASSO	6.613	5	0.940*	1.716	6	0.679*	0.002	8
SVR-G	7.421	14	0.066	1.741	10	0.121*	0.027	1
SVR-L	6.774	13	0.066	1.753	11	0.000	0.000	14
AR-X1	6.622	7	0.940*	1.720	8	0.478*	0.001	12
BMA-X1	6.611	3	0.940*	1.715	4	0.826*	0.002	6
LASSO-X1	6.609	1	1.000*	1.714	2	0.925*	0.002	7
SVR-G-X1	6.694	10	0.254*	1.725	9	0.250*	0.001	13
SVR-L-X1	6.773	12	0.374*	1.761	13	0.009	0.003	3
AR-X2	6.675	9	0.827*	1.757	12	0.002	0.002	9
BMA-X2	6.610	2	0.940*	1.715	3	0.917*	0.003	5
LASSO-X2	6.611	4	0.940*	1.715	5	0.726*	0.003	4
SVR-G-X2	6.650	8	0.702*	1.718	7	0.578*	0.003	2
SVR-L-X2	6.751	11	0.066	1.762	14	0.000	0.001	11

Notes: The table reports the results of the Model Confidence Set (MCS) test for comparing the out-of-sample forecast accuracy of Bitcoin daily returns. The analysis includes the benchmark models – which do not incorporate sentiment indicators and the models with lagged indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} , V_{low}^{lvl} , the values of MSE are multiplied by 10^4 , the values of MAE are multiplied by 10^2 , the lowest values of MSE and MAE and the highest for R^2 are in bold, p-value is for the MCS test, * indicates that models belong to MCS with a confidence level of 0.90. The analyzed period is August 1, 2022—March 31, 2024

Table 17 Evaluation of out-of-sample forecasts for BTC returns using the SPA test: models without additional variables vs. models with lagged sentiment indicators

Compared models	MSE	MAE
	P-value	P-value
AR vs. AR-X1	0.557	0.780
AR vs. AR-X2	0.760	0.500
AR vs. BMA-X1	0.417	0.652
AR vs. BMA-X2	0.414	0.636
LASSO vs. LASSO-X1	0.288	0.163
LASSO vs. LASSO-X2	0.404	0.377
SVR-G vs. SVR-G-X1	0.139	0.215
SVR-G vs. SVR-G-X2	0.125	0.147
SVR-L vs. SVR-L-X1	0.479	0.665
SVR-L vs. SVR-L-X2	0.400	0.699

Notes: The table reports the results of the Superior Predictive Ability (SPA) test for the out-of-sample estimation of Bitcoin daily returns. The test compares the benchmark models – without sentiment indicators against models incorporating lagged indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} , V_{low}^{lvl} , the table shows the p-values of the SPA test for pairs of models (listed on the left). A p-value below the significance level indicates that the forecasts from the second model are more accurate than those from the first model, which serves as the benchmark. The analyzed period is August 1, 2022—March 31, 2024

its returns and variances, and whether they improve prediction accuracy (Fiszeder et al. [22] forecast BTC volatility based on 62 variables, including the market sentiment, proxied by the Google Trends index). Based on our research, we draw several conclusions:

1. The daily number of BTC-related news headlines exhibits a cyclical pattern, with significantly fewer headlines published on Saturdays and Sundays. Moreover, a noticeable

Table 18 Evaluation of out-of-sample forecasts for BTC variance using the MCS test for models without and with lagged sentiment indicators

Model	MSE	Rank	P-value	MAE	Rank	P-value	R^2	Rank
HAR	7.061	9	0.502*	3.732	2	0.912*	0.177	11
BMA	7.036	7	0.559*	3.739	5	0.777*	0.178	7
LASSO	7.056	8	0.559*	3.731	1	1.000*	0.177	10
SVR-G	7.100	14	0.559*	3.738	4	0.912*	0.175	12
SVR-L	7.065	10	0.559*	3.735	3	0.912*	0.178	9
HAR-X1	6.982	4	0.559*	3.753	9	0.658*	0.185	4
BMA-X1	6.937	2	0.948*	3.746	6	0.658*	0.189	1
LASSO-X1	6.935	1	1.000*	3.746	7	0.658*	0.189	2
SVR-G-X1	7.100	13	0.413*	3.803	14	0.322*	0.172	14
SVR-L-X1	7.066	11	0.502*	3.799	13	0.285*	0.178	8
HAR-X2	7.092	12	0.413*	3.780	12	0.574*	0.172	13
BMA-X2	7.003	6	0.559*	3.748	8	0.628*	0.182	6
LASSO-X2	6.992	5	0.559*	3.753	10	0.658*	0.183	5
SVR-G-X2	7.376	15	0.217*	3.822	15	0.396*	0.142	15
SVR-L-X2	6.978	3	0.312*	3.757	11	0.355*	0.188	3

Notes: The table reports the results of the Model Confidence Set (MCS) test for comparing the out-of-sample forecasts accuracy of Bitcoin variance. The analysis includes the benchmark models – which do not incorporate sentiment indicators and the models with lagged indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} , V_{low}^{lvl} , the values of MSE are multiplied by 10^7 , the values of MAE are multiplied by 10^4 , the lowest values of MSE and MAE and the highest for R^2 are in bold, p-value is for the MCS test, * indicates that models belong to MCS with a confidence level of 0.90. The analyzed period is August 1, 2022—March 31, 2024

Table 19 Evaluation of out-of-sample forecasts for BTC variance using the SPA test: models without additional variables vs. models with lagged sentiment indicators

Compared models	MSE	MAE
	P-value	P-value
HAR vs. HAR-X1	0.311	0.760
HAR vs. HAR-X2	0.543	0.912
BMA vs. BMA-X1	0.205	0.690
BMA vs. BMA-X2	0.259	0.865
LASSO vs. LASSO-X1	0.210	0.749
LASSO vs. LASSO-X2	0.304	0.827
SVR-G vs. SVR-G-X1	0.575	0.509
SVR-G vs. SVR-G-X2	0.838	0.932
SVR-L vs. SVR-L-X1	0.408	0.920
SVR-L vs. SVR-L-X2	0.614	0.908

Notes: The table reports the results of the Superior Predictive Ability (SPA) test for the out-of-sample estimation of Bitcoin variance. The test compares the benchmark models – without sentiment indicators against models incorporating lagged indicators, grouped into two sets – X1 and X2. The X1 set includes three variables: P^{chg} , V^{chg} , V^{lvl} , the X2 set includes six variables: P_{inc}^{chg} , P_{dec}^{chg} , V_{inc}^{chg} , V_{dec}^{chg} , V_{high}^{lvl} , V_{low}^{lvl} , the table shows the p-values of the SPA test for pairs of models (listed on the left). A p-value below the significance level indicates that the forecasts from the second model are more accurate than those from the first model, which serves as the benchmark. The analyzed period is August 1, 2022—March 31, 2024.

increase in the total daily volume of information is observed toward the end of 2023 and the beginning of 2024.

2. The sentiment indicator representing the expected change in BTC price (P^{chg}), shows a very weak correlation with the indicators related to expected BTC volatility (V^{chg} and V^{lvl}). However, when considering the disaggregated indicators, it

becomes evident that P_{inc}^{chg} is strongly correlated with V_{inc}^{chg} and V_{high}^{lvl} , which represent indicators related to expected increases in volatility and high volatility, respectively.

3. According to correlation and the Granger causality analyses, and the significance of regressors in the extended AR, BMA, and LASSO models, most sentiment indicators related to the expected change in BTC price have a significant positive influence on BTC returns. In contrast, BTC returns are only marginally affected by the indicators related to expected volatility.

4. All considered return models, i.e., AR, BMA, LASSO, and SVR, extended with the contemporaneous sentiment indicators, estimate BTC returns in-sample more accurately than models that do not incorporate this information. This conclusion is less definitive in the case of the lagged indicators, since the addition of such variables improves the quality of estimates only in some models. These results show that BTC news headlines processed by ChatGPT have a significant influence on BTC returns; however, their impact is primarily confined to the day of publication.

5. The SVR model with a radial basis function kernel exhibits superior performance among return models with contemporaneous variables representing the sentiment indicators. However, it is impossible to identify a significantly better model for models extended with the lagged explanatory indicators.

6. The correlation analysis, the Granger causality analysis, and the significance of regressors in the extended HAR, BMA, and LASSO models indicate that most sentiment indicators significantly influence BTC volatility. The impact of the indicator related to the expected change in BTC price is negative, which aligns with the leverage effect observed in other financial time series, such as stocks or fiat currencies (see, e.g., Fałdziński et al. [16]).

7. All models augmented with the contemporaneous sentiment indicators show better performance in the in-sample estimation of BTC variance. In contrast, only certain models extended with the lagged indicators demonstrate improved accuracy in estimating BTC volatility compared to the models without additional information. This implies that BTC news headlines, when synthesized by ChatGPT and converted into numerical indicators, impact BTC variance mainly on the day of publication.

8. For volatility models incorporating both contemporaneous or lagged indicators, no single model emerges as superior across all evaluation metrics for in-sample estimation of BTC variance.

9. The models incorporating news sentiment indicators fail to produce significantly more accurate out-of-sample forecasts of returns and variances compared to the models without such variables. This finding is consistent with the efficient market hypothesis, which posits that prices reflect all available information, making it challenging to consistently generate abnormal returns based solely on past news.

10. The discrepancy between strong in-sample effects and the lack of out-of-sample forecasting gains is an important and economically meaningful result. The ChatGPT-derived sentiment indicators significantly improve model fit, suggesting that they capture relevant contemporaneous information embedded in news headlines. However, this information appears to be rapidly incorporated into Bitcoin prices, limiting its predictive

usefulness. This finding is consistent with the view that news-based sentiment primarily reflects market reactions rather than providing forward-looking signals. Moreover, the cryptocurrency market is characterized by frequent regime shifts, high volatility, and a low signal-to-noise ratio. Under such conditions, relationships identified in-sample may not remain stable out-of-sample, even when they are statistically significant. From this perspective, the absence of out-of-sample forecasting gains should not be interpreted as a failure of the sentiment indicators, but rather as evidence of market efficiency with respect to publicly available news.

Methodological limitations

While this study provides new evidence on the role of ChatGPT-derived news sentiment in modelling Bitcoin returns and volatility, several methodological limitations should be noted.

First, the analysis relies exclusively on news headlines rather than full article texts, abstracts, or additional metadata. This approach ensures a consistent information structure across a large dataset and is consistent with prior literature using headline-based sentiment indicators. However, sentiment extracted at the headline level may overlook contextual nuances contained in full articles, potentially leading to information loss or simplified emotional signals.

Second, the sentiment indicators are derived using manually designed prompts and a single large language model (ChatGPT 4.0). Although the prompts are conceptually motivated and applied consistently across the dataset, the resulting sentiment classifications may be sensitive to prompt wording and model-specific behavior. Moreover, the reliance on a single LLM version implies that the results may be sensitive to alternative model architectures or future updates.

Third, some sentiment indicators are constructed using contemporaneous news information, which requires caution in interpretation. While the absence of statistically significant out-of-sample forecasting gains mitigates concerns regarding predictive overfitting, contemporaneous sentiment measures may primarily capture information already reflected into market prices rather than genuinely predictive signals.

Finally, the study focuses exclusively on news headlines and does not incorporate other information channels such as social media content, on-chain metrics, or macro-economic announcements. As a result, the extracted sentiment indicators reflect only a subset of the information environment surrounding Bitcoin markets.

These limitations do not invalidate the empirical findings but delineate the scope within which the results should be interpreted.

Conclusions

ChatGPT is an advanced AI language model that serves as a valuable tool for investors. By leveraging its natural language processing capabilities, ChatGPT can identify historical patterns and synthesize vast amounts of information into clear summaries, helping investors efficiently navigate the overwhelming volume of news and data. It has

the potential to reduce information costs, facilitate access to financial knowledge, and enhance individual investors' ability to make informed investment decisions.

In this paper, we examine ChatGPT's capacity to process financial news and assess its effectiveness in uncovering practical investment opportunities. We compile an extensive dataset of Bitcoin-related news headlines and leverage ChatGPT to evaluate the content of each headline. The resulting assessments are translated into numerical sentiment indicators, which are subsequently applied in the statistical and machine learning models to analyze Bitcoin's returns and variances. In our study, we use AR, HAR, BMA, LASSO, and SVR models.

The correlation analysis, Granger causality analysis, and the significance of regressors in the applied return and volatility models indicate that most sentiment indicators related to the expected change in BTC price have a significant positive influence on BTC returns. Similarly, in most cases, all these indicators significantly impact BTC volatility.

All models incorporating contemporaneous indicators demonstrate improved performance in estimating BTC returns and variance within the sample. However, results differ for lagged indicators, as only certain models that include these variables provide more accurate estimates.

On the other hand, models that include the news sentiment indicators do not yield significantly more accurate out-of-sample forecasts for returns and variances compared to the models that lack these variables. This result suggests that ChatGPT-based sentiment measures primarily capture contemporaneous market-relevant information rather than persistent predictive signals, consistent with semi-strong market efficiency.

Our study can be extended in several possible directions in the future. Firstly, other statistical and machine learning models, such as autoregressive fractionally integrated moving average (ARFIMA), generalized autoregressive conditional heteroscedasticity (GARCH), random forest, or neural networks, could be explored. Different models vary in their ability to capture the characteristics of BTC time series and to represent the complex relationships among BTC returns, volatilities, and sentiment indicators.

Secondly, other kinds of information can be used and evaluated by ChatGPT, such as price quotes from other financial markets, macroeconomic news, uncertainty and risk indices, news extracted from the X platform, and Google Trends.

Thirdly, the approach used in the paper can be applied to other assets, such as stocks, traditional currencies, or commodities. Unlike cryptocurrencies, these assets possess fundamental values, which may make forecasting their returns and volatilities based on incoming news assessed by ChatGPT more attainable.

Appendix A

See Tables [20](#), [21](#), [22](#), [23](#).

Appendix B

See Figures [7](#) and [8](#).

Table 20 Contingency tables for ChatGPT-derived sentiment categories and Crypto News sentiment labels

QUERY 1				QUERY 2				QUERY 3			
	-1	0	1		-1	0	1		-1	0	1
-1	5819	20,277	1171	-1	1565	18,400	7302	-1	1383	22,680	3204
0	1033	26,184	1389	0	1530	24,003	3073	0	1663	25,782	1161
1	1019	48,756	8133	1	3928	42,172	11,808	1	3276	49,325	5307

Notes: Columns correspond to ChatGPT-derived sentiment categories (−1, 0, +1), while rows report sentiment labels provided by the Crypto News database

Table 21 Summary statistics for the sentiment indicators across bull and bear market regimes

Variable	Mean	Minimum	Maximum	Standard deviation	Skewness	Excess kurtosis
Bull market						
P^{chg}	0.050	-0.372	0.316	0.093	-0.429	1.389
V^{chg}	0.140	-0.095	0.471	0.083	0.446	0.493
V^{lvl}	0.030	-0.162	0.264	0.061	0.480	0.500
P_{dec}^{chg}	-0.059	-0.397	0.000	0.054	-2.052	6.350
P_{inc}^{chg}	0.105	0.000	0.329	0.061	0.873	0.595
V_{dec}^{chg}	-0.073	-0.273	0.000	0.034	-0.873	2.279
V_{inc}^{chg}	0.200	0.000	0.485	0.072	0.469	0.419
V_{low}^{lvl}	-0.059	-0.297	0.000	0.034	-1.165	3.908
V_{high}^{lvl}	0.085	0.000	0.289	0.048	0.898	0.931
Bear market						
P^{chg}	-0.024	-0.479	0.227	0.095	-0.757	2.024
V^{chg}	0.112	-0.059	0.389	0.076	0.610	0.798
V^{lvl}	0.018	-0.128	0.274	0.058	0.674	1.323
P_{dec}^{chg}	-0.095	-0.493	0.000	0.072	-1.578	3.893
P_{inc}^{chg}	0.066	0.000	0.242	0.045	1.235	1.799
V_{dec}^{chg}	-0.082	-0.200	0.000	0.035	-0.362	-0.024
V_{inc}^{chg}	0.178	0.038	0.438	0.065	0.664	0.811
V_{low}^{lvl}	-0.061	-0.231	0.000	0.036	-0.805	1.229
V_{high}^{lvl}	0.074	0.000	0.301	0.045	1.184	2.483

Notes: The analyzed period spans from January 1, 2021, to March 31, 2024. Bull market periods are defined as January 1, 2021–April 16, 2021; July 21, 2021–November 19, 2021; and November 22, 2022–March 31, 2024. Bear market periods are defined as April 17, 2021–July 20, 2021, and November 20, 2021–November 21, 2022

Table 22 Estimation results for the AR-X, BMA-X, and LASSO-X models for BTC returns

Model	Contemporaneous indicators from the X1 set							
	Const	$r_{d,t-1}$	P_t^{chg}	V_t^{chg}	V_t^{lvl}			
AR-X1	0.001 (0.591)	-0.252 (0.000)	0.179 (0.000)	-0.028 (0.033)	-0.001 (0.937)			
BMA-X1	0.000 (1.000)	-0.251 (1.000)	0.178 (1.000)	-0.017 (0.606)	-0.001 (0.118)			
LASSO-X1	6.98e-04	-0.244	0.176	-0.026	0			
Lagged indicators from the X1 set								
	Const	$r_{d,t-1}$	P_{t-1}^{chg}	V_{t-1}^{chg}	V_{t-1}^{lvl}			
AR-X1	-0.003 (0.157)	-0.074 (0.021)	0.019 (0.079)	0.025 (0.091)	-0.004 (0.854)			
BMA-X1	0.000 (1.000)	-0.002 (0.042)	0.000 (0.018)	0.002 (0.076)	0.000 (0.015)			
LASSO-X1	-0.002	-0.056	0.013	0.020	0			
Contemporaneous indicators from the X2 set								
	Const	$r_{d,t-1}$	$P_{inc,t}^{chg}$	$P_{dec,t}^{chg}$	$V_{inc,t}^{chg}$	$V_{dec,t}^{chg}$	$V_{high,t}^{lvl}$	$V_{low,t}^{lvl}$
AR-X2	0.005 (0.167)	-0.263 (0.000)	0.109 (0.000)	0.241 (0.000)	0.020 (0.327)	-0.031 (0.211)	-0.003 (0.930)	0.034 (0.172)
BMA-X2	0.006 (1.000)	-0.261 (1.000)	0.124 (1.000)	0.228 (1.000)	0.001 (0.051)	-0.001 (0.053)	0.000 (0.034)	0.002 (0.068)
LASSO-X2	0.005	-0.253	0.113	0.232	0.012	-0.024	0	0.028
Lagged indicators from the X2 set								
	Const	$r_{d,t-1}$	$P_{inc,t-1}^{chg}$	$P_{dec,t-1}^{chg}$	$V_{inc,t-1}^{chg}$	$V_{dec,t-1}^{chg}$	$V_{high,t-1}^{lvl}$	$V_{low,t-1}^{lvl}$
AR-X2	-0.005 (0.222)	-0.076 (0.019)	0.039 (0.167)	0.004 (0.882)	0.032 (0.180)	-0.008 (0.768)	-0.037 (0.262)	0.023 (0.430)
BMA-X2	0.000 (1.000)	-0.001 (0.026)	0.001 (0.032)	0.000 (0.005)	0.001 (0.040)	0.000 (0.005)	0.000 (0.008)	0.000 (0.006)
LASSO-X2	-0.003	-0.045	0.024	0	0.011	0	0	0.002

Notes: The table reports estimation results for the AR, BMA, and LASSO models extended with the sentiment indicators, applied to daily Bitcoin returns. Panels distinguish between contemporaneous and one-period lagged sentiment indicators, and between the aggregated indicator set (X1) and the disaggregated indicator set (X2). The dependent variable is the daily logarithmic return of Bitcoin $r_{d,t}$. For the AR models, p-values are reported in parentheses, whereas for the BMA models the corresponding posterior inclusion probabilities (PIPs) are shown. Coefficients of AR models with p-values below 0.1, indicating statistical significance, are highlighted in bold. The analyzed period is January 1, 2021—March 31, 2024

Table 23 Estimation results for the HAR, BMA, and LASSO models for BTC volatility

Model	Contemporaneous indicators from the X1 set								
	Const	$\ln RV_{d,t-1}$	$\ln RV_{w,t-1}$	$\ln RV_{m,t-1}$	P_t^{chg}	V_t^{chg}	V_t^{tol}		
HAR-X1	-0.703 (0.000)	0.389 (0.000)	0.312 (0.000)	0.240 (0.000)	-1.003 (0.000)	0.553 (0.106)	2.224 (0.000)		
BMA-X1	-0.706 (1.000)	0.387 (1.000)	0.317 (1.000)	0.232 (0.999)	-0.995 (0.999)	0.217 (0.392)	2.476 (1.000)		
LASSO-X1	-0.706	0.389	0.313	0.240	-1.000	0.548	2.220		
Lagged indicators from the X1 set									
Const	$\ln RV_{d,t-1}$	$\ln RV_{w,t-1}$	$\ln RV_{m,t-1}$	P_{t-1}^{chg}	V_{t-1}^{chg}	V_{t-1}^{tol}			
HAR-X1	-0.542 (0.009)	0.415 (0.000)	0.378 (0.000)	0.157 (0.003)	-0.042 (0.858)	0.229 (0.514)	-0.328 (0.496)		
BMA-X1	-0.623 (1.000)	0.415 (1.000)	0.404 (1.000)	0.115 (0.737)	-0.002 (0.036)	0.002 (0.039)	-0.012 (0.045)		
LASSO-X1	-0.690	0.406	0.375	0.146	0	0	0		
Contemporaneous indicators from the X2 set									
Const	$\ln RV_{d,t-1}$	$\ln RV_{w,t-1}$	$\ln RV_{m,t-1}$	$P_{inc,t}^{chg}$	$P_{dec,t}^{chg}$	$V_{inc,t}^{chg}$	$V_{dec,t}^{chg}$	$V_{high,t}^{tol}$	$V_{low,t}^{tol}$
HAR-X2	-0.557 (0.006)	0.375 (0.000)	0.320 (0.000)	0.289 (0.000)	3.203 (0.000)	-2.104 (0.000)	1.159 (0.067)	2.149 (0.004)	0.521 (0.418)
BMA-X2	-0.613 (1.000)	0.375 (1.000)	0.325 (1.000)	0.289 (1.000)	3.214 (1.000)	-1.900 (0.961)	0.520 (0.435)	1.823 (0.846)	0.114 (0.196)
LASSO-X2	-0.595	0.375	0.320	0.286	2.969	-1.876	1.073	2.062	0.463
Lagged indicators from the X2 set									
Const	$\ln RV_{d,t-1}$	$\ln RV_{w,t-1}$	$\ln RV_{m,t-1}$	$P_{inc,t-1}^{chg}$	$P_{dec,t-1}^{chg}$	$V_{inc,t-1}^{chg}$	$V_{dec,t-1}^{chg}$	$V_{high,t-1}^{tol}$	$V_{low,t-1}^{tol}$
HAR-X2	-0.588 (0.006)	0.425 (0.000)	0.372 (0.000)	0.149 (0.007)	-0.922 (0.173)	0.762 (0.185)	0.682 (0.226)	0.051 (0.948)	-0.157 (0.817)
BMA-X2	-0.663 (1.000)	0.416 (1.000)	0.417 (1.000)	0.096 (0.612)	-0.009 (0.025)	0.005 (0.021)	0.000 (0.020)	-0.007 (0.023)	-0.002 (0.018)
LASSO-X2	-0.721	0.404	0.374	0.142	0	0	0	0	0

Notes: The table reports estimation results for the HAR, BMA, and LASSO models extended with the sentiment indicators, applied to Bitcoin volatility. Panels distinguish between contemporaneous and one-period lagged sentiment indicators, and between the aggregated indicator set (X1) and the disaggregated indicator set (X2). The dependent variable is the logarithm of realized BTC variance $\ln RV_{d,t}$. The p-values for HAR and the PIP for BMA are given in parentheses. Coefficients of HAR models with p-values below 0.1, indicating statistical significance, are highlighted in bold. The analyzed period is January 1, 2021—March 31, 2024

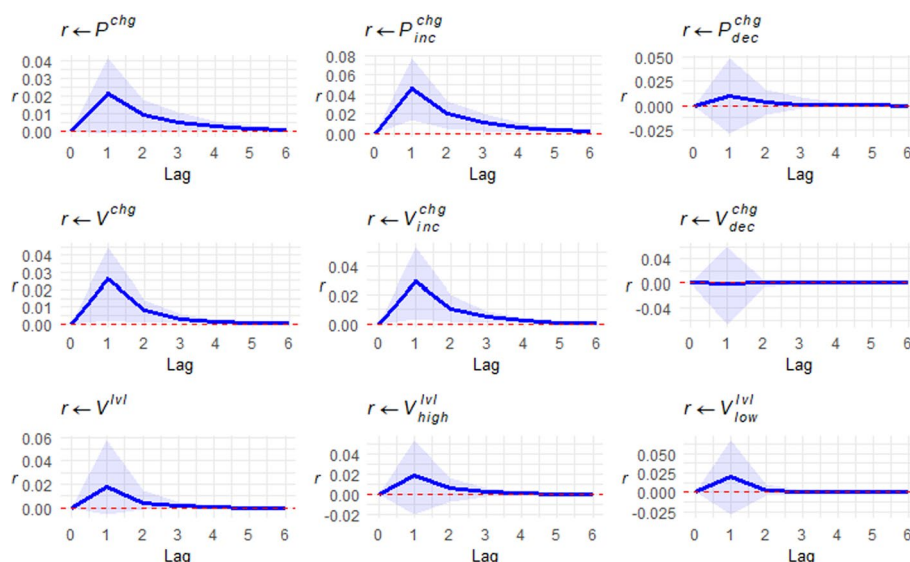


Fig. 7 Generalized impulse response functions of Bitcoin returns to sentiment indicators. The figure reports generalized impulse response functions (GIRFs) of Bitcoin returns to sentiment shocks derived from bivariate VAR models. Confidence bands correspond to 95% bootstrap intervals. The zero line indicates no response

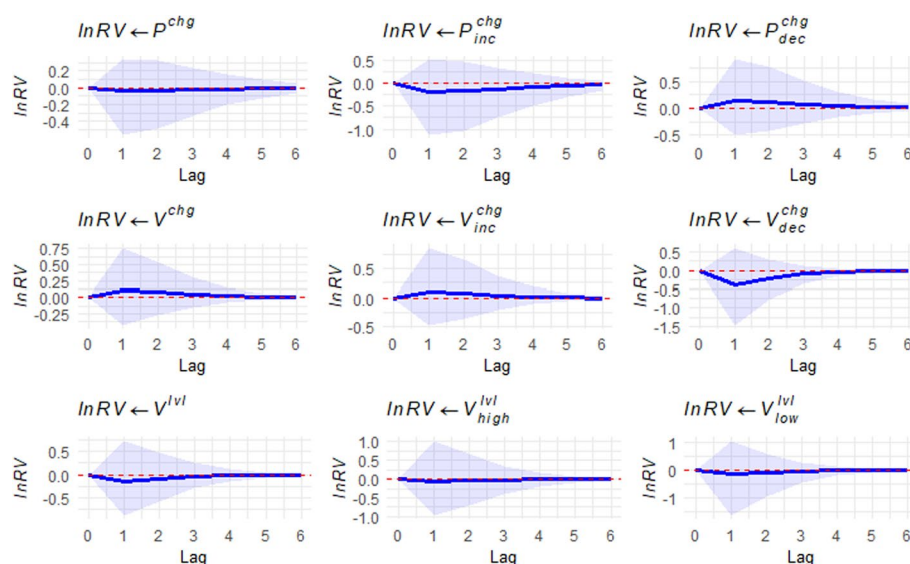


Fig. 8 Generalized impulse response functions of lnRV to sentiment indicators. The figure reports generalized impulse response functions (GIRFs) of lnRV to sentiment shocks estimated from VAR models. Confidence bands correspond to 95% bootstrap intervals. The horizontal zero line indicates no response

Acknowledgements

The authors thank four anonymous referees for their helpful comments and advice.

Author contributions

****PF:**** Conceptualization, Methodology, Validation, Investigation, Writing, Supervision, Funding acquisition, Project administration. ****WO:**** Conceptualization, Methodology, Software, Validation, Investigation, Writing, Visualization, Supervision, Funding acquisition. ****RP:**** Conceptualization, Methodology, Software, Validation, Investigation, Writing, Visualization, Supervision, Data curation.

Funding

Piotr Fiszeder and Radosław Pietrzyk were supported by the National Science Centre project number 2021/43/B/HS4/00353 entitled “Robust methods for range-based models—Risk and comovement analysis on the cryptocurrency market”; Piotr Fiszeder was supported by Institutional support of the Prague University of Economics and Business no. IP100040; Witold Orzeszko was supported by the National Science Centre project number 2019/35/B/HS4/00642 entitled “Forecasting financial markets using a machine learning approach”.

Data availability

The data are available at the following address: <https://doi.org/10.18150/WEATTF>.

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

Competing Interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI's ChatGPT in order to improve readability and polish the language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Received: 25 August 2025 / Accepted: 2 February 2026

Published online: 30 April 2026

References

1. Ahamad R, Mishra KN. Exploring sentiment analysis in handwritten and E-text documents using advanced machine learning techniques: a novel approach. *J Big Data*. 2025;12:11.
2. Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. *Knowl Based Syst*. 2021;226:107134.
3. Bond, S. A., Klok, H., Zhu, M., 2023. Large language models and financial market sentiment. Available at SSRN: <https://ssrn.com/abstract=4584928>.
4. Chen, Y., Kelly, B. T., Xiu, D., 2022. Expected returns and large language models. Available at SSRN: <https://ssrn.com/abstract=4416687>.
5. Chen, J., Tang, G., Zhou, G., Zhu, W., 2024. ChatGPT, stock market predictability and links to the macroeconomy. Olin Business School Center for Finance & Accounting Research Paper No. 2023/18. Available at SSRN: <https://ssrn.com/abstract=4660148>.
6. Chen B, Wu Z, Zhao R. From fiction to fact: the growing role of generative AI in business and finance. *J Chin Econ Bus Stud*. 2023;21(4):471–96.
7. Cheng, Y., Zeng, Y., Zou, J., 2024. Harnessing ChatGPT for predictive financial factor generation: A new frontier in financial analysis and forecasting. *The British Accounting Review*, 101507.
8. Chiu YL, Gao X, Liu HC, Zhai Q. Financial literacy of ChatGPT: evidence through financial news. *Financ Res Lett*. 2025;78:107088.
9. Chou WH, Feng Z, Li B, Liu F. A first look at financial data analysis using ChatGPT-4o. *J Risk Financ Manag*. 2025;18(2):99.
10. Cohen G, Aiche A. Predicting the Bitcoin's price using AI. *Front Artif Intell*. 2025;8:1519805.
11. Corsi F. A simple approximate long-memory model of realized volatility. *J Financ Econom*. 2009;7:174–96.
12. Dong MM, Stratopoulos TC, Wang VX. A scoping review of ChatGPT research in accounting and finance. *Int J Account Inf Syst*. 2024;55:100715.
13. Drašković D, Milanović S. Aspect-based sentiment analysis of user-generated content from a microblogging platform. *J Big Data*. 2025;12:186.
14. Du K, Zhao Y, Mao R, Xing F, Cambria E. Natural language processing in finance: A survey. *Information Fusion*. 2025;115:102755.
15. Dudek G, Fiszeder P, Kobus P, Orzeszko W. Forecasting cryptocurrencies volatility using statistical and machine learning methods: a comparative study. *Appl Soft Comput*. 2024;151:111132.
16. Faldziński M, Fiszeder P, Molnár P. Improving volatility forecasts: evidence from range-based models. *The North American Journal of Economics and Finance*. 2024;69:102019.
17. Fatouros G, Metaxas K, Soldatos J, Kyriazis D. Can large language models beat wall street? Evaluating GPT-4's impact on financial decision-making with MarketSenseAI. *Neural Comput Appl*. 2025;37:24893–918.
18. Fatouros G, Soldatos J, Kouroumali K, Makridis G, Kyriazis D. Transforming sentiment analysis in the financial domain with ChatGPT. *Machine Learning with Applications*. 2023;14:100508.
19. Fieberg, C., Hornuf, L., Streich, D., 2023. Using GPT-4 for financial advice. CESifo Working Paper No. 10529. Available at SSRN: <https://ssrn.com/abstract=4499485>.
20. Fiszeder P, Małecka M, Molnár P. Robust estimation of the range-based GARCH model: forecasting volatility, value at risk and expected shortfall of cryptocurrencies. *Econ Model*. 2024;141:106887.
21. Fiszeder P, Orzeszko W. Covariance matrix forecasting using support vector regression. *Appl Intell*. 2021;51:7029–42.
22. Fiszeder P, Orzeszko W, Pietrzyk R, Dudek G. Identification of Bitcoin volatility drivers using statistical and machine learning methods. *Appl Soft Comput*. 2026;188:14384.
23. Gao, T., Fisch, A., Chen, D., 2021. Making pre-trained language models better few-shot learners. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3816–3830.
24. Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. 1969;37(3):424–38.

25. Han D, Guo W, Chen H, Wang B, Guo Z. LEST: large language models and spatio-temporal data analysis for enhanced Sino-US exchange rate forecasting. *Int Rev Econ Financ*. 2024;96:103508.
26. Hansen PR. A test for superior predictive ability. *Journal of Business & Economic Statistics*. 2005;23(4):365–80.
27. Hansen PR, Lunde A, Nason JM. The model confidence set. *Econometrica*. 2011;79:453–97.
28. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Stat Sci*. 1999;14(4):382–417.
29. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R. Berlin: Springer; 2021.
30. Jeong W, Park S, Lee S, Son B, Lee J, Ko H. Influence and predictive power of sentiment: evidence from the lithium market. *Fin Res Lett*. 2024;68:105871.
31. Kang R, Hwang S, Shin J. 2024. Cryptocurrency prices and news sentiment: Which exerts influence on the other? Available at SSRN: <https://ssrn.com/abstract=4714113>.
32. Kang W, Yuan X, Zhang X, Chen Y, Li J. ChatGPT-based sentiment analysis and risk prediction in the Bitcoin market. *Procedia Computer Science*. 2024;242:211–8.
33. Kim JH. What if ChatGPT were a quant asset manager. *Fin Res Lett*. 2023;58:104580.
34. Kirtac K, Germano G. Sentiment trading with large language models. *Fin Res Lett*. 2024;62:105227.
35. Ko H, Lee J. Can ChatGPT improve investment decisions? From a portfolio management perspective. *Fin Res Lett*. 2024;64:105433.
36. Kuhn M, Johnson K. *Applied Predictive Modelling*. New York: Springer; 2013.
37. Li X, Feng H, Yang H, Huang J. Can ChatGPT reduce human financial analysts' optimistic biases? *Econ Polit Stud*. 2023;12(1):20–33.
38. Li X, Jia S, Xue F, Hu N. 2025. How investors' ChatGPT attention influence stock market? A liquidity perspective. *Research in International Business and Finance*. 77, Part B, 102939.
39. Kyriazis NA. A survey on volatility fluctuations in the decentralized cryptocurrency financial assets. *J Risk Financ Manag*. 2021;14:293.
40. Leamer EE. Regression selection strategies and revealed priors. *J Am Stat Assoc*. 1978;73(363):580–7.
41. LoGrasso M. Could ChatGPT have earned abnormal returns? A retrospective test from the U.S. stock market. *Mod Finance*. 2025;3(3):112–32.
42. Lopez-Lira A, Tang Y. 2023. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*.
43. Ma F, Lyu Z, Li H. Can ChatGPT predict Chinese equity premiums? *Financ Res Lett*. 2024;65:105631.
44. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. *J Am Stat Assoc*. 1988;83(404):1023–32.
45. Nguyen B T, Chu T T, Ha S X, Nguyen A T. 2025. Decoding market sentiment: the power of ChatGPT in explaining Bitcoin returns from X data. *China Finance Review International*. Vol. ahead-of-print No. ahead-of-print.
46. Niszczota P, Abbas S. GPT has become financially literate: insights from financial literacy tests of GPT and a preliminary test of how people use it as a source of advice. *Financ Res Lett*. 2023;58:104333.
47. Oehler A, Horn M. Does ChatGPT provide better advice than robo-advisors? *Financ Res Lett*. 2024;60:104898.
48. Pele DT, Bolovăneanu V, Lin MB, Ren R, Ginavar AT, Spilak B, et al. In the beginning was the word: LLM-VaR and LLM-ES. *Expert Syst Appl*. 2026;295:128676.
49. Pelster M, Val J. Can ChatGPT assist in picking stocks? *Financ Res Lett*. 2024;59:104786.
50. Polyzos E, Siriopoulos C. Autoregressive random forests: machine learning and lag selection for financial research. *Comput Econ*. 2024;64:225–62.
51. Qin C, Zhang A, Zhang Z, Chen J, Yasunaga M, Yang D. 2023. Is ChatGPT a general-purpose natural language processing task solver? In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1339–1384.
52. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models, *Journal of the American Statistical Association*. *J Am Stat Assoc*. 1997;92(437):179–91.
53. Romanko O, Narayan A, Kwon RH. Chatgpt-based investment portfolio selection. *Oper Res Forum*. 2023;4:91.
54. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol*. 1996;58(1):267–88.
55. Vapnik VN. *The nature of statistical learning theory*. Berlin/Heidelberg, Germany: Springer; 1995.
56. Vapnik VN, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer M, Jordan M, Petsche T, editors. *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press; 1997.
57. Yang K, Deng R, Wei Y, Wang S. The power of ChatGPT in processing text: evidence from analysis and prediction in the exchange rate markets. *Financ Innov*. 2025;11:118.
58. Zellner A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel PK, Zellner A, editors. *Basic Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland; 1986.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.