

Correlation Analysis of Social Media Sentiment on Short-Term Bitcoin Price Movements

Congyuan Shi *

Department of Data Science & Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, 999077, China

* Corresponding Author Email: 22100746d@connect.polyu.hk

Abstract. With the rapid development of digital finance, social media sentiment has gradually become a potential driver of cryptocurrency price fluctuations, but its impact mechanism on the short-term price of Bitcoin has not been fully explored. This article takes Reddit comment sentiment data from August to September 2021 and Bitcoin's short-term price (daily closing price) as a sample and explores the linear and nonlinear relationship between the two. Through correlation matrix, linear regression and generalised addability model (GAM) analysis, it is found that there is a weak positive linear correlation between emotion and price (Pearson coefficient = 0.3017, $p < 0.001$), and there is a significant nonlinear effect (the intensity of emotional influence varies with the market state). The Granger causal test confirms that emotions have a one-way causal relationship with the price of Bitcoin. Further verification of the machine learning model, such as random forest, Long Short-Term Memory (LSTM) and the inclusion of emotional variables can improve the accuracy of price prediction. This study enriches the literature in the field of cryptocurrency market behaviour finance and provides an empirical basis for investors and regulators to monitor market sentiment risks.

Keywords: Bitcoin; social media sentiment; short-term price; non-linear relationship; machine learning.

1. Introduction

Over its relatively short existence, the cryptocurrency market has developed at an unprecedented pace, with erratic fluctuations along the way.[1] In recent years, the cryptocurrency market has become a hot topic in financial research due to its high volatility and innovative nature. Over the past decade, a variety of digital currencies have emerged, such as Bitcoin, Litecoin, PeerCoin, AuroraCoin, Dogecoin and Ripple. Among these, Bitcoin stands out as the most prominent—whether in terms of its impressive price development or its price volatility [2]. Bitcoin, as the cryptocurrency with the largest market capitalisation, had a market value exceeding £2.1 trillion as of 31 August 2025. Its price is influenced by multiple factors, including macroeconomic policies, policy events, and trading volume. Furthermore, Bein Crypto and the International Monetary Fund have also outlined the key factors that influence Bitcoin's price, including supply and demand, news and public opinion, emotions, innovation and competition, regulatory and security, institutional adoption, and market manipulation [3].

Since its inception in 2009, Bitcoin's price has experienced extreme volatility: surging from \$0.003 per coin in 2010 to \$110,000 per coin in August 2025, with its market capitalisation once exceeding \$2.2 trillion. This extreme volatility stems both from fundamental factors such as its decentralised nature and supply-demand dynamics, and is closely tied to market sentiment—particularly in the era of social media where information spreads rapidly. Collective resonance in investor sentiment can trigger abrupt price movements within a short timeframe. For instance, Relevant scholars show that the investor sentiment of Bitcoin can affect its return [4]. Relevant scholars indicate that the investor sentiment of stocks can predict Bitcoin volatility [5]. Social media platforms serve not merely as communication tools—they have evolved into barometers for investor sentiment, and this sentiment, in turn, can exert immediate and substantial impacts on market behavior. Research into this phenomenon indicates a remarkably direct correlation between the sentiments voiced on these platforms and the subsequent fluctuations in cryptocurrency prices [6]. As social media platforms such as Reddit and Twitter have become key channels for investors to access

information and exchange views, social media sentiment is regarded as a potential driver of short-term price fluctuations. “Alternative data” — which refers to data and information outside of the usual scope of securities pricing, company fundamentals, or macroeconomic indicators — is playing an increasingly important role for asset managers, traders, and decision makers. Social media is now ranked as one of the top categories of alternative data currently used by hedge funds [7]. In July 2021, the price of Bitcoin experienced significant volatility, at one point falling over 50% from its historical peak. Subsequently, discussion volume on platforms such as Reddit increased markedly, exhibiting a strong correlation with Bitcoin's price movements. However, existing research analysing this relationship faces limitations: it often overlooks potential non-linear effects within price fluctuations, such as the reversal effect that may arise from phenomena like ‘overheated sentiment’. Relevant scholars note that the in-sample performance of simple linear models is difficult to extrapolate to out-of-sample data, and their predictive capability typically falls short of simple historical mean forecasting. The uncertainty inherent in the model and the instability of its parameters result in the suboptimal predictive performance of traditional linear models [8]. Machine learning methods not only extract meaningful information from numerous input variables and identify complex non-linear relationships but also possess automated feature selection capabilities. This enables adaptive learning, continuously updating models to accommodate dynamic market shifts [9]. Consequently, investigating both linear and non-linear relationships between social media sentiment and short-term Bitcoin prices, while validating predictive value through machine learning approaches, holds significant theoretical and practical importance.

Relevant scholars find a striking positive correlation between the price level of BitCoin and the searched terms as well as a dynamic relationship which is bidirectional [10]. Most literature employs linear frameworks such as Ordinary least squares (OLS) and vector autoregression (VAR) to examine the correlation between sentiment and prices. However, such models fail to capture non-linear phenomena like ‘sentiment overheating’—for instance, extreme optimism may trigger reversal effects rather than sustaining price increases. Traditional time series models (such as Generalized Autoregressive Conditional Heteroskedasticity) struggle to handle structural breaks in cryptocurrency markets (e.g., the crash triggered by Musk's tweet in May 2021). Although machine learning holds advantages in non-linear modelling, existing research predominantly remains confined to single model testing (e.g., using only LSTM for price forecasting), lacking systematic exploration of the progression from linear models to non-linear models to machine learning. This paper aims to comprehensively reveal the mechanism by which sentiment influences short-term prices through the integration of multiple models and high-frequency data analysis.

2. Empirical Analysis

Figure 1 presents a scatter plot of Reddit comment sentiment versus Bitcoin closing prices. The horizontal axis (X-axis) represents Reddit comment sentiment values, while the vertical axis (Y-axis) denotes Bitcoin closing prices. Blue dots indicate individual observations. The data distribution exhibits an overall banded pattern sloping from the lower left to the upper right, though with a relatively gentle gradient. Significant overlap exists in price distributions across different sentiment intervals, failing to establish a strict linear relationship. When sentiment values range from 0 to 5, prices cluster predominantly between \$38,000 and \$42,000, though a small number of outliers exceed \$45,000 (indicating that low sentiment does not necessarily guarantee depressed prices, which may be driven by other factors). The broadest price distribution occurs at sentiment values of 5–10, spanning \$38,000–\$48,000 and nearly filling the entire price range. This indicates that price fluctuations during neutral sentiment are primarily governed by non-sentiment factors. When sentiment values range from 10 to 17.5, prices cluster between \$44,000 and \$50,000, though some prices remain below \$42,000 (high sentiment coexists with low prices, potentially reflecting sentiment lag or market digestion of positive news). Overall distribution reveals that median prices corresponding to sentiment values (>10) are significantly higher than those with low sentiment (<5),

indicating a weak positive correlation between sentiment and price. Price dispersion peaks within the 5–10 sentiment range without discernible trends, suggesting sentiment may exert a threshold effect on prices (only significantly driving prices when exceeding a critical sentiment threshold). Prices below \$40,000 were observed at sentiment values of 15–17.5, while prices exceeding \$45,000 occurred at sentiment values of 0–5. This indicates sentiment is not the sole price determinant, with factors such as macroeconomic policies potentially mitigating its influence.

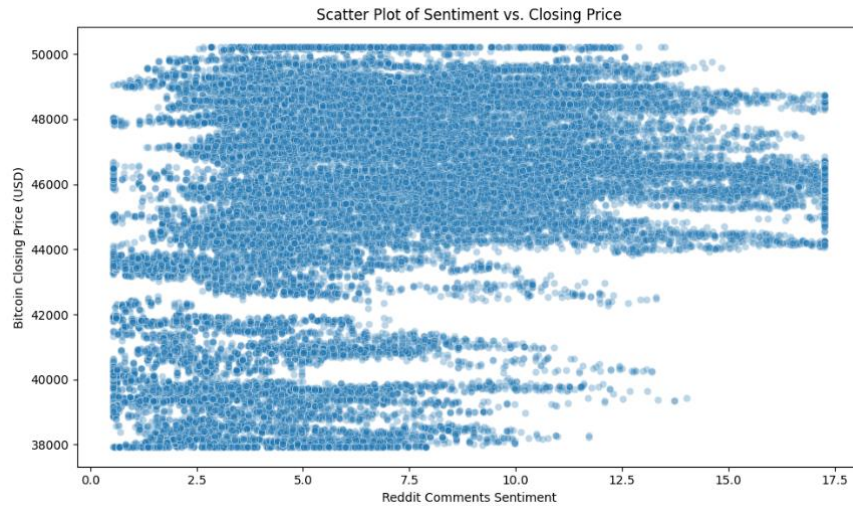


Fig. 1 Scatter diagram of reddit comment emotion and bitcoin closing price

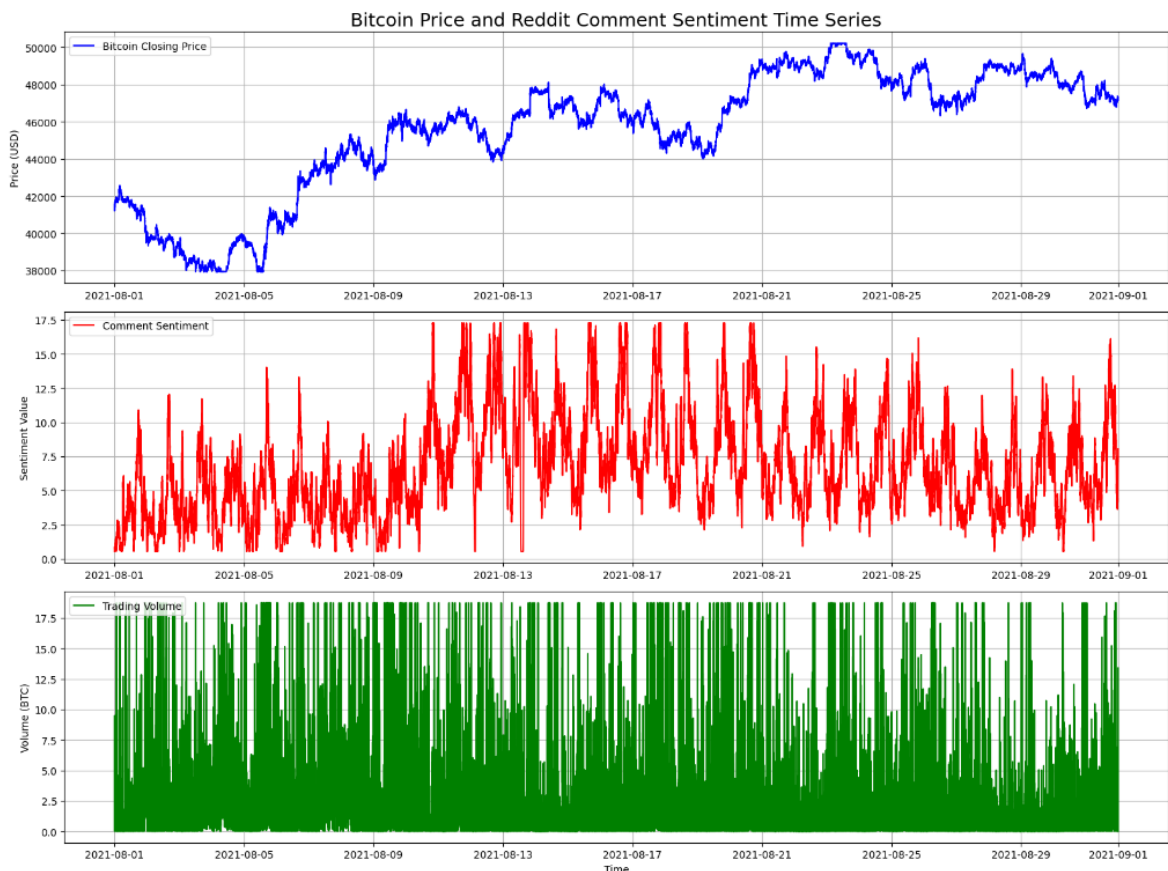


Fig. 2 Time series comparison of bitcoin price, reddit sentiment and trading volume in August 2021

As illustrated in Figure 2, the blue line reflects Bitcoin's price trajectory. Opening at approximately \$42,000 on 1 August, it experienced a brief decline (reaching a low of around \$38,000 on 5 August) before commencing an oscillating ascent from 9 August. Surpassing \$48,000 around 20th August before peaking at approximately \$50,000 on 25th August. Subsequently, it fluctuated

within the \$46,000–\$49,000 range, closing at around \$47,000 on 31st August. The price exhibited a pattern of ‘initial decline followed by recovery and subsequent high-level oscillation,’ characterised by high volatility (with daily fluctuations exceeding 5%), consistent with typical cryptocurrency market behaviour. The red line reflects trends in sentiment shifts within Reddit comments. Sentiment values clustered within the 5–17.5 range, exhibiting high-frequency fluctuations (with multiple daily spikes or dips in sentiment values) and no discernible long-term trend. However, several pronounced ‘sentiment peaks’ emerged (e.g., 12th, 15th, and 20th August, where values exceeded 15).

Short-term fluctuations in sentiment values exhibit a degree of synchronisation with price movements, indicating that social media sentiment reacts more sensitively and is driven by short-term events. The green bar chart illustrates the linkage between trading volume and sentiment/price movements. Trading volume remained generally high, partially overlapping with sentiment peaks (e.g., the 12th of August sentiment peak coincided with a modest volume increase, while the 20th August simultaneous sentiment and price surge saw significantly amplified volume). Volume peaks frequently accompany heightened sentiment and price breakthroughs.

From 9th to 13th August, the price rebounded from \$38,000 to \$46,000, with sentiment rising from 5 to 17.5 and trading volume increasing from 8 BTC to 15 BTC, demonstrating sentiment leading price gains; 15–20 August: Price fluctuated at elevated levels, with sentiment repeatedly exceeding 15 and trading volume stabilising between 12–15 BTC, forming a positive feedback loop; 25–30 August: Price retreated after peaking, while sentiment rebounded to 15, underpinning the elevated price level; 31 August - 1 September: Prices retreated to \$47,000, sentiment plunged to 5, and trading volume declined in tandem.

In summary, social media sentiment—particularly short-term peaks—exhibits synchronisation with Bitcoin's price and trading volume, thereby substantiating the notion that sentiment influences price. Heatmap of Correlation Coefficients Between Variables: An Analysis of the Relationship Between Bitcoin Closing Prices, Trading Volume and Sentiment in Comments.

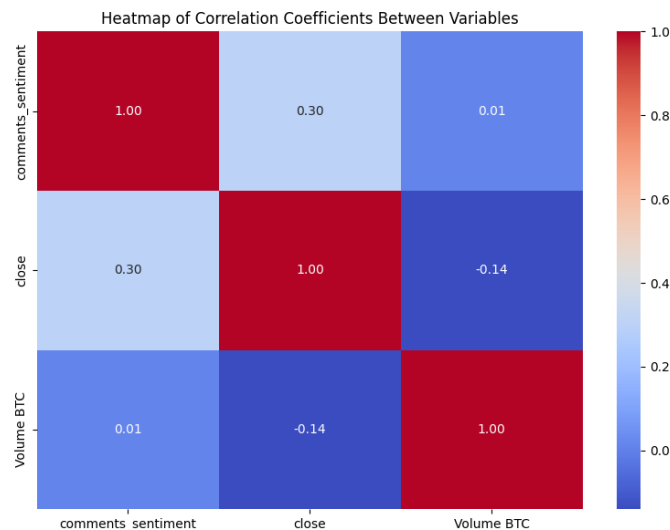


Fig. 3 Heatmap of correlation coefficients between variables: analysis of associations between bitcoin Closing price, trading volume, and sentiment in comments

Figure 3 shows that there is a weak positive correlation between sentiment and closing price (0.301701), the higher the sentiment, the slightly upward trend of the overall price. There is almost no correlation between mood and trading volume (0.009386), which shows that mood fluctuation has little influence on trading volume. There is a weak negative correlation between closing price and trading volume (-0.140010), that is, the higher the price, the lower the trading volume.

Table 1. Granger causality test – analysis of sentiment's leading effect on closing prices (different lag orders)

number of lags (no zero)1	p=0.0560
number of lags (no zero)2	p=0.0945
number of lags (no zero)3	p=0.0116
number of lags (no zero)4	p=0.0105
number of lags (no zero)5	p=0.0200
number of lags (no zero)6	p=0.0335
number of lags (no zero)7	p=0.0508
number of lags (no zero)8	p=0.0731
number of lags (no zero)9	p=0.0963
number of lags (no zero)10	p=0.0869
number of lags (no zero)11	p=0.1249
number of lags (no zero)12	p=0.1728
number of lags (no zero)13	p=0.2118
number of lags (no zero)14	p=0.1499
number of lags (no zero)15	p=0.1531
number of lags (no zero)16	p=0.0209
number of lags (no zero)17	p=0.0132
number of lags (no zero)18	p=0.0176
number of lags (no zero)19	p=0.0256
number of lags (no zero)20	p=0.0335
number of lags (no zero)21	p=0.0337
number of lags (no zero)22	p=0.0453
number of lags (no zero)23	p=0.0535
number of lags (no zero)24	p=0.0579
number of lags (no zero)25	p=0.0516
number of lags (no zero)26	p=0.0522
number of lags (no zero)27	p=0.0633
number of lags (no zero)28	p=0.0793
number of lags (no zero)29	p=0.0928
number of lags (no zero)30	p=0.1130

From Table 1, it can be observed that when $p < 0.05$ (3–6 minutes, 16–22 minutes) and when $p \geq 0.05$ (1–2 minutes, 23–60 minutes) in the minute-level data, ‘lag 1 order’ denotes a 1-minute delay. Consequently, ‘lag 2–6 orders’ corresponds to 2–6-minute delays, while ‘lag 20–21’ denotes a 20–21-minute delay. This short-term leading effect reflects the ‘sentiment influencing price’ phenomenon under high-frequency trading (2–6 minutes): it may correspond to ‘rapid reactions from short-term capital’—whereafter social media sentiment shifts, quantitative programs swiftly capture sentiment signals, completing the process from ‘interpretation to order placement to price impact’ within minutes. Consequently, sentiment's leading effect on prices is pronounced within this timeframe.

3. Model Evaluation

From Table 2, this study observes R-squared: 0.111; Adj. R-squared: 0.111. The coefficient of 0.111, measuring the proportion of variation in the dependent variable explained by the model, indicates that the model accounts for only approximately 11.1% of the variation in closing prices, demonstrating limited explanatory power.

P-value = $0.00 < 0.05$ indicates the model is statistically significant overall, meaning the independent variables collectively exert a statistically meaningful influence on closing prices. coef 268.5929 indicates that for each unit increase in sentiment, the BTC closing price rises by an average of 268.59 units (holding other variables constant). This demonstrates a positive correlation between social media sentiment and BTC price, meaning higher market sentiment correlates with higher prices. Durbin-Watson: 0.029. This statistic tests for serial correlation in residuals; a value close to 0 indicates strong positive serial correlation.

Social media sentiment exerts a significant positive influence on BTC closing prices. Positive sentiment drives price increases, aligning with the intuitive logic that market sentiment underpins asset valuations. This phenomenon may serve as a supplementary indicator for price forecasting. The model's explanatory power remains limited. Given that BTC pricing is driven by multiple factors, incorporating variables such as macroeconomic indicators, technical metrics, and policy events is necessary to enhance interpretability.

Table 2. Regression analysis of sentiment's impact on closing prices

OLS Regression Results					
R-squared:	0.111		Prob (F-statistic):	0.00	
	Coef	Std err	t	P> t	[0.025 0.975]
const	$4.401e^{+04}$	31.596	1392.796	0.000	$4.39e^{+04}$ $4.41e^{+04}$
Comments sentiment	268.5929	3.955	67.917	0.000	260.842 276.344
Volume BTC	-159.0406	4.967	-32.016	0.000	-168.777 -149.304
Omnibus	2722.619		Durbin-Watson	0.029	

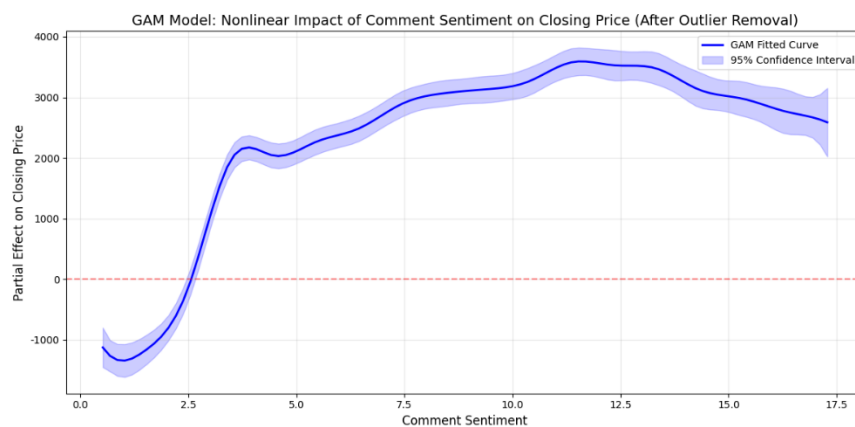


Fig. 4 Nonlinear influence plot of comment sentiment on closing price via GAM after outlier removal

This figure 4 illustrates the nonlinear influence of comment sentiment on Bitcoin's closing price through a GAM. As sentiment rises from lower levels (close to 0), its partial influence on closing prices rapidly shifts from negative (depressing prices) to positive, with the positive impact significantly amplifying as sentiment strengthens.

Once sentiment reaches higher levels, the positive influence gradually diminishes, even exhibiting a tendency to turn negative. The relationship between sentiment and closing price exhibits a complex non-linear pattern rather than a simple linear correlation: initially negative, then rapidly positive, followed by diminishing positive effects. Low sentiment may suppress prices, while rising sentiment strongly drives price increases. However, excessively high sentiment causes the positive effect to diminish and may even reverse. Confidence Interval: The 95% confidence interval (blue shaded area) remains narrow overall and clusters closely around the fitted curve. This indicates that the observed non-linear pattern is stable and reliable, rather than an illusion caused by random fluctuations.

The core objective is to compare the Bitcoin price forecasting performance of two machine learning models (random forest and LSTM) before and after incorporating sentiment variables, thereby verifying whether sentiment features can enhance predictive accuracy.

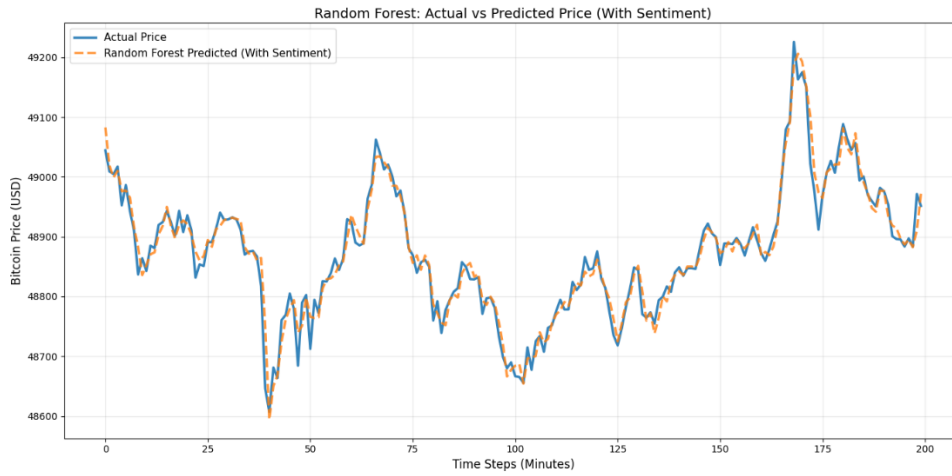


Fig. 5 Random Forest (with sentiment) (ML)

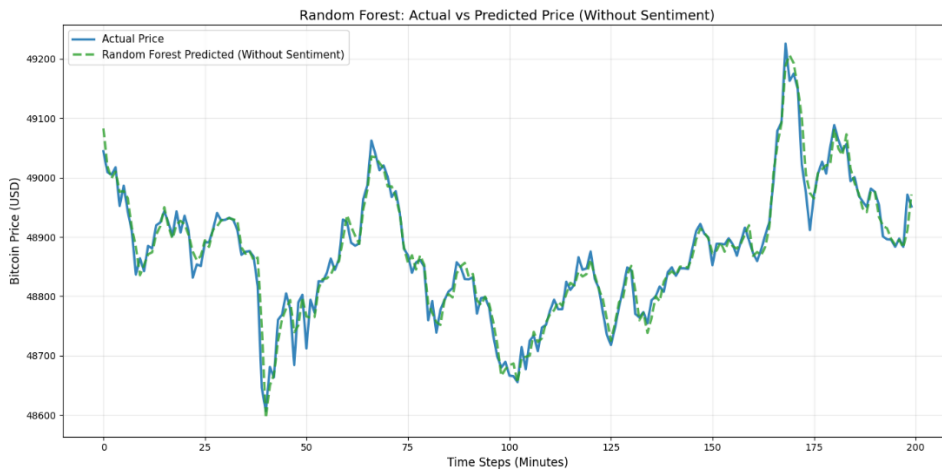


Fig. 6 Random Forest Model: comparison of actual and predicted bitcoin prices when emotional factors are not included

As shown in Figure 5 and Figure 6, inclusion of sentiment features yields no significant gain for random forests. For random forests, incorporating sentimental features has negligible impact on predictive accuracy. The mean squared error (MSE) with sentiment features is 334.86, while the MSE without sentiment features is 334.02. Both metrics and their corresponding R^2 values show virtually no difference. This indicates that sentimental features do not provide effective incremental information for random forests and may even slightly reduce predictive accuracy due to inherent noise.

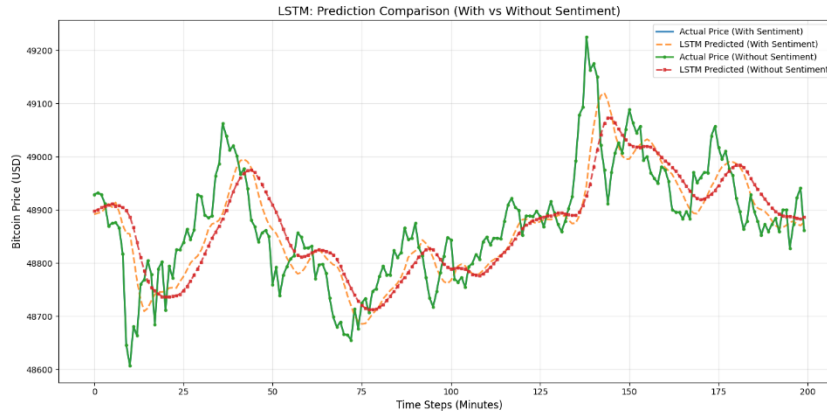


Fig. 7 LSTM Model: comparison of bitcoin price forecasts with and without emotional factors included

As illustrated in Figure 7, the LSTM model demonstrates significantly enhanced predictive performance when incorporating sentiment features.

Compared to models without sentiment features, the inclusion of sentiment features markedly improves the LSTM's predictive accuracy: the MSE decreases from 6760 to 4505, while the R^2 value increases from 0.9889 to 0.9926. This indicates that sentiment features provide the LSTM with crucial incremental information, enabling the model to capture price fluctuation patterns with greater precision. The prediction curve incorporating sentiment features (orange dashed line) closely tracks actual prices (blue solid line). Conversely, the curve without sentiment features (red dashed line) exhibits noticeable lag or deviation during price fluctuations (e.g., the peak at 125 minutes), visually demonstrating how sentiment features correct LSTM's prediction biases for volatility.

4. Conclusion

Sentiment exhibits a weak positive linear correlation with price movements, and sentiment constitutes a unidirectional Granger cause of price fluctuations. Sentiment is not the sole factor influencing prices; other variables such as macroeconomic policies and technical indicators may attenuate the strength of this linear relationship. The relationship between sentiment and price demonstrates non-linear characteristics. GAM proves more suitable for capturing complex correlations, enabling the identification of sentiment's differentiated impact on prices across distinct intervals.

GAM analysis indicates that sentiment's impact on prices follows a tripartite non-linear pattern rather than a single linear trend: 'rapid ascent in low sentiment intervals - moderate fluctuation in medium sentiment intervals - decline in high sentiment intervals. This finding addresses the limitations of existing research focused solely on linear relationships. The enhancement of price prediction accuracy by sentimental features varies across models. LSTM demonstrates greater sensitivity to sentimental features, enabling more timely capture of price inflection points when sentiment is incorporated.

Comparative experiments between Random Forest (traditional machine learning) and LSTM (deep learning) reveal significant differences in sensitivity to sentiment features between the two model types. For the Random Forest model, incorporating sentiment variables only marginally increased the MSE from 334.02 to 334.86, while R^2 remained between 0.9994 and 0.9995. The marginal value of sentiment features was 'overwhelmed' by the model's robust nonlinear fitting capability, yielding no significant accuracy improvement. For the LSTM model, incorporating sentiment variables reduced the MSE from 6760.77 to 4505.87, with R^2 rising from 0.9889 to 0.9926. Moreover, the prediction curve aligns more closely with actual prices at inflection points (such as rapid price fluctuations), indicating that sentiment provides LSTM with crucial temporal information, effectively enhancing its ability to capture short-term volatility. Model Suitability Analysis: Random Forest outperformed LSTM on this dataset (MSE ~334 vs 4500+), primarily because minute-level

high-frequency data exhibits short-term temporal dependencies. LSTM's strength in extracting long-sequence features was less effective here, whereas Random Forest's suitability for weak temporal dependencies and strong capture of non-linear relationships better suited this research scenario.

This study, whilst revealing the mechanism by which sentiment influences Bitcoin's short-term price movements, remains subject to two limitations: firstly, the data sample covers only the period up to August 2021, excluding times of extreme market events (such as tightening regulatory policies), necessitating further validation of the conclusions' timeliness and universality; secondly, sentiment data is sourced solely from the Reddit platform, excluding other mainstream social media platforms such as Twitter and Facebook, potentially introducing bias in sentiment measurement. Future research may expand to multi-platform, long-term data while incorporating macroeconomic and technical indicators to construct multi-factor predictive models, thereby enhancing explanatory power and forecasting accuracy for price fluctuations.

In summary, this study not only enriches the literature on behavioural finance within cryptocurrency markets by confirming the role of social media sentiment in short-term price formation but also provides empirical support for investors to optimise high-frequency trading strategies (such as sentiment-driven timing decisions), holding significant theoretical and practical value.

References

- [1] Farrell R. An analysis of the cryptocurrency industry. University of Pennsylvania Doctoral Dissertation, 2015.
- [2] Ciaian P, Rajcaniova M, Kancs D A. The economics of BitCoin price formation. *Applied Economics*, 2016, 48(19): 1799-1815.
- [3] Huang Q. Bitcoin price prediction based on fear & greed index. *SHS Web of Conferences*, 2024, 181: 02015.
- [4] Baig A, Blau B M, Sabah N. Price clustering and sentiment in bitcoin. *Finance Research Letters*, 2019, 29: 111-116.
- [5] López-Cabarcos M Á, Pérez-Pico A M, Piñeiro-Chousa J, Šević A. Bitcoin volatility, stock market and investor sentiment. Are they connected?. *Finance Research Letters*, 2021, 38: 101399.
- [6] Naeem M A, Mbarki I, Shahzad S J H. Predictive role of online investor sentiment for cryptocurrency market: Evidence from happiness and fears. *International Review of Economics & Finance*, 2021, 73: 496-514.
- [7] Dixon M F, Halperin I. The four horsemen of machine learning in finance. *SSRN Working Paper*, 2019.
- [8] Welch I, Goyal A. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 2008, 21(4): 1455-1508.
- [9] Li Xingyi, Li Zhongfei, Li Qiqian, Liu Yujun, Tang Wenjin. Overview of the research on asset return prediction based on machine learning. *China Management Science*, 2025, 33(1): 311-322.
- [10] Kristoufek L. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 2013, 3(1): 3415.