Can Reddit Posts Forecast Bitcoin Return?

Minwoo Song

CUNY, The Graduate Center

May 17, 2022

Abstract

I quantitatively measure sentiment from the Reddit forum on Bitcoin (r/Bitcoin) and study its short-term relationship with the return of Bitcoin, controlling for other potential explanatory variables. I find that the significance of regression models for the Bitcoin returns have been changing over time and it show that increased positive postings about Bitcoin on the Reddit community are associated with an increase in the price of Bitcoin. Also, incorporating the sentiment improves the predictive ability for some time periods.

Keywords: Bitcoin, prediction, sentiment analysis, textual analysis.

JEL Codes: C32(Multiple or Simultaneous Equation Models: Time-Series Models), C58(Financial Econometrics), G17(Financial Forecasting and Simulation)

1. Introduction

Investors have an interest in predicting the return of Bitcoin because it is regarded as a decentralized digital currency, and it might be considered an investment. When Bitcoin stands out as a speculative investment, rather than a currency, as Yermack (2015) concludes, investors are interested to predict its future price for capital gain. When Bitcoin is instead used as a currency,[1] its future price is relevant to gauge whether it is beneficial to use Bitcoin as payment instead of using other currencies.

Theoretical papers such as Bolt and Van Oordt (2020) show that illegal usage such as black market and gambling, and speculative motives are biggest driver of demand of Bitcoin. On the other hand, empirical papers (Chen at al 2021; Garcia et al 2014; Mai et al 2015) investigate factors that explain the price of Bitcoin, especially using public attention measures such as Google Trends, Wikipedia views (the number of visits on Bitcoin's page on Wikipedia), Twitter posts, and a Bitcoin discussion forum (Bitcointalk.org). They observe that search volume is positively associated with the price of Bitcoin. I find that the relationship, however, is not consistent over time, and sometimes it becomes negative. When there were few Bitcoin users, interested buyers would search Bitcoin through Google and Twitter before they purchased it. Therefore, previous papers (based on the data from 2008 to 2015) show a positive relationship. However, the relationship becomes negative starting in 2018. One possible explanation is that people who hold Bitcoin do web query Bitcoin when there are drastic price changes, and in the presence of negative news, then they might sell Bitcoin. While Kaminski (2014) uses sentiment analysis of Twitter posts by sorting tweets into three categories based on their emotional content: positive, negative or uncertain. Although a dynamic Granger causality analysis does not show a causal effect of emotional tweets on Bitcoin prices, there is a significant positive correlation between emotional tweets and the close price, trading volume, and intraday price spread of Bitcoin.

The aim of this paper is to investigate if sentiment from Reddit posts is a potential determinant of the Bitcoin returns. In the analysis, I use several controllers. First, I consider the lagged returns of Bitcoin and other cryptocurrencies whether there exists time-series momemtum in Bitcoin. Second, I also consider the returns of the top four cryptocurrencies in the market[2] (Bitcoin, Ethereum, Litecoin, and Bitcoin cash) since investors of Bitcoin tend to trade other cryptocurrecies. Third, I adopt the number of active Bitcoin addresses as a proxy for the network effect to capture user adoption. Lastly, I consider Google Trends and the sentiment scores from a Reddit community of Bitcoin to examine the role of investors' attention about Bitcoin.

Google Trends analyzes the popularity of search queries in Google, and I obtaine a standardized numerical

[1] Note that PayPal announced that cryptocurrencies including Bitcoin are supported as a form of payment (March 30, 2021).

[2] It is based on market capitalization as the time of April 2021.

measure of how many times people look up Bitcoin in a given period. This variable has been widely used to proxy for investor attention (Chen et al. 2021; Kristoufek 2013; Liu and Tsyvinski 2021). I collect the Reddit posts from the Bitcoin community (https://www.reddit.com/r/Bitcoin/) and construct sentiment scores of them by using natural language processing (NLP). There are several studies that adopt sentiment analysis of Twitter, Facebook, and newspapers, but no one has yet considered Reddit posts. In light of the "GME short squeeze" event[3], we regard the Reddit community as having enough power to ignite changes in the prices of stocks or cryptocurrencies. One of the distinguishing features of the Reddit community is that posts are created and responded to by the people ("redditors") who are really paying attention to Bitcoin, and this implies that the effect of sentiment would have a more considerable impact on Bitcoin than other channels, such as Twitter. Therefore, it is worthwhile to utilize the Reddit posts for the returns of Bitcoin, as it would capture the shifts in demand for Bitcoin.

I find that using the Reddit posts contributes to explaining the movement of return of Bitcoin. First, I find that if more people posts positive on the Bitcoin community, then it would increase the demand of Bitcoin and raise the future price. Second, my results indicate that increased positive postings about Bitcoin on the Reddit community are associated with an increase in the price of Bitcoin. Lastly, incorporating the Reddit sentiment moderately improves the out-of-sample forecast of the Bitcoin returns.

The paper is structured as follows. Section 2 describes the data, including the sentiment analysis for the posts on the Reddit community. The in-sample and out-of-sample forecasting exercises are carried out in Sections 3 and 4, respectively. Section 5 summarizes and concludes.

2. Data

The cryptocurrency price data series are obtained from TradingView (https://www.tradingview.com/) I selected the top four cryptocurrencies as of April 2021 – Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), and Bitcoin cash (BCH) – based on the value of transactions, the price times its transaction volume, on average. Unit prices are measured in US dollars, and the return is measured in percentage return. Table 1 displays the descriptive statistics of cryptocurrencies' prices and returns at daily frequency from 11/22/2018 to 3/1/2021.

[3]In January 2021, a short squeeze resulted in a 1,500% increase in GameStop's share price over the course of two weeks, reaching an all-time intraday high of US$483.00 as of January 29, 2021, on the New York Stock Exchange. This effect was mainly attributed to a coordinated effort by the Reddit community r/wallstreetbets, a subreddit dedicated to stocks with high market risk.

Descriptive statistics of the returns of cryptocurrencies

	Return of BTC	Return of ETH	Return of LTC	Return of BCH
Min	-0.3881	-0.4332	–0.3801	-0.4505
Max	0.1949	0.2646	0.2965	0.5154
Mean	0.0036	0.0042	0.0033	0.0026
Median	0.0021	0.0015	0.0004	-0.0004
SD	0.0405	0.0511	0.0539	0.0622

<Table 1. Descriptive statistics of cryptocurrencies>

I use the number of active addresses of Bitcoin (BTCAAD) to capture the demand of Bitcoin for a form of payment or transferring money. It is the sum count of unique addresses that were active in the network interval (either as a recipient or originator of a ledger change) which is offered by Coinmetrics.io. To put it simply, the higher the active number of addresses, the more people use Bitcoin as a form of payment or to transfer money.
I also consider the S&P500 index to explain the Bitcoin since the cryptocurrency market is compared with the stock market in the literature frequently. Note that I handle missing data (holidays and weekends) by replacing it with previous values to put the S&P500 index at a daily frequency.

Google Trends represents the search volume of Bitcoin on Google and Youtube globally; I use it to proxy for investor attention. I normalize the value of search queries for Bitcoin for the period from 11/22/2018 to 3/1/2021[4]. Lastly, I use the Reddit posts from the Bitcoin forum (https://www.reddit.com/r/Bitcoin/), which is a novel dataset in this paper. From the GME short squeeze situation[5], we notice that a movement of the price of stocks or cryptocurrencies can be captured by the posts of a Reddit community. As of 4/1/2021, there are 2.8 million subscribers of the Bitcoin community. It is worthwhile to investigate the Reddit posts to model Bitcoin returns since they might capture the shifts of demand for the cryptocurrency. Consequently, it is possible that the Reddit posts would affect the price of Bitcoin. In order to collect the posts for more than two years at the daily frequency, I use the Python packages, `Pushshift` and `PSAW`, and collect 179,234 posts for 1,156 days (1/1/2018 to 3/1/2021). On average, there have been 155 posts per day. The main parts of posts are title, content, karma score, the number of comments, and time. I do not use filters such as the number of comments and karma score to construct the daily frequency scores.

[4]The resulting numbers are then scaled on a range of 0 to 1 based on a topic's proportion to all searches on all topics.

[5]The short squeeze was initially and primarily triggered by users of the internet forum on Reddit, r/wallstreetbets. See Abhinav and Pathak (2022)

To utilize the Reddit posts and use them as an explanatory variable, it is necessaty to transform the text into numerical measure. First, I collect titles of Reddit posts instead of contents to extract valuable information, as most of the posts do not have valid text. Second, I use a Sentiment Intensity Analyzer (SIA) of the Natural Language Toolkit (NLTK) to obtain a polarity score for the sentiment strength based on the input text. Table 2 shows examples of titles and their scores, and Table 3 and Figure 1 show descriptive statistics of the scores of Reddit posts and the distribution of them, respectively. Lastly, Figure 2 shows the time series datasets.

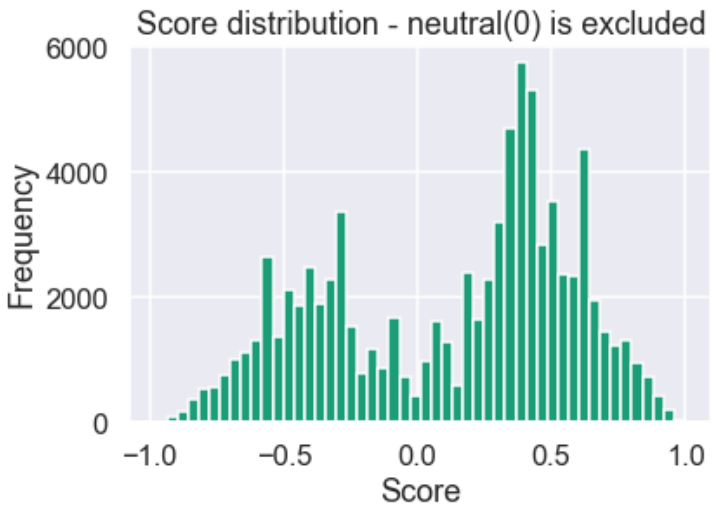
Time	Title of Post	Score
2018-01-03 20:39:50	*"**Devil**'s advocate: how long can bitcoin continue to **grow/appreciate** in **value** if it's continuously expensive to use? (transaction fees)"* devil =–3.6, grow = 0.7, appreciate = 1.7, value = 1.4, negative: 0.0, neutral: 0.882, positive: 0.118, compound: 0.34	0.34
2018-01-03 22:31:22	*"Egypt's **top** imam has declared that bitcoin is **forbidden** under Islam"* top = 0.8, forbidden =–1.8, negative: 0.206, neutral: 0.662, positive: 0.132, compound: -0.25	-0.25
2018-01-04 01:11:09	*"There are currently 1,000 more people **actively** viewing this sub than there are **active** Bitcoin Nodes globally"* actively = 1.3, active = 1.7, negative: 0.0, neutral: 0.74, positive: 0.26, compound: 0.6461	0.65
2018-01-04 03:03:09	*"It's only a **matter** of time before China lifts crypto exchange **ban**, entrepreneur says"* matter= 0.1, ban =–2.6 negative: 0.216, neutral: 0.719, positive: 0.066, compound: -0.5423	-0.54

The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized ($z = x/\sqrt{x^2+15}$) to be between -1 (most extreme negative) and +1 (most extreme positive).
The pos, neu, and neg scores are ratios for proportions of text that fall in each category.

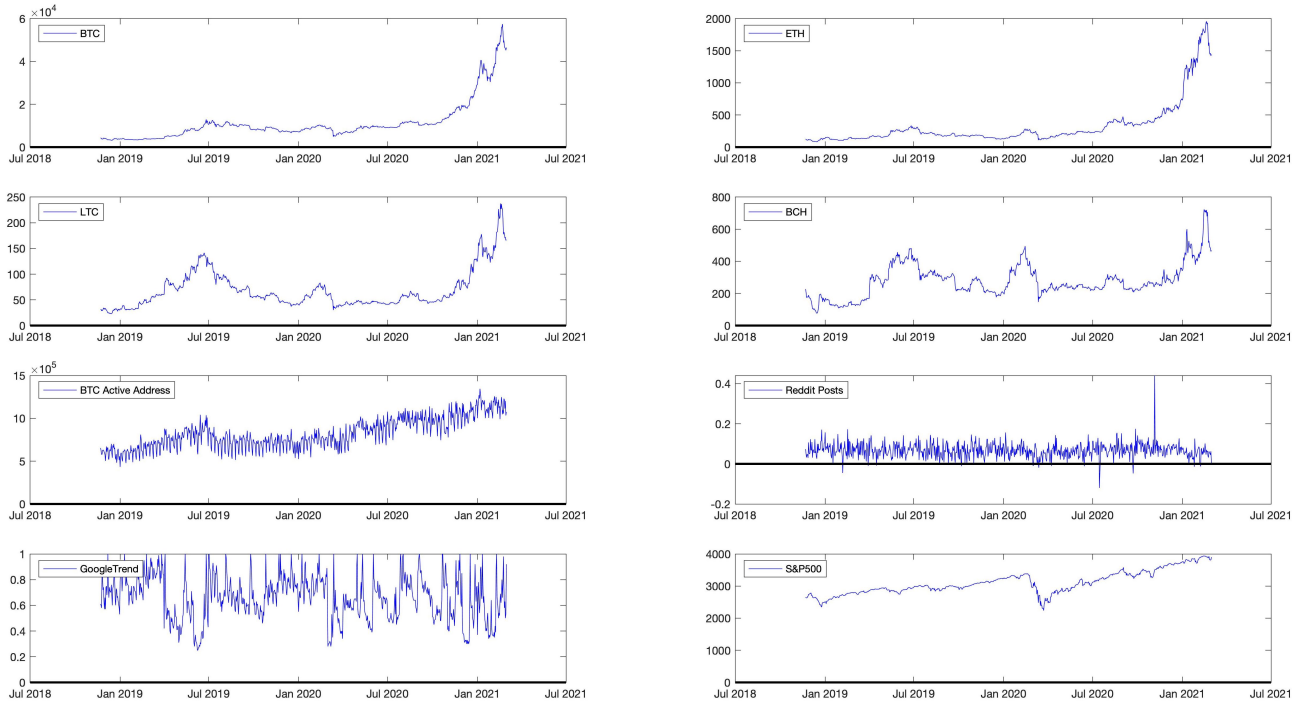
<Table 2. Sample posts' title and their score>

	Min	Max	Mean	Median	Mode	SD
Scores of Reddit Post	-0.9752	0.9969	0.062	0	0	0.319

<Table 3. Descriptive statistics of scores of Reddit posts>



<Figure 1. Distribution of scores of the Reddit posts of Bitcoin community>



<Figure 2. Historical data for the prices of cryptocurrencies, active address of BTC, scores of Reddit posts, Google Trends, and S&P500>

3. In-Sample Analysis

3.1 Linear Probability Model

First, I investigate the directional predictability for the return of Bitcoin by converting it to a binary variable, Y_t, (positive return=1, zero or negative return=0) [6]. Given that the noise in the observed returns of Bitcoin is too high comparing with other assets, checking the directional probability with a linear probabilty model is better to find consistent and reliable predictors. By considering a set of potential explanatory variables with a simple binary regression model, we can examine the predictive ability of certain variables.

Linear Probability Model :

$$P(Y_t = 1 \mid X_{t-1}) = \beta_0 + \sum_{j=1}^{J} \beta_j R_{j,t-1}, \tag{1}$$

where $R_{j,t-1}$ is lagged log return of variable j, and $j = \{BTC, ETH, LTC, BCH, SP500, BTCAAD, RedditScore, Google\}$.

Table 4 reports the estimation result of model (1) with the binary response variable, the return of Bitcoin. First of all, for the full sample, there is no time-series momentum effect for the return of Bitcoin since the lagged return of Bitcoin is not statistically significant. On the other hand, two variables are statistically significant at the 1% level: the first lag of return of Ethereum (-1.83) and the lagged growth rate of the score of Reddit post (0.20). It implies that Bitcoin investors consider Ethereum as a substitute – if today's return of Ethereum is negative, people would sell it and buy Bitcoin, leading to an increase in the price of Bitcoin. Positive and significant coefficient of the lagged growth rate of the score of Reddit post tell us, if Reddit users are optimistic about the cryptocurrency, then it would increase the demand of Bitcoin and raises its price.

[6] Note that there is no zero return for the period.

	Full Sample	Pre-COVID-19	COVID-19
	11/22/2018 -3/1/2021	11/22/2018 - 2/29/2020	3/1/2020 -3/1/2021
BTC	0.75 (0.87)	-0.31 (1.19)	2.05 (1.06)*
ETH	-1.83 (0.68)***	-1.71 (1.10)	-1.81 (0.88)**
LTC	0.54 (0.65)	0.64 (0.89)	0.71 (0.99)
BCH	-0.42 (0.47)	0.13 (0.60)	-1.63 (0.71)**
S&P 500	-0.01 (0.01)	0.02 (0.03)	-0.03 (0.01)*
BTC Active Address	0.00 (0.14)	-0.07 (0.19)	0.05 (0.22)
Score of Reddit Post	0.20 (0.06)***	0.12 (0.05)**	0.78 (0.30)***
Google Trends	-0.03 (0.10)	0.11(0.16)	-0.18 (0.13)

Note that explanatory variables are the first lagged log return. Standard errors are shown in parantheses. *, **, and *** denote significance levels at the 10%, 5%, and 1% levels based on the robust standard t-statistics. The data frequency is daily.

<Table 4. LPM estimation>

To evaluate the robustness of the results, I divide the periods into before and after COVID-19 impacts – 11/22/2018 to 2/29/2020 and 3/1/2020 to 3/1/2021. The Pre-COVID-19 column reports the estimated result of the model (1) for the period before the COVID-19 impacts (11/22 2018 to 2/29/2020). Interestingly, only the lagged growth rate of the Reddit score (0.12) is statistically significant under $\alpha = 5\%$ while other variables are not statistically significant, and even signs are not consistent.

For the period of the COVID-19 impacts (3/1/2020 to 3/1/2021), however, it shows a different result with pre-COVID-19, as the last column presents. First, there is time-series momentum in Bitcoin returns, and the current returns predict future returns one day ahead. Second, estimates of Ethereum and Bitcoin cash (BCH) are negative and statistically significant, which means these two cryptocurrencies are regarded as alternatives to Bitcoin[7]. That being said, if the stock market crashed today, investors would adjust their portfolios, selling their stocks and buying Bitcoin instead, and is supported by the estimation result (-0.03). Lastly, the score of Reddit posts plays a more important role during the wave of the COVID-19 crisis. In fact, the number subscribers of the Bitcoin community on Reddit has been drastically increased through the COVID-19 period[8], and it implies that the Reddit posts would affect the demand of Bitcoin for investors more than before.

[7] According to a survey from Cardify (https://www.cardify.ai/reports/crypto-revisited), there has been a shift from traditional investments to cryptocurrencies.

[8] The subscriber of r/Bitcoin was 1.2 million at the time of Dec 2019, and 2.6 million at the time of March 2021. Source. https://subredditstats.com/r/Bitcoin

Overall, the estimation results of a linear probability model show that the sentiment score of Reddit posts is useful to predict the Bitcoin returns regardless of the market situation. Halaburda et al. (2020) point out that the extreme volatility and price increase of Bitcoin is a sign of a bubble and its price does not reflect the fundamentals – blockchain technology, permisionless access, and decentralized database management–, and many empirical papers examine the demand of Bitcoin to explain Bitcoin price. Therefore, the sentiment score of Reddit posts is a good predictor to capture the shift of demand for Bitcoin. Increases in demand of Bitcoin could be caused by (i) an increased number of users, (ii) people expect the price to increase, (iii) people's preferences (e.g., more people accept cryptocurrency as a currency or a digital asset), (iv) increase in income, and (v) changes in the prices of related currencies and assets. The sentiment score of Reddit post reflects the shifts of demand for the cases of (i), (ii), and (iii), while lagged returns of other cryptocurrencies and S&P 500 index capture (v).

3.2 Rolling window regression

In addition to a binary regression model, we estimate a rolling window regression to see how the coefficients of the linear regression model and its significance change over time for the continuous Bitcoin returns.

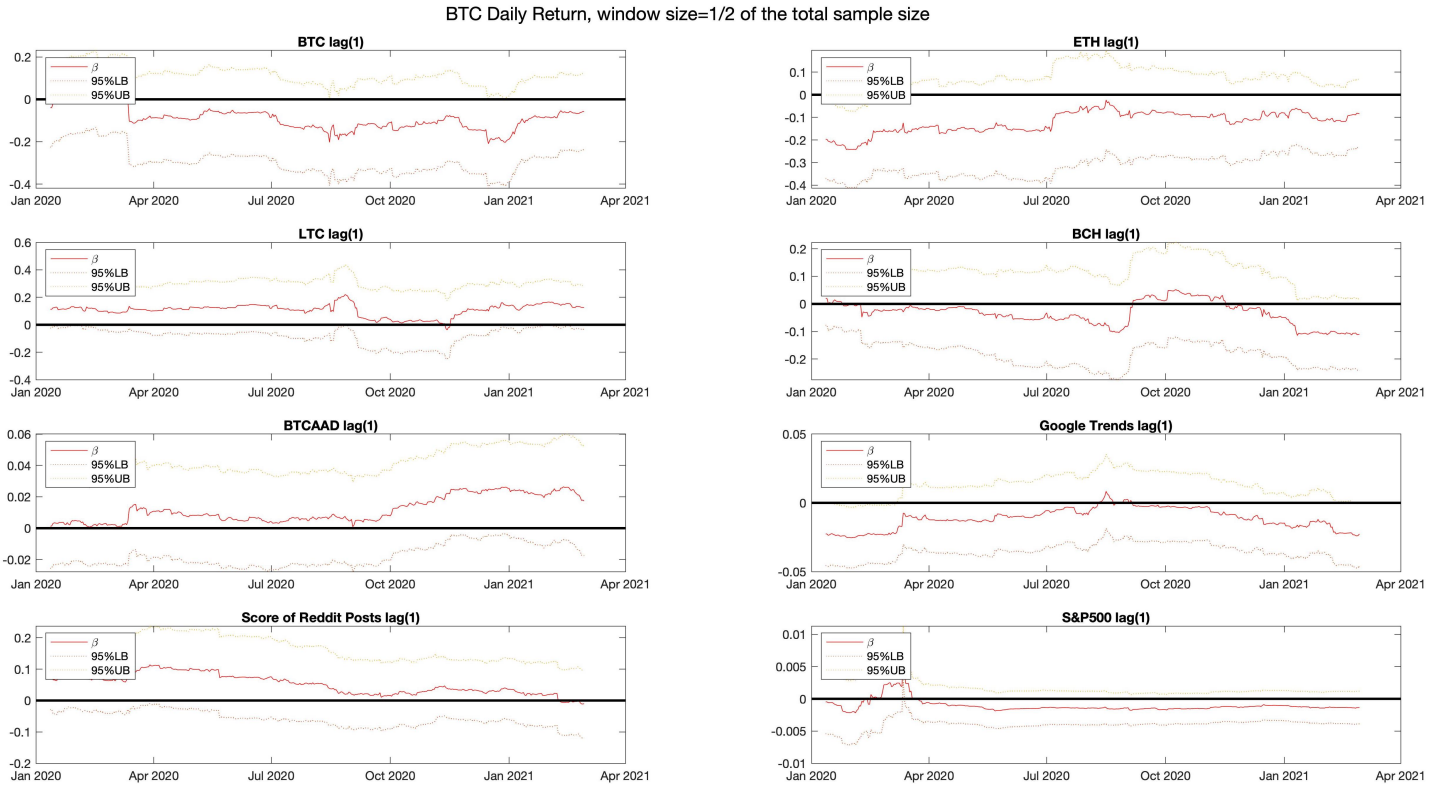
Regression Model :

$$R_{BTC,t} = \alpha_t + \beta_{X,t-1} X_{t-1} + e_t, \tag{2}$$

where $X_{t-1} = [R_{BTC,t-1}, R_{ETH,t-1}, R_{LTC,t-1}, R_{BCH,t-1}, R_{SP500,t-1}, R_{BTCAAD,t-1}, R_{RedditScore,t-1}, R_{Google,t-1}]$, $e_t \sim WN(0,1)$

We estimate the parameters of the model (2) with rolling in-sample windows (window size = 1/2 of the total sample size)[9], and Figure 3 reports the result of it.

[9]BTC daily return, rolling window regression(window size=1/2 of the total sample size), 11/22/2018-03/01/2021(T=831).



<Figure 3. Rolling window regression with 95% confidence intervals>

First of all, there is no variable that is statistically significant for the windows, and even the signs of coefficients for some variables have changed:

1. The lagged Bitcoin returns are positive before the COVID-19 impacts, then it becomes negative (today's return is positive, then tomorrow's return would be negative), and only a few periods are statistically significant.

2. The sign of the coefficients for the lagged return of Ethereum is consistent as negative, in more instances than it is significant than the lagged Bitcoin returns. Even though it is not always statistically significant, it still shows a consistent negative sign. This implies that investors consider Ethereum as an alternative cryptocurrency of Bitcoin, especially as a speculative asset.

3. In the same way as Ethereum, the score of the Reddit post shows consistent positive coefficients, and it was statistically significant for a while, from Jan 2020 to May 2020. When there are more positive Reddit posts on the Bitcoin forum, we would expect positive returns in the following day.

4. The coefficient of lagged growth rate of Google Trends is instead negative than positive and especially for the periods when it is statistically significant. This might happen if people notice some negative news about Bitcoin and they look up "Bitcoin" in Google Search to check what is happening to the Bitcoin, and then they sell it.

3.3 Impulse response function(IRF) analysis

To examine how much the Reddit posts affect Bitcoin returns, I estimate an impulse response function based on the local projection method proposed by Jordà (2005).

First, consider a reduced form VAR model

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \qquad t = 1, \ldots, T, \tag{3}$$

where A_i are $(k \times k)$ matrices of coefficients, $u_t = (u_{1t,}, \ldots, u_{kt})'$ is a K-dimensional innovation process such that $E(u_t) = 0$, and $E(u_t u_t') = \Sigma_u$.

From a reduced form VAR model, we can recover a structural VAR model with an identification. Then, we have a structural VAR model

$$B_0 y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + \omega_t, \qquad t = 1, \ldots, T \tag{4}$$

,where ω_t is assumed that K-dim white noise such that

$E(\omega_t) = 0$, $E(\omega_t \omega_t') = \Sigma_\omega$, and $E(\omega_t \omega_t') = 0$ for $s \neq t$.

Following Jordà (2005), I project y_{t+h} onto the linear space generated by $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})'$, specifically

$$y_{t+h} = \alpha^h + B_1^h y_{t-1} + \cdots + B_p^h y_{t-p} + u_t^h, \qquad h = 0, 1, \ldots, H-1 \tag{5}$$

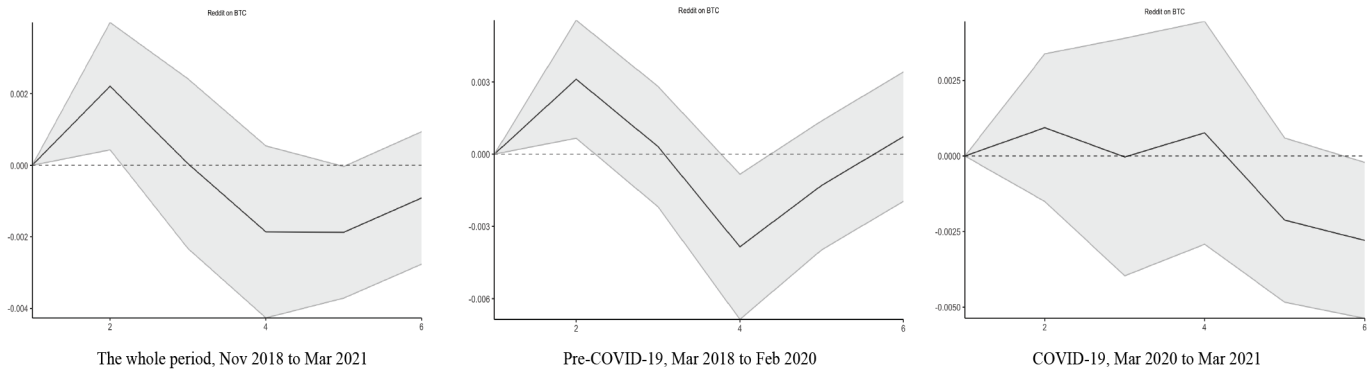
,where α^h is a vector of constants, and B_i^h are parameter matrices for lag p and forecast horizon h.

The slope matrix B_1^h can be interpreted as the respone of y_{t+h} to a reduced form shock in t, and structural responses are then estimated by following:

$$\hat{IR}(t, h, d_i) = \hat{B}_1^h d_i, \tag{6}$$

where $d_i = B_0^{-1}$.

The lag-length, $p = 3$, is chosen to make the estimation parsimonious. Figure 4 displays the impulse responses of Bitcoin due to a shock of the Reddit score for the whole period, pre-COVID-19, and COVID-19, respectively. The first graph shows that an increase of positive postings about Bitcoin on the Reddit community (an unanticipated shock) is associated with an increase in the Bitcoin returns up to two days ahead for the whole period.



The whole period, Nov 2018 to Mar 2021 | Pre-COVID-19, Mar 2018 to Feb 2020 | COVID-19, Mar 2020 to Mar 2021

<Figure 4. Impulse responses to a shock in Bitcoin>

Notes: Light grey areas correspond to the 90% confidence intervals.

The second graph displays the impulse responses of Bitcoin for the subsample, pre-COVID-19. Similarly to the full sample, the score of Reddit is strictly positive for the first two days. During the pandemic, however, it shows different impulse responses. A Reddit score shock cannot contemporaneously impact the Bitcoin returns. Since $\beta_{Reddit\ Score,t}$ of rolling window regression for the period of COVID-19 is close to zero and not significant a Reddit score shock would not be associated with the Bitcoin returns.

4. Out-of-Sample Analysis

4.1 Binary Prediction with Classification Models

In the previous section, we found that the lagged Reddit score is useful to predict the return sign of Bitcoin. The goal of this section is to evaluate the out-of-sample power of the Reddit score to forecast the Bitcoin return sign. I adopt

a hold-out validation method to evaluate the forecasting performance, and use 50% of the data for training and the remaining 50% of the data for testing. The forecast period (i.e., testing data set) is 1/11/2020 to 3/1/2021.

First, I compare the accuracy of classifications models (see Appendix for the descriptions of models) with and without the lagged return of Reddit score. As Table 5 shows, models that include the Reddit score as a predictor outperform models without it on average (18 models perform better with the Reddit score, and 7 models perform worse), but none of them are statistically significant under $\alpha = 10\%$ of the likelihood ratio test. For instance, the likelihood ratio test shows that the p-value of the null hypothesis,

H_0: Accuracy of Boosted trees model with the Reddit scores = Accuracy of Boosted trees model without the Reddit scores,

is 0.124[10]. However, the improvement 5.3 percentage points is non-negligible in finance.

[10]The z-test statistic of the likelihood test is $\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{2p_c(1-p_c)/n}}$, where $p_c = \frac{\hat{p}_1 + \hat{p}_2}{2}$

Model	Accuracy (%)		Model	Accuracy (%)	
	w/ Reddit Score	w/o Reddit Score		w/ Reddit Score	w/o Reddit Score
Fine Tree	52.0	51.1	Fine KNN	50.2	53.8
Medium Tree	52.1	51.1	Medium KNN	50.7	52.9
Coarse Tree	54.2	54.2	Coarse KNN	51.6	52.0
Linear Discriminant	52.9	51.6	Cosine KNN	56.9	52.9
Quadratic Discriminant	56.4	52.9	Cubic KNN	50.7	52.9
Logistic Regression	52.9	51.1	Weighted KNN	55.1	53.3
Gaussian Naive Bayes	49.3	50.7	Boosted Trees	58.7	53.3
Kernel Naive Bayes	53.3	52.9	Bagged Trees	54.2	54.7
Linear SVM	53.8	53.3	Subspace Discriminant	52.5	54.2
Quadratic SVM	57.8	56.4	Subspace KNN	53.3	52.4
Cubic SVM	56.9	53.3	RUS Boosted Trees	52.0	49.8
Fine Gaussian SVM	54.2	52.0			
Medium Gaussian SVM	56.9	56.9			
Coarse Gaussian SVM	52.0	52.0			

Classification models for binary response variable, BTC Return (0-negative return and 1-positive return), with hold-out validation with 50% held out.
Training data set: 11/22/2018~1/10/2021, and testing data set: 1/11/2021-3/1/2022.
Features: Lagged return of BTC, ETH, LTC, and BCH. Lagged Active address of BTC and ETH, Lagged return of S&P500, Lagged GoogleTrends, and Lagged Sentiment score of Reddit.

<Table 5. Forecasting accuracy of classification models with and without the Reddit scores>

4.2 Autoregressive Distributed Lag (ADL) Model

In this section, I evaluate the predictive ability of the scores of Reddit post using the ADL(p, q) model:

$$R_{BTC,t} = c + A_1 R_{BTC,t-1} + \cdots + A_p R_{BTC,t-P} + B_1 R_{Reddit,t-1} + \cdots + B_q R_{Reddit,t-q} + u_t, \qquad t = 1, \ldots, T, \quad (7)$$

where $R_{BTC,t}$ and $R_{Reddit,t-1}$ are the first difference of the logarithm of respectively the price of Bitcoin and the scores of Reddit posts; $T$is the total sample size, and $u_t \sim WN(0,1)$.

By comparing model (7) with the restricted model that excludes the scores of Reddit post, we can see whether the score of Reddit post improves the predictive ability.

The forecasting exercise is performed in pseudo real-time, using a rolling window regression, and projecting the models forward up to 5 steps ahead ($h = 1$ to $h = 5$). The window size is 1/2 of the total sample size.

I evaluate the results in terms of Mean Squared Forecast Error (MSFE) generated by the model (5). Defining $\widehat{R}_{BTC,t+h|t}$ as the h-step-ahead forecast of $R_{BTC,t+h}$, given the information available at time t, the h-step-ahead forecast error at time t is :

$FE_{h,t} = R_{BTC,t+h}, -\Delta\widehat{R}_{BTC,t+h|t}$,

and the h-step-ahead MSFE is calculated as:

$MSFE_h = \frac{1}{T_0}\sum_{t=1}^{T_0}(FE_{h,t})^2$, where T_0 is the total number of computed forecasts.

To facilitate the comparison, I provide results in terms of the Root Mean Squared Forecast Error (RMSFE) against the AR (p) model with lagged the Bitcoin returns[11]:

$RMSFE_h = \frac{MSFE_h^{ADL}}{MSFE_h^{AR}}$.

The results of the forecasting exercise are summarized in Table 6. The first three columns show $RMSFE_h$ with recursive (an expanding window) forecasts method with different time periods (whole data period, pre-COVID-19, and COVID-19)[12], and columns (4) to (6) show rolling (a moving window) forecast methods. I set $p = 2$, and $q = 2$, based on AIC and BIC.

First of all, a ADL(2,2) model is slightly better than the AR(2) model on average, and statistically significantly outperforms to forecast 2-steps ahead for the period of before COVID-19 regardless of the estimation scheme. Note that the recursive forecasts method yields a better performance than a rolling forecast method on average. In general, a recursive scheme is better than a rolling scheme in the absence of structural breaks. Through COVID-19, there were multiple breaks, however, a recursive scheme is still better than a rolling scheme for forecasting Bitcoin returns.

[11]Note that model (5) statistically significantly outperforms the random walk (with/without drift) under $\alpha = 1\%$ of Diebold-Mariano test statistic. Since Bitcoin returns are highly volatile the random walk models perform very poorly. As an extreme, Bitcoin returns on 3/12/2020 is -38.81%, and the error by random walk without drift is -39.36%.

[12]Note that the Akaike (AIC) or Bayes (BIC) information criteria suggest $p = 6$ and $p = 1$, repectively.

Forecasting Horizon	Recursive scheme			Rolling scheme		
	(1)	(2)	(3)	(4)	(5)	(6)
	Whole period	Pre-COVID-19	COVID-19	Whole period	Pre-COVID-19	COVID-19
$h = 1$	1.0000	1.0007	1.0012	1.0008	1.0026	0.9985
$h = 2$	0.9806	0.9218***	0.9986	0.9905	0.9301**	0.9993
$h = 3$	0.9932	1.0079	1.0003	0.9916	1.0064	1.0026
$h = 4$	1.0003	1.0004	0.9994	1.0012	0.9999	0.9997
$h = 5$	0.99957	1.0010	0.9994	0.9996	1.0003	0.9992

Notes. The table reports the ratio of RMSFE of the ADL model to the AR model.

*, **, and *** denote significance levels at the 10%, 5%, and 1% levels based on the Diebold-Mariano test statistics.

Forecast window covers 1/13/2020-3/1/2021 for whole period, 7/7/2019-2/15/2020 for before COVID-19, and 9/14/2020-3/1/2021 for COVID-19.

<Table 6. Forecasting results, RMSFE of ADL model against the AR model>

5. Conclusion

I find that Bitcoin returns respond to various variables, although the significance of regression models has been changing over time, and even signs of some estimates are not consistent. It implies that Bitcoin returns are very sensitive, and it is hard to find a robust predictor of Bitcoin returns.

I use the novel data of posts from the Bitcoin forum on the Reddit community, and construct the daily frequency scores to predict Bitcoin returns–the scores of Reddit posts. The biggest difference from other sentiment analysis for Bitcoin returns is that subscribers of Reddit community represent investors that have a high degree of interest. I find that their posts on the Reddit are more informative than Google Trends. It turns out the scores of Reddit posts play an important role in forecasting the Bitcoin returns, and it implies that investor's decision-making of buying and selling Bitcoin is affected by a market sentiment that comes from a Reddit community where people share their opinions and feelings about Bitcoin. That is, the sentiment score of Reddit posts is a good predictor to capture the shift of demand for Bitcoin. Unlike common stocks, Bitcoin does not have fundamentals (cash flow, debt-to-equity ratio, etc.), so its price is essentially driven by non-fundamental demand factors. Therefore, as we study in this paper, it is helpful to investigate people's opinions when we want to know the movement of Bitcoin.

References

[1] Anand, Abhinav, and Jalaj Pathak. "The role of Reddit in the GameStop short squeeze." *Economics Letters* 211 (2022): 11024ret

[2] Bring, Johan. "A geometric approach to compare variables in a regression model." *The American Statistician* 50.1 (1996): 57-62.

[3] Bolt, Wilko, and Maarten RC Van Oordt. "On the value of virtual currencies." *Journal of Money, Credit and Banking* 52.4 (2020): 835-862.

[4] Chen, Wei, Huilin Xu, Lifen Jia, and Ying Gao. "Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants." *International Journal of Forecasting* 37, no. 1 (2021): 28-43.

[5] Ciaian, Pavel, Miroslava Rajcaniova, and d'Artis Kancs. "The economics of BitCoin price formation." *Applied Economics,* 48.19 (2016): 1799-1815.

[6] Clark, Todd E., and Michael W. McCracken. "Improving forecast accuracy by combining recursive and rolling forecasts." *International Economic Review* 50.2 (2009): 363-395.

[7] Garcia, David, et al. "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy." *Journal of the Royal Society Interface* 11.99 (2014): 20140623.

[8] Garcia, Diego. "Sentiment during recessions." *The Journal of Finance* 68.3 (2013): 1267-1300.

[9] Halaburda, Hanna, Guillaume Haeringer, Joshua S. Gans, and Neil Gandal. The Microeconomics of Cryptocurrencies. No. w27477. *National Bureau of Economic Research*, 2020.

[10] Harvey, David, Stephen Leybourne, and Paul Newbold. "Testing the equality of prediction mean squared errors." *International Journal of Forecasting* 13.2 (1997): 281-291.

[11] Mai, Feng, Qing Bai, Jay Shan, Xin Shane Wang, and Roger HL Chiang. "The impacts of social media on Bitcoin performance." (2015).

[12] Pagnotta, Emiliano. "Decentralizing money: Bitcoin prices and blockchain security." *Review of Financial Studies*, 2020.

[13] Tetlock, Paul C. "Giving content to investor sentiment: The role of media in the stock market." *The Journal of Finance* 62.3 (2007): 1139-1168.

[14] Uhlig, Harald. "What are the effects of monetary policy on output? Results from an agnostic identification procedure." *Journal of Monetary Economics* 52.2 (2005): 381-419.

[15] Yermack, David. "Is Bitcoin a real currency? An economic appraisal." Handbook of digital currency. Academic Press, 2015. 31-43.

[16] Liu, Yukun, and Aleh Tsyvinski. "Risks and returns of cryptocurrency." *The Review of Financial Studies* 34.6 (2021): 2689-2727.

Appendix

This section describes classification models and examine out-of-sample forecasting for pre-COVID-19 and COVID-19 periods.

(1) Classification Trees

Classification tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches; Fine Tree – few leaves to make coarse distinctions between classe (maximum number of split is 4); Medium Tree – medium number of leaves for finer distinctions between classes (maximum number of splits is 20); Coarse Tree – many leaves to make many fine distinctions between classes (maximum number of splits is 100).

(2) Discriminant Analysis

Discriminant analysis classifies sets of patients or measures into two ore more groups (prior) on the basis of multiple characteristics simultaneously; Linear Discriminant – creates linear boundaries between classes; Quadratic Discriminant – creates nonlinear boundaries between classes (ellipse, parabola or hyperbola).

(3) Logistic Regression

Logistic regression uses a logistic function ($\frac{1}{1+exp(\frac{-x+\mu}{s})}$,where μ is a location parameter and s is a scale parameter) to classify a binary dependent variable.

(4) Naive Bayes Classifiers

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' theorem; Gaussian Naive Bayes – using normal distribution to classify the continuous values to each class ; Kernel Naive Bayes – kernel type options are Box, panechnikov, or Triangle for a classitication.

(5) Support Vector Machines

Support vector machines construct a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection; Linear SVM – makes a simple linear separation between classes; nonlinear SVM – specify the kernel function (Gaussian, quadratic, or cubic) to compute the classifier.

(6) $K-$Nearest Neighbor (KNN) Classifiers

Given a positive integer K and a test observation x_0, the KNN classifier identifes the K pointes in the training data that are closest to x_0; Fine KNN – K=1; Medium KNN – K=10; Course KNN – K=100; Cosine KNN – Medium distinctions between classes, using a Cosine distance metric; Cubic KNN – Medium distinctions between classes,

using a cubic distance metric; Weighted KNN –Medium distinctions between classes, using a distance weight.

(7) Ensemble Classifiers

Ensemble classifier is to combine the predictions of several base estimators, and it is distinguished by two families of averaging methods and boosting methods.

i. Boosted Trees: It fits consecutive trees by using different weights for every iteration, and derive the weighted final model to compute the classifier.

ii. Bagged Trees: It creates several subsets of data from n samplings of trading data with replacement, and each subset data is used to train their decision trees. The averaged final model to compute the classifier.

iii. Subspace Discriminant: It is an automated staging method by using random ensembles (subspace) with discriminant learners.

iv. Subspace KNN: It is an automated staging method by using random ensembles (subspace) with nearest neighbor learners.

v. Random Undersampling (RUS) Trees: It takes N, the number of members in the class with the fewest members in the training data, as the basic unit for sampling. Classes with more members are under sampled by taking only N observations of every class.

Table 7 and Table 8 report the out-of-sample forecast the Bitcoin return sign with and without Reddit score for pre-COVID-19 and COVID-19 , respectively. For pre-COVID-19, eighteen models perform better with the Reddit score than without of it, but these are not significant ($p-$value >10%) of the likelihood test.

Model	Accuracy (%)		Model	Accuracy (%)	
	w/ Reddit Score	w/o Reddit Score		w/ Reddit Score	w/o Reddit Score
Fine Tree	50.9	53.9	Fine KNN	53.0	50.9
Medium Tree	50.4	49.1	Medium KNN	51.3	49.6
Coarse Tree	54.7	54.7	Coarse KNN	50.4	50.0
Linear Discriminant	54.3	53.9	Cosine KNN	47.8	46.1
Quadratic Discriminant	48.7	47.8	Cubic KNN	53.9	46.1
Logistic Regression	54.7	54.3	Weighted KNN	53.0	51.3
Gaussian Naive Bayes	50.4	49.6	Boosted Trees	53.0	52.6
Kernel Naive Bayes	53.4	53.0	Bagged Trees	54.3	53.0
Linear SVM	52.6	50.0	Subspace Discriminant	53.4	52.2
Quadratic SVM	53.0	55.6	Subspace KNN	50.4	52.2
Cubic SVM	46.1	47.8	RUS Boosted Trees	53.0	49.6
Fine Gaussian SVM	54.0	50.4			
Medium Gaussian SVM	48.3	53.0			
Coarse Gaussian SVM	50.9	50.9			

Classification models for binary response variable, BTC Return (0-negative return and 1-positive return), with hold-out validation with 50% held out.
Training data set: 11/22/2018~7/12/2019, and testing data set: 7/13/2019-2/29/2020.
Features: Lagged return of BTC, ETH, LTC, and BCH. Lagged Active address of BTC and ETH, Lagged return of S&P500, Lagged GoogleTrends, and Lagged Sentiment score of Reddit.

<Table 7. Forecasting accuracy of classification models with and without the Reddit scores, pre-COVID-19>

For COVID-19 period, however, only eleven models outperform with the Reddit score. Although, none of them is significant same as the pre-COVID-19.

Model	Accuracy (%)		Model	Accuracy (%)	
	w/ Reddit Score	w/o Reddit Score		w/ Reddit Score	w/o Reddit Score
Fine Tree	57.1	60.4	Fine KNN	57.1	51.6
Medium Tree	58.2	59.3	Medium KNN	59.3	60.4
Coarse Tree	62.6	62.6	Coarse KNN	56.0	57.1
Linear Discriminant	58.2	57.1	Cosine KNN	59.3	58.2
Quadratic Discriminant	58.2	57.1	Cubic KNN	59.3	57.1
Logistic Regression	58.2	58.2	Weighted KNN	52.7	52.7
Gaussian Naive Bayes	59.3	59.3	Boosted Trees	59.3	59.3
Kernel Naive Bayes	59.3	58.2	Bagged Trees	58.2	54.9
Linear SVM	57.1	58.2	Subspace Discriminant	58.2	56.0
Quadratic SVM	59.3	57.1	Subspace KNN	58.2	52.7
Cubic SVM	64.8	57.1	RUS Boosted Trees	58.2	57.1
Fine Gaussian SVM	57.1	59.3			
Medium Gaussian SVM	58.2	60.4			
Coarse Gaussian SVM	53.8	54.9			

Classification models for binary response variable, BTC Return (0-negative return and 1-positive return), with hold-out validation with 25% held out.
Training data set: 3/1/2020~6/1/2020, and testing data set: 6/2/2020-3/1/2022.
Features: Lagged return of BTC, ETH, LTC, and BCH. Lagged Active address of BTC and ETH, Lagged return of S&P500, Lagged GoogleTrends, and Lagged Sentiment score of Reddit.

<Table 8. Forecasting accuracy of classification models with and without the Reddit scores, COVID-19>