Sentiment Analysis for Blockchain and Beyond

Jiayi Li

Supervisor: Prof. Luyao Zhang, PhD, Prof. Xin Tong, PhD.

Duke Kunshan University

2021.12

Table of Contents

Data Sources

We collected data from four sources. **English Tweets** #Cryptocurrency, Litecoin for example, are obtained from Twitter through Snscrape API, sustain real-time updates. Snscrape API outweighs Twitter API in terms of comprehensiveness of queries because Twitter API has an upper limit of 180 queries per 15 minutes. Besides the cryptocurrency, we will also collect data for tokens and NFT for future research. Cryptocurrency **Close Prices** are crawled from Coinmarketcap through a historical market price data scraper cryptocmd API on a daily interval. **Tweets Volume** is crawled from BitInfoCharts through Beautiful Soup on a daily interval. It records the number of related tweets per day for crypto, representing absolute popularity. **Google Trends** data is crawled from Google Trends through Pytrends API. The data point is divided by the total searches of the geography and time range it represents to compare relative popularity.

Meta Data Information

Table 1 A summary of Variables Collected for Further Analysis:

ID	Sources	Frequency	Unit	Description
Date	/	1 day	YYYY-MM-DD	Real-time update
Tweets	Twitter Crawler: Snscrape API	1 tweet	Strings (in English)	Tweets that contain the phrase #cryptocurrency, such as #Litecoin

Tweets Volume	BitInfoCharts Crawler: Beautiful Soup	1 day	number	Number of Tweets per day for a cryptocurrency
Close Price	Coinmarketcap Crawler: cryptocmd API	1 day	in dollar$	The daily close price of the specified cryptocurrency
Google trends data	Google Trends Crawler: Pytrends API	1 day	a range of 0 - 100	The resulting numbers are then scaled on a range of 0 - 100 based on a topic's proportion to all searches on all topics (Google trend support FAQ, 2021).

Data Processing

We use Valence Aware Dictionary for Sentiment Reasoning (Gilbert and Hutto, 2014), combining with the dictionaries that embed financial words, Loughran & McDonald financial corpus (Loughran and McDonald, 2011), and Harvard IV-4 psychological corpus, to carry out lexicon-based processing of raw sentiments.

Another option is to use Word2Vec.

2. Statistical Tests

Augmented Dickey-Fuller test

The augmented Dickey-Fuller test (ADF) (Cheung & Lai, 1995) is used for testing the stationarity of the data. Code for the ADF test refers to Statsmodels.

Vector Autoregressions

Vector Autoregressions (Stock & Watson, 2001) can be used for predicting multivariable time series. Combined with the Akaike information criterion (Akaike, 1973), Bayesian information criterion (Schwarz, 1978), or Final Prediction Error (FPE) criterion (Akaike, 1969), one is able to find optimal lags for the series. One needs to pay attention that the VAR class assumes that the passed time series is stationary. Code for the vector autoregressions refers to Statsmodel.

Causality testing

Granger-causality (Granger, 1969) can analyze the statistically significant patterns in lagged values of Tweets and Litecoin, and determine whether Twitter sentiment has predictive power (Mao et al., 2011). This test is able to figure out the causal relationship

between sentiment and crypto prices as well. Does A cause B or does B cause A? As long as the sentiment reflects or causes crypto price changes, we can conclude that, with several lags, the sentiment can have predictive power on the cryptocurrency market. Code for the Granger-causality test refers to Statsmodel.

Correlation Test and Heatmap

According to Figure 2, Google trend has a high correlation (0.63) with Tweet Volume. To avoid intercorrelation between the two independent variables, we only take Google Trends as final variable input.

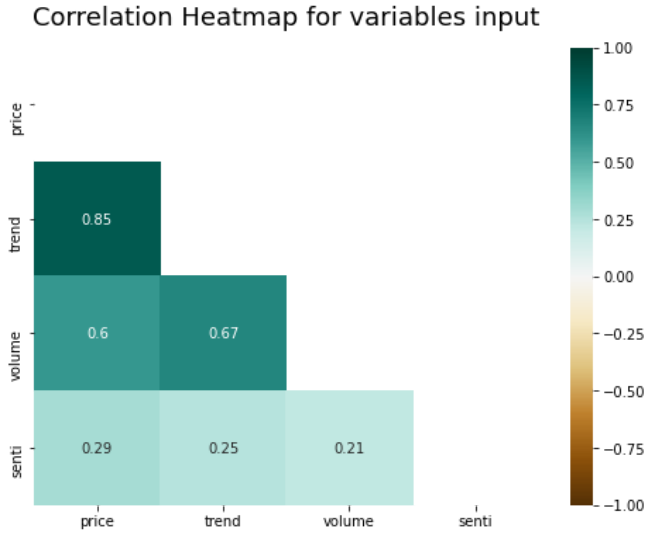


Figure 2: Correlation Heatmap for variables

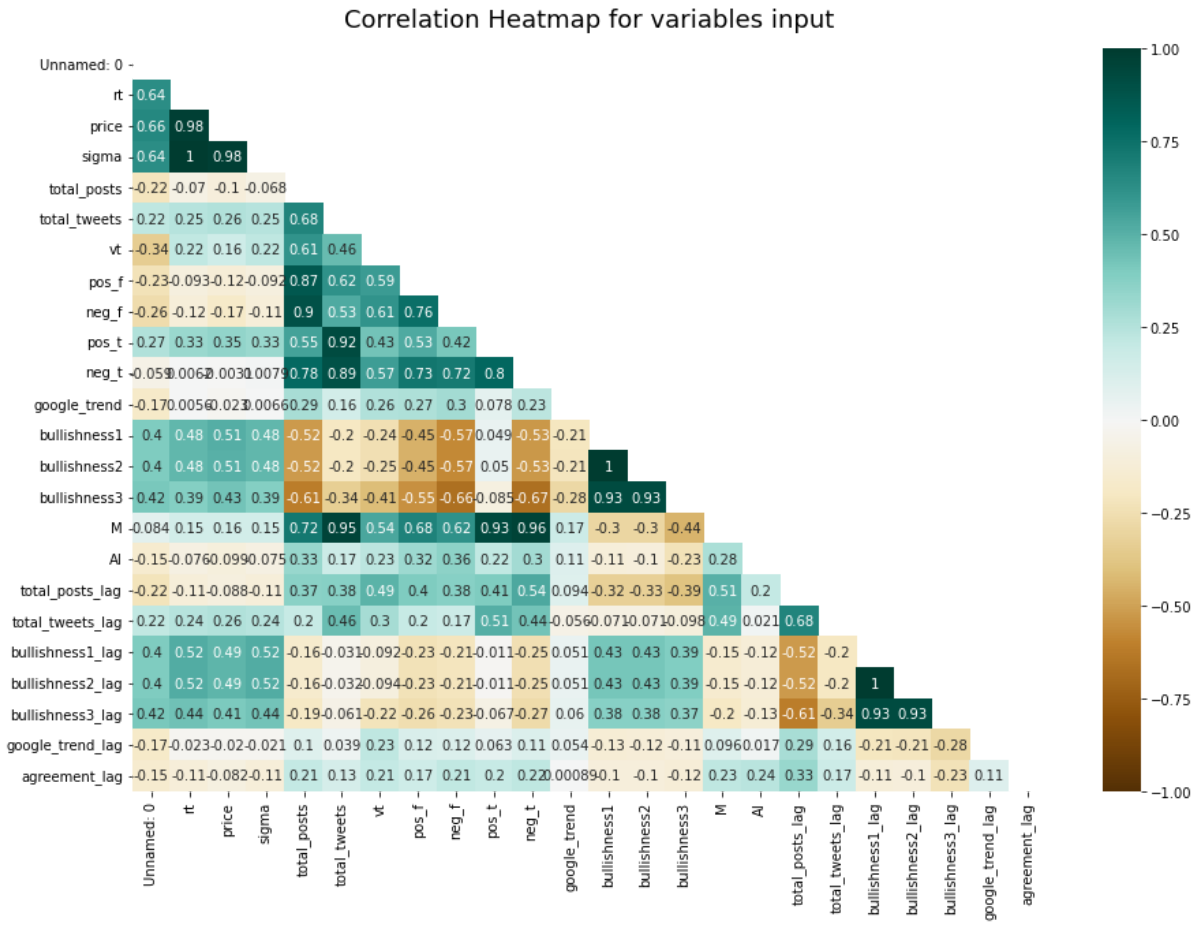


Figure 2: Please see replication3 for more details

3. Machine Learning for Market Trend Prediction

Based on the Granger-causality testing results, if the relationship between the token price and the public mood turns out to be non-linear as previous researchers have concluded, we will apply neural networks (NN) and compare with other models such as SVM, and RF. In addition, we will also simulate Bollen (2011) et al. 's approach of using Self Organizing Fuzzy Neural Networks (SOFNN) which obtained an accuracy of 86.7% in predicting daily changes in the closing values of the stock market.

4. Algorithmic Trading Strategy

In this session, we simulate algorithmic trading strategies and evaluate the ROI.

- **Strategy initiated by buy signal:** When cash is available, buy a certain number of shares. The buy signal is initiated when the predicted value is higher than the previous day.
- **Strategy initiated by sell-signal:** When shares are available, sell a certain number of shares, the sell signal is initiated when the predicted value is lower than the previous day.

- **Strategy at the beginning of the trading period:** Nothing if there is no buy/sell signal
- **Strategy at the end of the trading period:** Sell everything
- **Transaction fees**: Zero transaction fees
- **Return of Investment:** =(all the holdings on the last day-initial capital)/initial capital
- **Comparative Studies:** We make a comparison of the Newsreader Algorithm and the buy and hold strategy (invest all cash in the Adj Close on the first day, and sell all the stocks on the last day).

Pilot Results

In this pilot study, we replicate three literatures: All of them work on researching the influence of media on the cryptocurrency, especially bitcoin.

1. **Glaser (2014):**

Glaser, F., Zimmermann, K., Haferkorn, M., Weber, M. C., and Siering, M. (2014). Bitcoin-asset or currency? revealing users' hidden intentions. In ECIS 2014 Tel Aviv.

2. **Kristoufek (2015)**

x L. (2015). What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. PloS one, 10(4):e0123923.

3. **Mai et al. (2015)**

Mai, F., Bai, Q., Shan, Z., Wang, X. S., and Chiang, R. H. (2015). The impacts of social media on bitcoin performance. In Proceedings of the 2015 International Conference on Information Systems

Replication of current literature:

Replication1

https://colab.research.google.com/drive/1r5AehztY8LYHfd_HmxlKN8kq9k1Y754j?usp=sharing

1. Data

	Glaser(2014)	My Replication
Data Variables	Wikipedia Searches Bitcoin Prices Positive & Negative events Exchange Volume Network Volume	Tweet Volume Bitcoin Prices Positive & Negative Events
Data Range	2011-01-01 ~ 2013-10-02 (1004 observation)	2019-01-01 ~ 2021-10-28 (1031 observations)
Data Frequency	Daily	Daily
Method	ARCH/GARCH with Exogenous Regressors	ARCH/GARCH with Exogenous Regressors

2. MODEL : ARX+ARCH/GARCH

a) Main Model: AR model with exogenous variables(ARX) + ARCH/GARCH

This notebook provides examples of the accepted data structures for passing the expected value of exogenous variables when these are included in the mean. For example, consider an AR(1) with 2 exogenous variables. The mean dynamics are

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \beta_0 X_{0,t} + \beta_1 X_{1,t} + \epsilon_t.$$

Lag return · Exogeneous Variable · error term (ARCH, GARCH)

The h-step forecast, $E_T[Y_{t+h}]$, depends on the conditional expectation of $X_{0,T+h}$ and $X_{1,T+h}$,

$$E_T[Y_{T+h}] = \phi_0 + \phi_1 E_T[Y_{T+h-1}] + \beta_0 E_T[X_{0,T+h}] + \beta_1 E_T[X_{1,T+h}]$$

where $E_T[Y_{T+h-1}]$ has been recursively computed.

ARCH/GARCH (documentation)

ARCH models are a popular class of volatility models that use observed values of returns or residuals as volatility shocks. A basic GARCH model is specified as

$$r_t = \mu + \epsilon_t \quad (1.1)$$
$$\epsilon_t = \sigma_t e_t \quad (1.2)$$
$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \quad (1.3)$$

A complete ARCH model is divided into three components:

a) a mean model, e.g., a constant mean or an **ARX** (Autoregressive model with optional exogenous regressors estimation and simulation); (1.1) μ: $mean, \epsilon_t$: $error\ term$

b) a distribution for the standardized residuals (1.2). σ_t: $Error\ Variance$

c) a volatility process, e.g., a GARCH or an EGARCH process; (1.3)

$ARCH: \sigma_t^2 (Variance) = \omega(Constant) + \epsilon_{t-1}^2 (Lagged\ Squared\ Error)$

$GARCH: \sigma_t^2 (Variance) = \omega(Constant) + \epsilon_{t-1}^2 (Lagged\ Squared\ Error) + \sigma_{t-1}^2$

b) Glaser's Article

Glader's article uses the model with Exogenous Regressors, in the article the author listed the regression formula: ARX+ARCH/GARCH

$$r_t = \underbrace{a_0 + \sum_{i=1}^{7} a_i r_{t-i} + a_8 \Delta Wiki_t + a_9 Igood_t + a_{10} Ibad_t + \sum_{j=11}^{n} a_j \Delta C_{j,t}}_{\text{ARX}} + e_t, \quad \text{ARCH/GARCH} \quad (4)$$

$$e_t \sim N(0, h_t), \quad (5)$$

$$h_t = b_0 + b_1 e_{t-1}^2 + b_2 h_{t-1}. \quad (6)$$

Again, Δ designates first order differenced values, where r_t is the open-to-close Bitcoin return at date t. We then rely on the previous introduced exogenous variables starting with one to seven autoregressive terms of past open-to-close returns, changes in Wikipedia Bitcoin traffic, the event dummies and exchange as well as network volume changes.

c) Glaser's Result

The first model in Table below indicates that the interest of new users, proxied by the amount of Wikipedia traffic, does not, on average, influence future Bitcoin returns. These results remain consistent throughout all models. However, Model 2 to 4 indicate that there is an asymmetric bias of Bitcoin returns towards Bitcoin related events. Within all Models, the coefficient of positive events remains statistically larger than zero, indicating that such events temporarily drive Bitcoin returns. However, focusing on the coefficient of the negative events dummy, we do not observe any price correction effect after negative news within each of our models. While Model 2 only includes the events and the time series dummies, Model 3 and 4 further add additional controls and further lags of the endogenous variable, the results

however remain robust. Thus, we can conclude that Bitcoin users seem to be positively biased towards Bitcoin, while serious negative events, like thefts and hacks, are apparently not serious enough to lead to significant price corrections (H2 is accepted).

	Bitcoin Open-to-Close Return			
Variable	Model 1	Model 2	Model 3	Model 4
$OC - Return_{t-1}$	0.044 (0.348)	0.049 (0.306)	0.058 (0.212)	0.061 (0.182)
$OC - Return_{t-2}$				0.041 (0.254)
$OC - Return_{t-3}$				-0.024 (0.536)
$OC - Return_{t-4}$				0.033 (0.395)
$OC - Return_{t-5}$				-0.003 (0.913)
$OC - Return_{t-6}$				0.088 (0.017)
$OC - Return_{t-7}$				0.000 (0.994)
$\Delta Wiki_t$	0.000 (0.225)	0.000 (0.290)	0.000 (0.303)	0.000 (0.411)
$\Delta ExchangeVolume_t$			-0.001 (0.000)	-0.000 (0.000)
$\Delta NetworkVolume_t$			0.000 (0.530)	0.000 (0.460)
$PositiveEvents_t$		0.015 (0.027)	0.013 (0.072)	0.014 (0.045)
$NegativeEvents_t$		0.000 (0.967)	0.002 (0.637)	0.003 (0.543)
Constant	0.023 (0.000)	0.022 (0.000)	0.026 (0.000)	0.022 (0.000)
Time Dummies	Yes	Yes	Yes	No
ARCH - Coefficients				
ARCH	0.458 (0.000)	0.437 (0.000)	0.325 (0.000)	0.320 (0.000)
GARCH	0.693 (0.000)	0.703 (0.000)	0.755 (0.000)	0.756 (0.000)
Statistics				
BIC	-2,971	-2,959	-2,972	-2,922
AIC	-3,094	-3,092	-3,114	-3,094
Observations	1,005	1,005	1,005	999

Table 2. Regression Results on the Bitcoin Exchange Rate (p-values in parentheses).

3. My Replication

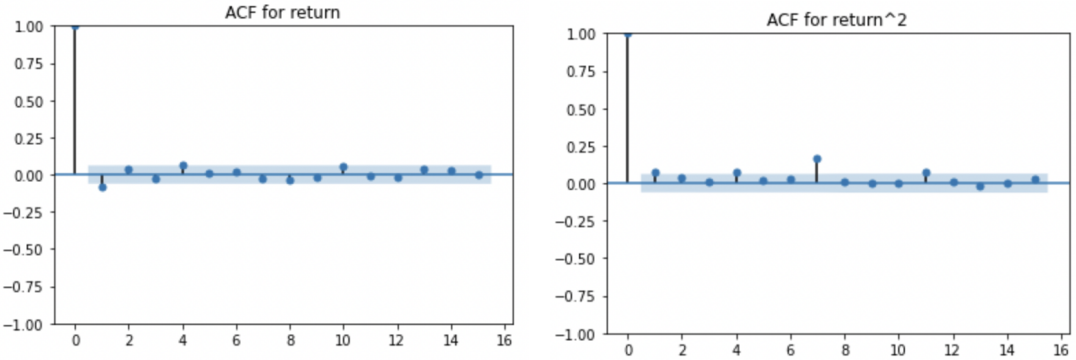
a) ACF/PACF TEST

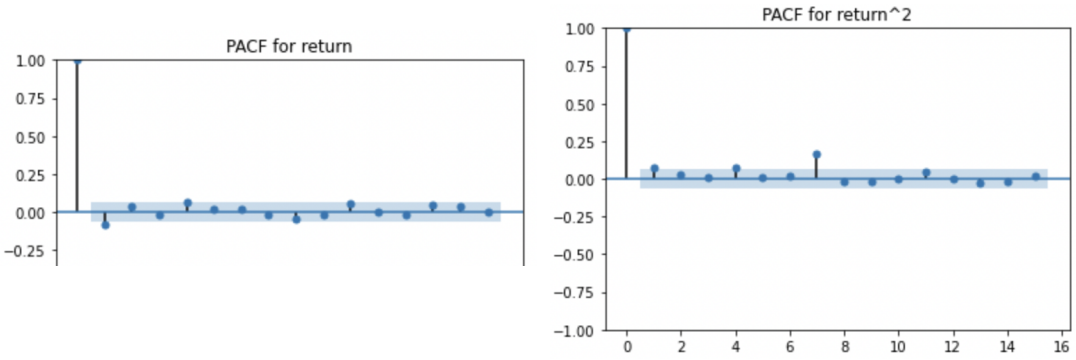
identify an ARCH/GARCH Model in Practive (Psu.edu)

The best identification tool may be a time series plot of the series. It's usually easy to spot periods of increased variation sprinkled through the series. It can be fruitful to look at the ACF and PACF of both y_t and y_t^2. For instance, if y_t appears to be white noise and y_t^2 appears to be AR(1), then an ARCH(1) model for the variance is suggested. If the PACF of the y_t^2 suggests AR(m), then ARCH(m) may work. GARCH models may be suggested by an ARMA type look to the ACF and PACF of y_t^2. In practice, things won't always fall into place as nicely as they did for the simulated example in this lesson. You might have to experiment with various ARCH and GARCH structures after spotting the need in the time series plot of the series.

≈

PACF, ACF tests for returns:





From the above plots, we figure out that $return$ appears to be white noise and $return^2$ is auto-correlated with the lag term t-7. Therefore, an ARCH(7) model for the variance is suggested. GARCH(7,7) is also suggested.

b) Model Preparation

	Input	Output
Model1	Lag return 1 Tweet Volume Difference	return
Model2	Lag return 1 Tweet Volume Difference Positive Events Negative Events	return

Model3	Lag return 1,2,3,4,5,6,7 Tweet Volume Difference Positive Events Negative Events	return

c) Result

The first model in Table below indicates that popularity of bitcoin, proxied by the Tweet Volume can influence future Bitcoin returns (P-value quite small). These results remain consistent throughout all models. Unlike Glaser's result, Model 2 to 3 indicate that the coefficients of both positive events are not statistically significant.

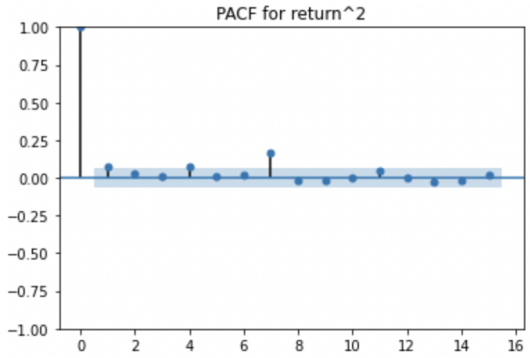
Alpha (1) and beta (1) of the GARCH were all significant (see table below for the model output). This shows that the variance did depend on the lag structure.

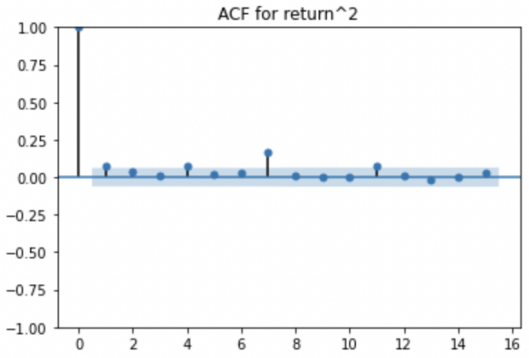
Variable		Bitcoin Open-to-Close Returns		
		Model 1	Model 2	Model 3
Lagged Return	OC – Return t-1	-0.020 (0.959)	-0.021 (0.839)	-0.023 (0.948)
	OC – Return t-2			0.035 (0.706)
	OC – Return t-3			0.010 (0.961)
	OC – Return t-4			0.110 (0.887)
	OC – Return t-5			0.012 (0.944)
	OC – Return t-6			0.020 (0.713)
	OC – Return t-7			0.008 (0.943)
Exogeneous Variables	Diff Twitter Vol-1	0.136 (0.001)	0.131 (0.186)	0.137 (0.001)
	Positive Events		-0.016 (0.648)	-0.016 (0.649)
	Negative Events		-0.030 (0.354)	-0.030 (0.343)
	Constant	0.005	0.005 (0.843)	0.007 (0.793)

ARCH Coefficients				
Error Terms	ARCH	0.101 (0.053)	0.103 (0.055)	0.104 (0.056)
	GARCH	0.806 (0.000)	0.805 (0.000)	0.804 (0.000)
Statistics				
AIC		2757.79	2760.48	2754.82
BIC		2821.59	2834.10	2857.76
Observation		1000	1000	994

4. Possible Errors and Changes

a) Glaser's model uses GARCH(1,1), while I use GARCH(7,7) by observing the lags in my dataset through PACF test.

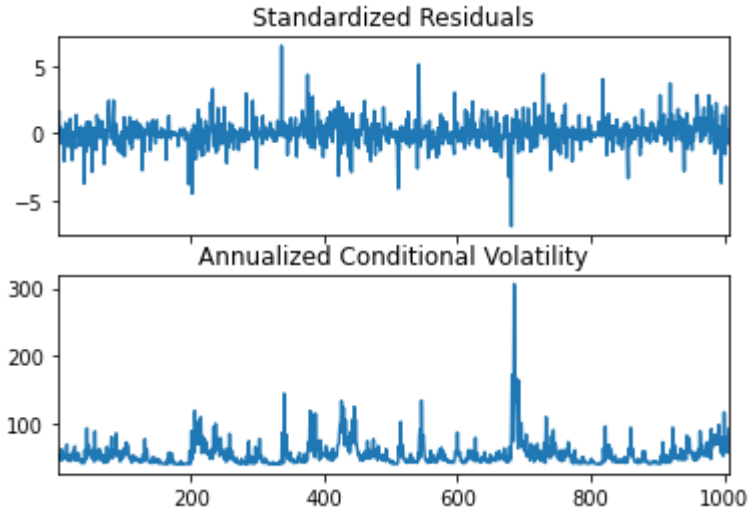




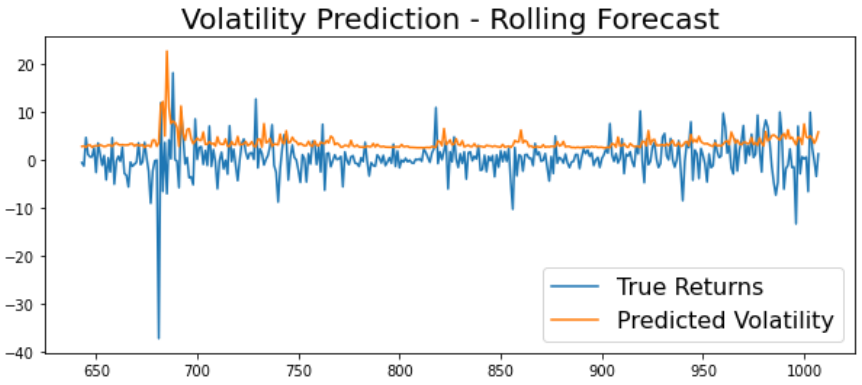
b) Glaser's model gain positive/negative events from https://en.bitcoin.it/wiki/History (only 24 events), while I manually gain the data from https://cryptoq8.com/cryptoq8-portal/cryptocurrency/bitcoin/ (37 events). Both work's data for positive/negative is not big.
Therefore I suspect that the results Glaser gains, that *"Bitcoin users are positively biased towards Bitcoin, while serious negative events, like thefts and hacks, will not lead to significant price corrections"* might just be a coincidence!

Explorations1 :

a) Only applies GARCH, INPUT: Daily Return ; OUTPUT: conditional volatility

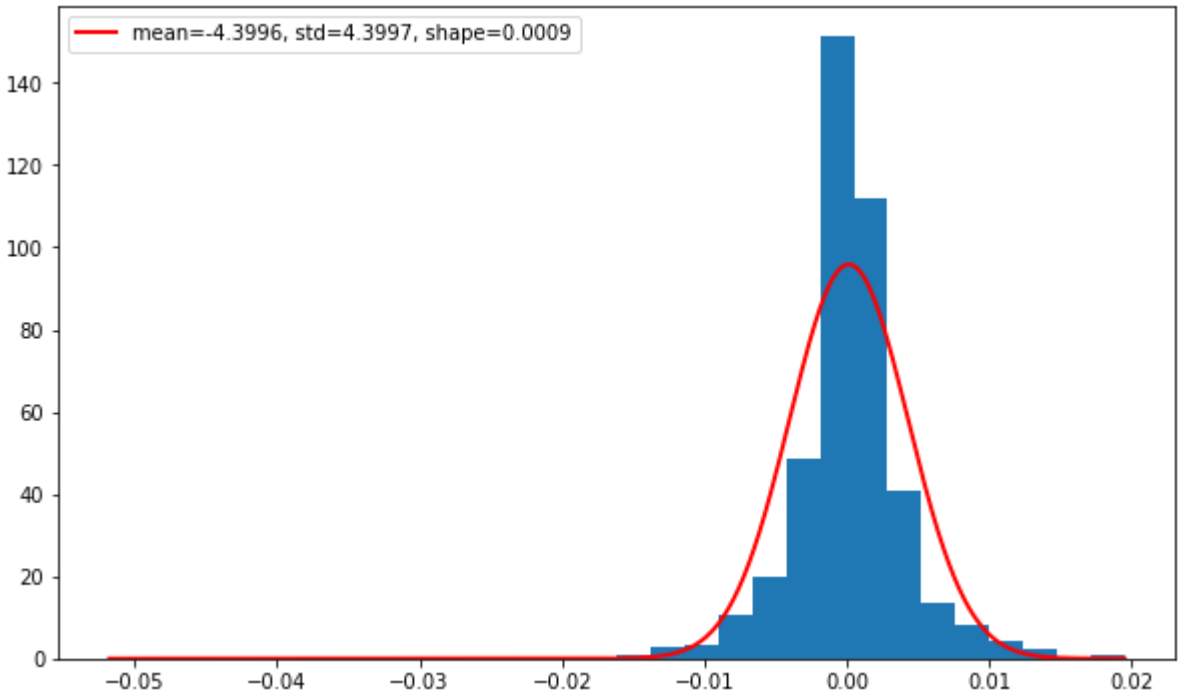


Rolling Prediction of Volatility(365天)



Explorations2 :

a) Use Log returns to run a regression (with no ARCH/GARCH)
Log returns distribution:



b) Results

	Input	Output
Model1	Lag Log return 1 Tweet Volume Difference	return
Model2	Lag Log return 1 Tweet Volume Difference Positive Events Negative Events	return
Model3	Lag Log return 1,2,3,4,5,6,7 Tweet Volume Difference Positive Events Negative Events	return

We can see from the Table below that the Tweet Volume Difference and Past Log returns have small P-value, which means that they can influence the future returns. These results remain consistent throughout all models.

Variable		Bitcoin Open-to-Close Returns		
		Model 1	Model 2	Model 3
Lagged Return	Log Return t-1	-0.065 (0.035)	-0.067 (0.033)	-0.063 (0.046)
	Log – Return t-2			0.064 (0.041)

	Log – Return t-3			-0.008 (0.789)
	Log – Return t-4			0.052 (0.097)
	Log – Return t-5			0.024 (0.442)
	Log – Return t-6			0.042 (0.182)
	Log – Return t-7			-0.039 (0.214)
Exogeneous Variables	Diff Twitter Vol-1	0.126 (0.000)	0.128 (0.000)	0.133 (0.000)
	Positive Events		-0.015 (0.625)	-0.015 (0.641)
	Negative Events		-0.024 (0.431)	-0.028 (0.369)
	Constant	0.000 (0.998)	0.000 (1.000)	0.000 (0.991)
Statistics				
AIC		2842	2845	2830
BIC		2857	2870	2884
Observation		1006	1006	1000

Colab Link:
https://colab.research.google.com/drive/1sscpsOEGmVU1R8wPvRrvcQw7_y3P-00j?usp=sharing

REPLICATION2 : Kristoufek (2013) & Kristoufek (2015)

1. Data

	Kristoufek (2013)	Kristoufek (2015)	My Replication

Data Variables	The Logarithmic Returns of Stock Indices (NASDAQ, FTSE, CAC, DAX, HSI, NIKKEI)	Wikipedia Searches, Google Trend, Bitcoin Price	Wikipedia Searches, Google Trend, Bitcoin Price, Logarithmic Returns of Bitcoin Price
Data Range	2000-2013	2011-09-14 ~ 2014-02-28	2017-01-01 ~ 2019-11-23
Data Frequency	Daily	Daily	Daily
Method	Continuous wavelet transform	Wavelet Coherence	Continuous wavelet transform; https://github.com/mabelcalim/waipy Wavelet Coherence (Model Code Reference)

2. Objective:

(Kristoufek, 2013)

"Fractal Markets Hypothesis and the Global Financial Crisis:Wavelet Power Evidence"

The author analyzes whether the prediction of the fractal markets hypothesis about a dominance of specific investment horizons during turbulent times holds. To do so, the author utilizes the continuous wavelet transform which gives crucial information about the variance distribution across scales and its evolution in time. He shows that the most turbulent times of the Global Financial Crisis can be very well characterized by the dominance of short investment horizons which is in hand with the assertions of the fractal markets hypothesis.

The fractal markets hypothesis (FMH) argues that when What is considered negative information and thus a selling signal for an investor with a short horizon might be a buying opportunity for an investor with a long horizon, and vice versa. If a sufficient number of buyers and sellers trade and are efficiently cleared in the market mechanism, a smooth functioning of this market is guaranteed. If investment horizons are uniformly represented in the market. However, if an investment horizon (or a group of horizons) becomes dominant, buying and selling orders are not efficiently cleared and extreme events are likely to occur. Therefore, FMH directly predicts that critical events are connected to dominating investment horizons.

To uncover whether this prediction of FMH is correct for the Global Financial Crisis (GFC) of the late 2000s, the author applied the continuous wavelet transform procedure on developed and liquid world stock indices. In terms of financial economics, the wavelet power can be understood as a scale-specific variance.

According to the FMH arguments, one should observe increased power at low scales (high frequencies) during the critical periods. Moreover, one might observe a changing structure of variance across frequencies.

(Kristoufek, 2015)

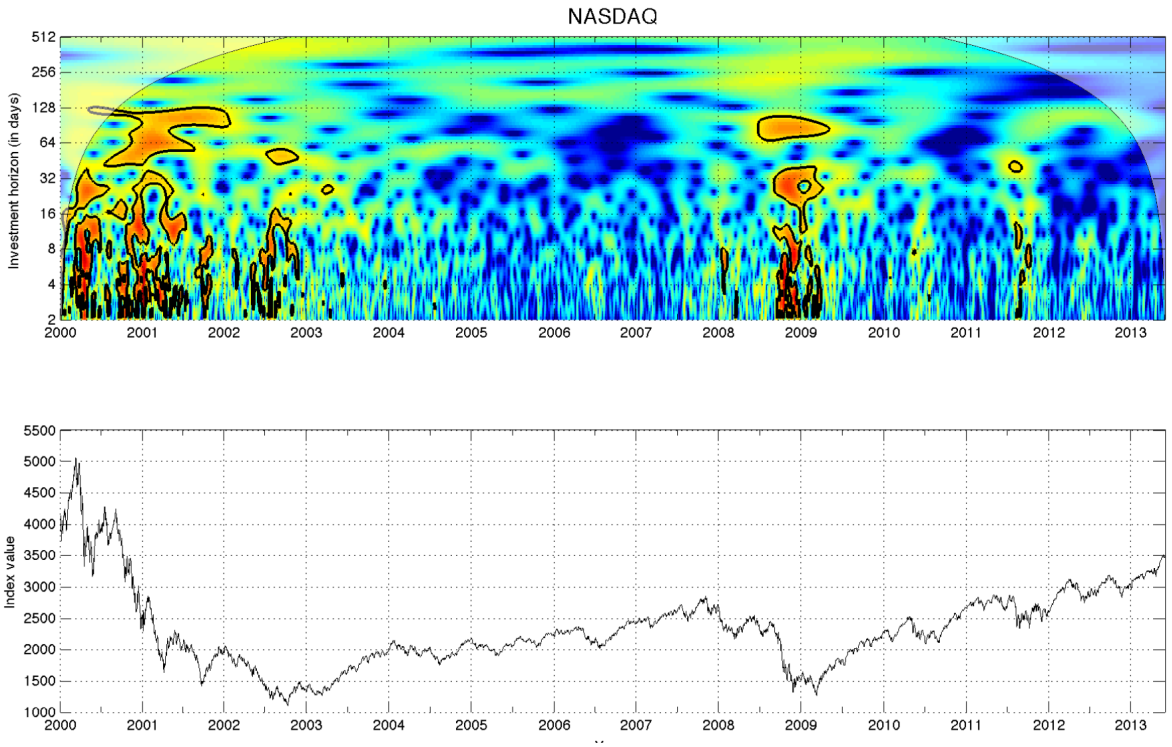
"What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis"

The author wants to focus on various possible sources of price movements, ranging from fundamental sources to speculative and technical sources, and examine how the interconnections behave in time but also at different scales (frequencies). To do so, he utilized continuous wavelet analysis, specifically wavelet coherence, which can localize correlations between series and evolution in time and across scales.

1. It must be stressed that both time and frequency are important for Bitcoin price dynamics because the currency has undergone a wild evolution in recent years, and it would thus be naive to believe that the driving forces of the prices have remained unchanged during its existence.

2. In addition, the frequency domain viewpoint provides an opportunity to distinguish between short and long-term correlations. We show that the time and frequency characteristics of the dynamics are indeed both worth investigating, and various interesting relationships are uncovered.

3. Kristoufek's Result - (2013)



Wavelet power spectrum for NASDAQ. Significant wavelet powers against the null hypothesis of a red noise are marked off by a bold black line. Cone of influence separates the spectrum into two – a top (pale) part where the inference is less reliable and a bottom part (colorful)

where the results are reliable. The hotter the color, the higher the power (variance) at the specific scale (y-axis) and time (x-axis). Statistically significant regions are bordered by a thick black curve. As the data resolution is daily, the scales are in day units. Wavelet power is shown for the whole analyzed period (upper chart) together with evolution of the index value (bottom chart). Dominance of high frequencies is evident for the most turbulent times of the GFC but also for other critical trend changes and corrections so that the results are in hand with fractal markets hypothesis prediction.

Two results:

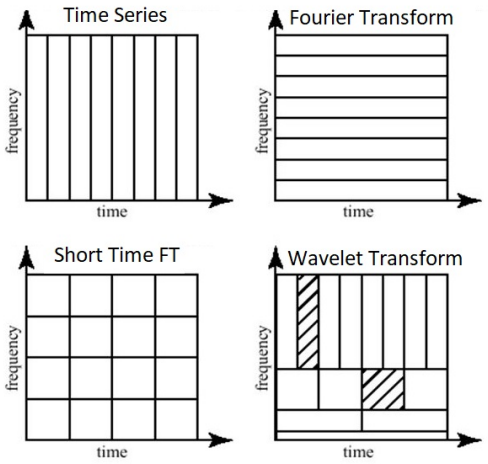
1. In most of the time, there is no scale-characteristic power which can be distinguished from a pure noise so that no investment horizon dominates. The whole period between 2003 and 2008 is characterized by low total variance at all frequencies and the colors of the power spectra remain cold. The period after the Dot-Com bubble is connected to a uniformly distributed activity at separate investment horizons (or at least, there is no dominating horizon), which is well in hand with FMH.
2. The period between September and December 2008 is strongly dominated by an increased energy at the highest frequencies. Even though the variance increases at more scales, the dominance of the very low scales is apparent (very hot colors even bordering with black for October and November 2008) and is again in hand with the predictions of FMH – the turbulent times are connected with the dominance of specific investment horizons so that the efficient market clearing is not possible.

Replication- Kristoufek (2013)

Wavelet power spectrum for Bitcoin returns. Cone of influence separates the spectrum into two – a top (colorful) part where the inference is reliable and a bottom part (stripe) where the results are less reliable. The hotter the color, the higher the power (variance) at the specific scale (y-axis) and time (x-axis). Y axis signifies investment horizons in days, representing low scales (which corresponds to high frequency) indicating time information, to higher scales (which corresponds to high frequency) indicating frequency information;

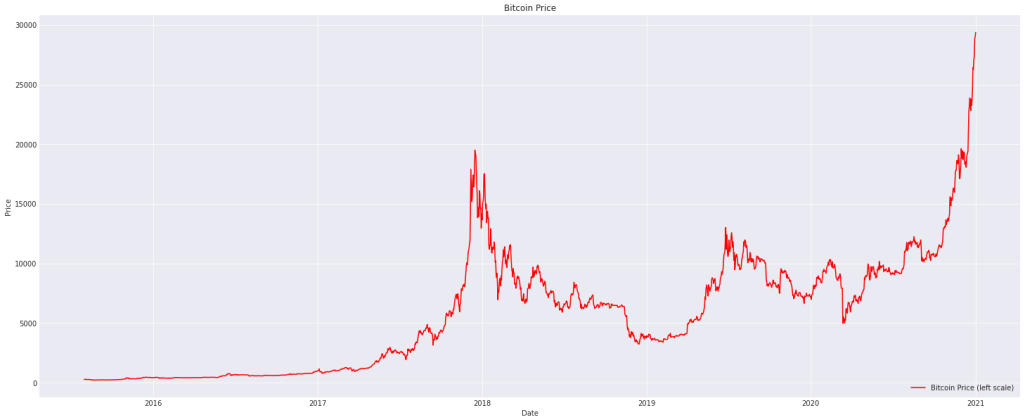
Scale is a useful property of signals. For example, you can analyze temperature data for changes on different scales. You can look at year-to-year or decade-to-decade changes. Of course, you can examine finer (day-to-day), or coarser scale changes as well. Some processes reveal interesting changes on long time, or spatial scales that are not evident on small time or spatial scales.

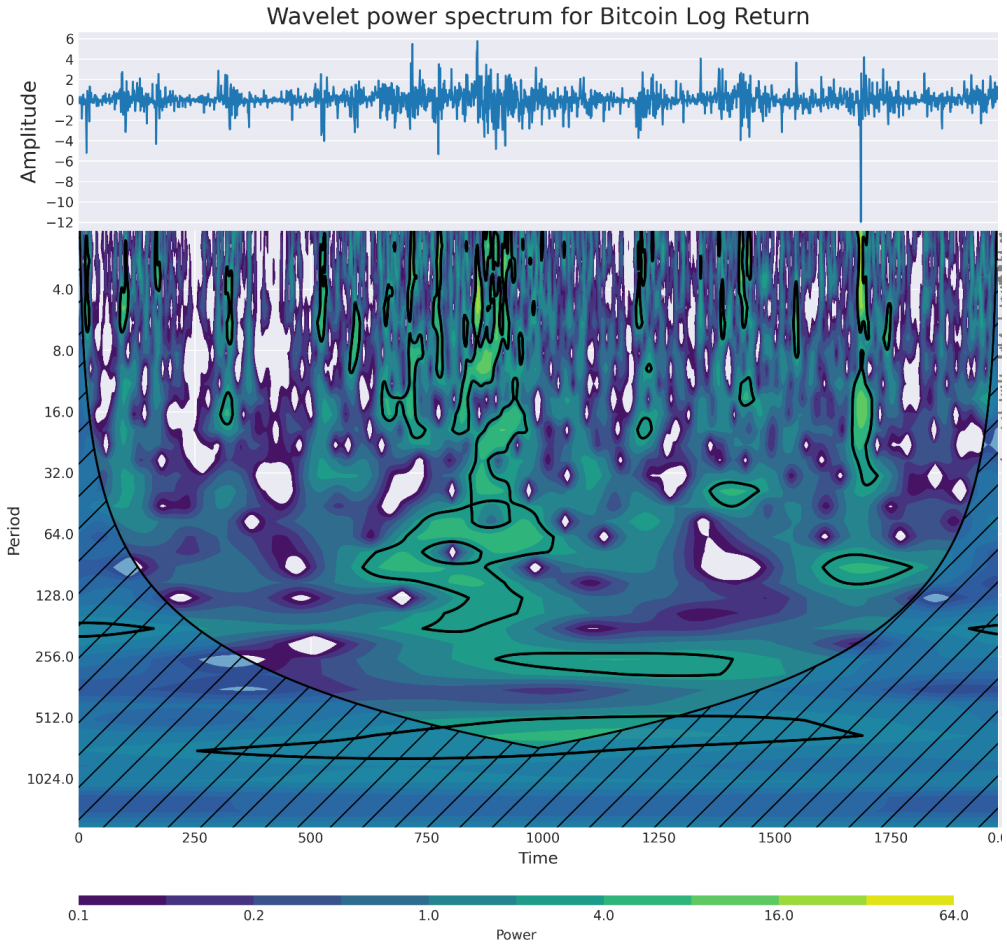
The below picture illustrates that, Fourier Transformation obtains Frequency Spectrum, Time Series obtains time information, STFT approach has both frequency and time but the window size is constant. In this case, wavelet transform stands out because: Wavelet analysis allows the use of long time intervals where we want more precise low-frequency (high-scale) information, and shorter regions where we want high-frequency(low-scale) information



In our case, continuous wavelength transformation coefficient metrics consists of 1024 rows representing the scales, and almost 5000 columns representing the days, each day represented by one price in our time series, each price is considered under 1024 scales. More precise definition of scales please see here: https://ww2.mathworks.cn/help/wavelet/gs/continuous-wavelet-transform-and-scale-based-analysis.html.

Statistically significant regions are bordered by a thick black curve. As the data resolution is daily, the scales are in day units. Wavelet power is shown for the whole analyzed period (lower chart) together with evolution of the price return value (upper chart).





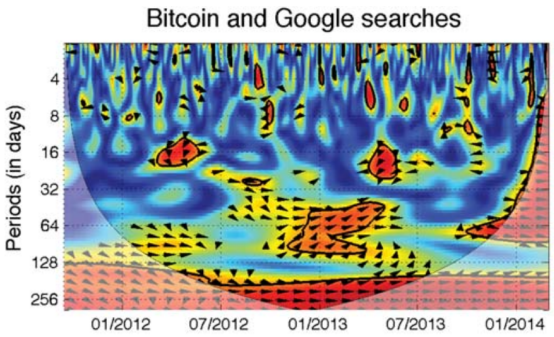
Results:

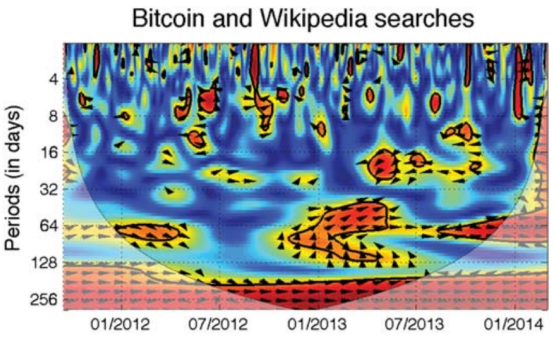
1. In most of the time, there is no scale-characteristic power which can be distinguished from a pure noise so that no investment horizon dominates.
2. The period of late 2017 is strongly dominated by an increased energy at the highest frequencies. Even though the variance increases at more scales, the dominance of the very low scales is apparent (warmer colors bordering with black for late 2017 and early 2008) and is in hand with the predictions of FMH – the turbulent times are connected with the dominance of specific investment horizons so that the efficient market clearing is not possible.

4. Kristoufek's Result - (2015)

Time Range: 2011-09-14 ~ 2014-02-28

Kristoufek (2015) who uses wavelet analysis to analyze the fluctuations of Bitcoin's price. He finds that until the first half of 2012, Bitcoin's price jumps boosted Google searches, while after 2013 the relation is reversed, i.e The number of Google searches seems to influence the price positively.





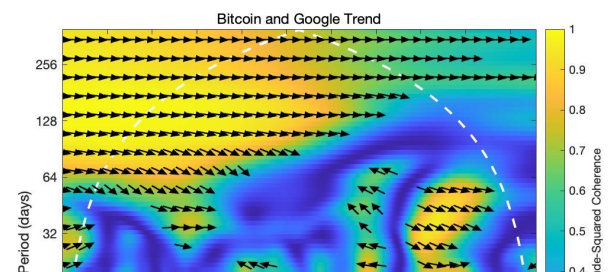
* How to interpret the Wavelet Coherence Graph?

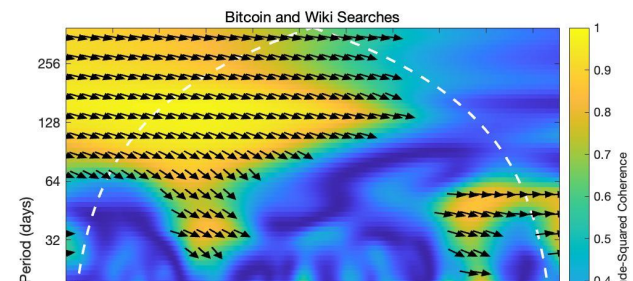
Interpretation:

- Time is displayed on the horizontal axis, while the y axis shows the scale, in this case, the length of days or period (the lower the scale, the higher the frequency).
- Continuous wavelength transformation coefficient metrics which consists of 256 rows representing the scales, and almost 1000 columns representing the days, each day represented by one price in our time series, each price is considered under 256 scales.
- Regions in time-frequency space where the two time series co-vary are located by the wavelet coherence.
- Warmer colors (red) represent regions with significant interrelation, while colder colors (blue) signify lower dependence between the series. Cold regions beyond the significant areas represent time and frequencies with no dependence in the series.
- An arrow in the wavelet coherence plots represents the lead/lag phase relations between the examined series. A zero phase difference means that the two time series move together on a particular scale. Arrows point to the right (left) when the time series are in phase (anti-phase).
- When the two series are in phase, it indicates that they move in the same direction, and anti-phase means that they move in the opposite direction. Arrows pointing to the right-down or left-up indicate that the first variable(in my case, Bitcoin) is leading, while
- arrows pointing to the right-up or left-down show that the second variable(in my case, Google Trend or Wiki Searches) is leading.

Replication- Kristoufek (2015)

1. For Google Trend, in the long run, Google Trend is positively correlated with the Bitcoin price, with no evident leaders. However, in the short run, Bitcoin leads the Google Trend positively.
2. For Wiki Searches, in the long run, especially in the two thirds of the period, Wiki Search is positively correlated with the Bitcoin price, with Bitcoin leading the Wiki Search. However, in the last third of the period, there is no significant relationship between Wiki search and Bitcoin price in the long run, while there are several significant episodes at the lower scales illustrating the positive correlation between Bitcoin and Wiki Search.





Code:

```
google = replication22(:,{'google_trend'});
wiki = replication22(:,{'wiki'});
price= replication22(:,{'price'});
google= table2array(google);
wiki= table2array(wiki);
price=table2array(price);
wcoherence(price,wiki,days(1));
title("Bitcoin and Wiki Searches")
wcoherence(price,google,days(1));
title("Bitcoin and Google Trend")
```

Replication 3

Mai, Feng & Bai, Qing & Shan, Jay & Wang, Shane & Chiang, Roger. (2015). From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance. SSRN Electronic Journal. 10.2139/ssrn.2545957.Colab Link: https://colab.research.google.com/drive/1w5FLOA7QxR07feUYR5QYU-YVSLT15mwv?usp=sharing

Data Collection

- Daily Bitcoin Price and trading volume is crawled from cryptocmd API
- Reddit Comments are crawled through PushShift API (20, 000+)
- Tweets are crawled through snscrape API (300, 000+)
- Google trend is crawled through pytrend API

Sentiment analysis - Lexicon

- Finance sentiment dictionary (Loughran and McDonald 2014),

Table 1: Key Measures:

Variable	Meaning
P_t	Daily bitcoin price in USD
r_t	Bitcoin returns, continuously compounded
σ_t^2	Volatility of bitcoin returns, updated daily
V_t	Daily trading volume (logged, linear trend removed) from top
POS^F	Number of positive Reddit posts in a day
NEG^F	Number of negative Reddit posts in a day
POS^T	Number of positive tweets in a day
NEG^T	Number of negative tweets in a day

we seek to determine whether variation in social media activities is just noise or is associated with underlying market activities.

We carry out three bullishness measurement:

$$BullishnessI = (POS^T - NEG^T)/(POS^T + NEG^T)$$

$$BullishnessII = In(1 + POS^T)/(1 + NEG^T)$$

$$BullishnessIII = POS^T - NEG^T$$

$$M = POS^T + NEG^T$$

To measure disagreement among forum contributors, we constructed an agreement index (Antweiler and Frank, 2004)

$$AI = 1 - \sqrt{1 - bullishness\,I^2}$$

Dynamic Relationships

Model Summary:

Model	Endogenous Variable	Exogenous Variables	Sample Period	Sampling Frequency
1	P_t, σ_t^2, V_t, POS^T, NEG^T	Yes	5/1/2021-12/1/2021	Daily
2	P_t, σ_t^2, V_t, M, $bullishness$, AI	Yes	5/1/2021-12/1/2021	Daily
3	P_t, σ_t^2, V_t, POS^F, NEG^F, POS^T, NEG^T	Yes	5/1/2021-12/1/2021	Daily

Stationarity: Number of positive & negative reddit posts, Number of positive & negative Tweets are stationary. Bitcoin Price, Returns, Volatility and Volume are non-stationary, and have one order of integration.

We determine the Lag length p using Akaike's information criterion (AIC) for each model.

We run the model through a coin integration test and select the coin rank for each model.

We use the VECM Model:

Results:

Pairwise Correlations

	Return	Volatility	Trading Volume
Volatility			
Trading Volume	0.217	0.219	
#Post (Tweets)	0.247	0.248	0.464
#Post (Tweets), 1d lag	0.244	0.244	0.295
#Post (Reddits)	-0.070	-0.068	0.609
#Post (Reddits), 1d lag	-0.114	-0.112	0.492
Bullishness I	0.480	0.478	-0.244
Bullishness I, 1d lag	0.521	0.519	-0.092
Bullishness II	0.478	0.477	-0.246
Bullishness II, 1d lag	0.519	0.517	-0.093

Bullishness III	0.388	0.387	-0.407
Bullishness III, 1d lag	0.438	0.436	-0.221
Agreement	-0.076	-0.074	0.230
Agreement,1d lag	-0.112	-0.110	0.210
Google Trend	0.005	0.006	0.259
Google Trend, 1d lag	-0.022	-0.021	0.225

Model 1: Dynamic Relationship among Bitcoin Returns, Volatility and Tweet Polarity

Dependent Variable	Independent Variable			
	ΔP_{t-1}	ΔV_{t-1}	$\Delta POS^{T}{}_{t-1}$	$\Delta NEG^{T}{}_{t-1}$
ΔP_{t}	0.006	619.7	-12.10***	9.249***
ΔV_{t}	0.000	-0.272***	0.000	0.000
ΔPOS^{T}	0.000	-31.16	0.112	0.046
ΔNEG^{T}	0.000	-14,87	0.105	0.163

Model 2: Dynamic Relationship among Bitcoin Returns, Volatility, Bullishness and M

	ΔP_{t-1}	$\Delta \sigma_{t-1}^{2}$	ΔM_{t-1}	$\Delta Bullishness_{t-1}$	ΔAI_{t-1}
ΔP_{t}	-0.023	-575.5	-1.061	-11.60***	-5458
$\Delta \sigma_{t}^{2}$	0.000***	0.055***	0.000	0.000	-0.165*
ΔM_{t}	-0.006	81.40	-0.112	0.043	-126.4

$\Delta Bullishness_t$	0.006***	185.1	-0.011	-0.166***	-90.98
ΔAI_t	0.000	0.001	0.000	0.000	0.008

Model 3: Dynamic Relationship among Bitcoin Returns, Volatility, Bullishness and M

	ΔP_{t-1}	$\Delta\sigma_{t-1}^2$	ΔV_{t-1}	$\Delta NEG^T{}_{t-1}$	$\Delta POS^T{}_{t-1}$	$\Delta NEG^F{}_{t-}$	$\Delta POS^F{}_{t-1}$
ΔP_t	-0.003	1935	1149	12.21***	-12.73***	-39.13***	-0.160
$\Delta\sigma_t^2$	0.000***	0.050***	0.005***	0.000	0.000	0.000	0.000
ΔV_t	0.000	0.332	-0.380***	0.000	0.000***	0.002	0.007***
$\Delta NEG^T{}_t$	0.000	-36.34	-60.73*	-0.247**	0.038	0.067	3.231***
$\Delta POS^T{}_t$	0.004**	86.96	-53.33***	-0.084	-0.066	0.330	1.589*
$\Delta NEG^F{}_t$	-0.001***	5.421	-9.060***	-0.005	0.006	-0.003	0.205*
$\Delta POS^F{}_t$	0.000***	9.081	-2.579*	-0.016***	0.015***	0.006	0.056