**Nijmegen School of Management**
**Department of Economics and Business Economics**
**Master's Thesis Economics (MAN-MTHEC)**

# Cryptocurrency price prediction: A narrative-based approach to sentiment analysis

By Paul Hoekstra (s4308565)

Nijmegen, 30 June 2022

Program: Master's Program in Economics
Specialisation: Corporate Finance and Control
Supervisor: dr. J. Schmitz

Radboud Universiteit

# Abstract

This paper takes a novel approach to sentiment analysis of cryptocurrency-related Twitter data by applying concepts from narrative economics. The aim is to determine whether tweet engagement and sentiment metrics are predictors of cryptocurrency price returns and trading volumes. The machine-learning algorithm latent Dirichlet Allocation (LDA) was used on a dataset consisting of cryptocurrency-related tweets to unveil the following four narratives: Decentralised Finance (DeFi), Non-fungible tokens (NFTs), Gaming and Memecoins. Empirical analysis consisting of Granger causality testing and OLS regressions revealed a complex relationship between tweet engagement and cryptocurrency prices, where both are predictive of each other. Out of the identified narratives, the Memecoin narrative was found to hold the most predictive power over cryptocurrency prices and trading volumes. A strong association between the S&P500 stock market index and cryptocurrency prices was also revealed, which goes against the common belief that cryptocurrencies are able to act as a hedge against traditional markets.

# Table of Contents

# 1 Introduction

The concept of narrative economics, as introduced by Nobel Laureate Robert Shiller in his influential 2017 paper, concerns the contagion of narratives (i.e., stories and ideas). Drawing similarities to the spread of infectious diseases, Shiller describes how contagious stories have the potential to drive major economic events and argues for the incorporation of this important mechanism for economic change into economic theory (Shiller, 2017). Building further on this concept in his subsequent book, he argues that financial bubbles and crashes are largely driven by narratives and devotes an entire section of the book to the important role which narratives have played in the rise of Bitcoin and cryptocurrency (Shiller, 2020). While this paper does not limit itself to the specific narratives surrounding Bitcoin set forth by Shiller, the aim is to do as Shiller suggests and incorporate narrative economic theory into existing methods, specifically by applying the concept of narrative contagion to an empirical sentiment analysis study.

Despite the parabolic growth of Bitcoin and other cryptocurrencies in the past year, many people still consider them to have little to no inherent value. Bitcoin price has been likened to tulip mania (Taskinsoy, 2019), where a vicious circle occurs after price increases and causes new investors to participate. Others suggest that technological developments in cryptocurrencies and blockchain are causing them to gain higher penetration among different industries and people (Akar & Akar, 2020). While this study does not tackle the complex question of what value, if any, cryptocurrencies possess, it does attempt to determine the relationship between social media narrative contagion and cryptocurrency prices. This paper contributes to a growing body of work on cryptocurrency sentiment analysis through its novel application of topic modelling techniques to Twitter data. The first step in doing so is identifying the leading narratives in cryptocurrency social media spheres. A set of 353.171 tweets on the topic of cryptocurrency are scraped and prepared for analysis. A machine-learning algorithm called latent Dirichlet allocation (LDA) is used to uncover the narratives present in this set of text data in a quantifiable manner. Through this method, four leading narratives are found over the study period of 2020–2021, referred to herein as Decentralised Finance (DeFi), Non-Fungible Tokens (NFTs), Gaming and Memecoins. Further information on the uncovering of these narratives is

provided in Subsection 3.2. After categorising tweets according to these four narratives, grouping the data into a time series, and gathering price data on two of the largest cryptocurrencies, empirical analysis was performed. Through Granger causality tests it was revealed that tweet engagements are in a complex two-way predictive relationship with price returns and trading volumes. Out of the uncovered narratives memecoin engagement was found to have the most predictive power over cryptocurrency price returns. OLS regressions reveal a strong relationship between S&P500 stock market index returns and cryptocurrency prices.

The remainder of this paper proceeds as follows. Section 2 presents a review of the related literature that leads to the development of hypotheses to be tested. Section 3 describes the data collection process, the machine learning algorithm used to find narratives, and the empirical framework used to study the data. The results are presented in Section 4. Limitations are discussed in Section 5, and conclusions are presented in Section 6.

# 2 Literature Review

## 2.1 The evolution of cryptocurrency

Our common understanding of what a cryptocurrency is and can do has broadened in recent years. What started with the invention by the pseudonymous Satoshi Nakamoto (Nakamoto, 2008) of a peer-to-peer electronic cash system which solves the double-spending problem through the use of cryptography has grown into an entire asset class. Until the cryptocurrency bull market of 2017, Bitcoin always made up the overwhelming majority of the overall cryptocurrency market cap, with Bitcoin's so-called 'dominance' of the total market capitalisation never dropping below 75%[1]. It is thus no surprise that early studies on cryptocurrency prices were all strictly concerned with the price of Bitcoin (Greaves & Au, 2015; Madan et al., 2015; McNally, 2016). The rise of various altcoins (an abbreviation of 'alternative coin') with different consensus mechanisms and features has drastically changed this trend. As

[1] https://coinmarketcap.com/charts/

of June 2022, Bitcoin's market capitalisation dominance sat at 43%, a steep drop from a few years prior.

The introduction of smart contracts[2] by Ethereum (the cryptocurrency which currently holds the position of second-largest market cap) has added a great number of potential use cases for blockchain technology. Simply put, smart contracts are programs which run on the blockchain. They are addresses which are not controlled by private keys but have code that determines what happens when transactions are sent to them. All four narratives uncovered by the LDA algorithm for this study are built upon this smart contract technology. Decentralised finance applications, blockchain gaming and NFTs use smart contract technology, and the vast majority of so-called 'memecoins' are ERC-20 coins built upon the Ethereum blockchain. For that reason, as well as examining the relationships between our uncovered narratives and the US dollar-denominated cryptocurrency prices, in this study, the relationship between narratives and cryptocurrency prices is also analysed through the ETH/BTC trading pair. The ETH/BTC pair could be seen as a representation of how the market values the newer evolving smart contract technology brought forth by Ethereum versus the censorship-resistant store of value which Bitcoin aspires to be.

### 2.2 Price prediction through sentiment analysis

In an informationally efficient market with rational participants, as described by Fama (1970), prices should fully reflect the available information, and thus price predictions based on sentiment should not be possible. Under the Efficient Market Hypothesis assumptions, prices determined by unpredictable news follow a random walk pattern and are not predictable with more than 50 per cent accuracy. In practice, however, models with higher prediction accuracy have been developed in many studies (Qian & Rasheed, 2007). Various studies have found that sentiment analysis performed on text data scraped from Twitter is useful in the prediction of financial market instruments, which illustrates that not all of the assumptions for an

[2] https://ethereum.org/en/developers/docs/smart-contracts/

informationally efficient market are being met (Bollen et al., 2011; Pagolu et al., 2016; Rao & Srivastava, 2012).

There is consensus among researchers that cryptocurrency markets are generally inefficient. Al-Yahyaee et al. (2018) find cryptocurrency markets to be less efficient than gold, stock and currency markets, and Grobys et al. (2020) note that there are various arbitrage opportunities and simple rules-based trading strategies which can generate excess returns in cryptocurrency markets. While some may argue that this inefficiency is a result of the infancy of the asset class and should improve over time, Zhang et al. (2018) found that inefficiencies grew in the period of the 2017–2018 bull market compared to earlier years, which indicates that this might not necessarily be the case. As a result of these large inefficiencies in cryptocurrency markets, it is no surprise that so many studies have been able to predict cryptocurrency prices successfully using sentiment analysis. Colianni et al. (2015) were among the first to develop advantageous Bitcoin trading strategies using text data scraped from Twitter. More recently, Abraham et al. (2018), Valencia et al. (2019) and Kraaijeveld & De Smedt (2020) were able to predict prices accurately for various other cryptocurrencies using sentiment analysis. This paper contributes to this collection of literature by implementing a narrative-based topic modelling approach in addition to sentiment analysis, which is novel in its application to cryptocurrency.

### 2.3 Social media topic modelling

Topic modelling is a machine learning technique which can be used to discover abstract topics in a set of documents. There is a growing body of literature on the use of topic models to uncover narratives in social media data, with the recent COVID-19 pandemic proving to be a popular case study (Sha et al., 2020). Research appears lacking on the application of such techniques to cryptocurrency-related topics on social media. There is, however, a fairly rich collection of literature on topic modelling based sentiment analysis in traditional financial markets, starting initially with studies applying such techniques to data acquired from various blog sources such as yahoo finance message boards (Nguyen & Shirai, 2015; O'Hare et al., 2009). Recently, much of the discussion on stock market investments and trading has moved to larger

social media platforms such as Twitter and Reddit. Si et al. (2013) studied Twitter data for stock prediction and found that their topic-based approach to sentiment analysis performed better than existing state-of-the-art non-topic methods. These findings support the approach taken in this study, which is to apply such methods to cryptocurrency markets, which have not yet been studied using Twitter data. Azqueta-Gavaldón (2020) researched the relationship between narratives and cryptocurrency prices and found a strong bi-directional causal relationship between cryptocurrency prices and various narratives using machine learning and a complex dynamic system. However, their study analysed data from traditional mainstream media sources rather than social media. Aside from traditional media, the popular forum BitcoinTalk[3] has also been examined using topic modelling techniques. Linton et al. (2017) walk through the steps for constructing such a model but perform no further empirical study using topics scraped from the forum.

The micro-blogging social media website Twitter[4] is currently home to the most discussion on the topic of cryptocurrency. Whereas the earlier mentioned forum BitcoinTalk held this number one spot for the earliest years of Bitcoin's existence, it has since lost this position and seen a large decrease in message volume in recent years, while the overall trend of tweets on the topic of cryptocurrency has been increasing. With regard to predictive power, social media such as Twitter have been found to be superior to traditional media outlets because they are such a rich source of data on the emotions of market participants, with investors frequently expressing their sentiments (Kraaijeveld & De Smedt, 2020). Our study thus takes a novel approach and makes a valuable contribution to the existing literature by applying topic modelling techniques to cryptocurrency-related text data gathered from Twitter and empirically studying the relation to cryptocurrency prices and volumes.

[3] https://bitcointalk.org/

[4] https://twitter.com/

## 2.4 Hypothesis development

To summarise the preceding subsections, we find that sentiment analysis has been used to establish models which can predict stock and cryptocurrency prices accurately. Both tweet sentiment and volume have been found to be associated with cryptocurrency prices and trading volume. However, on the matter of whether sentiment can predict prices, findings are mixed, with some studies finding evidence in favour (Kraaijeveld & De Smedt, 2020), while other studies have found that Twitter sentiment is dictated by trading volumes and holds no predictive power over price (Kaminski, 2014). Additionally, topic-based approaches to sentiment analysis have been found to perform better than non-topic methods (Si et al., 2013). As explained in more detail in Section 3 on the research method, empirical tests are performed in this study using overall tweet engagement and sentiment metrics and individually for the uncovered narratives. With respect to the overall effects of tweet engagement and bullishness, without taking the specific narratives into account, we expect to find results in line with the most recent non-topic-based sentiment analysis studies. Therefore the following two hypotheses are formulated:

*H1: Tweet engagement is a predictor of Bitcoin and Ethereum price returns and volume*
*H2: Tweet bullishness is a predictor of Bitcoin and Ethereum price returns and volume*

Furthermore, the argument is brought forward that the ETH/BTC trading pair is able to act as a proxy for the market's valuation of newer smart contract technology versus the censorship-resistant store of value which is Bitcoin. Since all four of the narratives uncovered through the LDA topic modelling algorithm appear to be related to some degree to smart contracts and Ethereum blockchain technology, the expectation is that higher engagement and bullishness in tweets specific to each narrative will be associated with an increase in the ETH/BTC trading pair. This leads to the formulation of the third and fourth hypotheses:

*H3: Narrative-specific tweet engagement is a predictor of ETH/BTC price returns and volume*
*H4: Narrative-specific tweet bullishness is a predictor of ETH/BTC price returns and volume*

# 3 Methodology

## 3.1 Data collection

Data were gathered from two sources: text data consisting of tweets were scraped from Twitter, and financial data consisting of cryptocurrency prices and trade volumes were obtained through the CoinMarketCap API[5]. An overview containing summary statistics of the collected data can be found in Table 1**Error! Not a valid bookmark self-reference.**.

TABLE 1. SUMMARY STATISTICS

A. Financial data[a]

| | *Mean* | *Median* | *Max* | *Min* | *N* |
|---|---|---|---|---|---|
| *Daily close BTC* | 29.251,81 | 29.001,72 | 67.556,83 | 4970,79 | 731 |
| *Daily close ETH* | 1541,26 | 738,80 | 4812,09 | 110,61 | 731 |
| *Daily volume BTC (billion)* | 40.08 | 36.15 | 350.97 | 12.25 | |
| *Daily volume ETH (billion)* | 20.75 | 18.20 | 84.48 | 5.11 | |

B. Twitter text data[b]

| | *Mean* | *Median* | *Max* | *Min* | *N* |
|---|---|---|---|---|---|
| *Overall tweet volume* | 483 | 282 | 2862 | 76 | 353.171 |
| *Overall tweet engagement* | 239.639 | 91.772 | 1.981.196 | 11.124 | 175.175.854 |
| *Overall weighted bullishness[c]* | 1.62 | 1.63 | 3.50 | 0.46 | |
| *NFT engagement* | 17566 | 1934 | 219.256 | 0 | 12.840.932 |
| *DeFi engagement* | 14225 | 6610 | 211.055 | 109 | 10.399.197 |
| *Gaming engagement* | 9254 | 3071 | 183.158 | 0 | 6.764.859 |
| *Memecoin engagement* | 14850 | 1445 | 661.855 | 0 | 10.855.330 |
| *NFT weighted bullishness[c]* | 2.58 | 2.11 | 9.58 | -6.29 | |
| *DeFi weighted bullishness[c]* | 2.93 | 2.40 | 10.27 | -1.98 | |
| *Gaming weighted bullishness[c]* | 3.37 | 2.59 | 10.23 | -5.87 | |
| *Memecoin weighted bullishness[c]* | 3.20 | 2.86 | 10.51 | -6.79 | |

[a] Prices and volumes are presented in USD denomination [b] Twitter data is presented per day [c] See section 3.3 for further explanation of the bullishness variable

### *3.1.1 Twitter data*

[5] https://coinmarketcap.com/api/

The tweets were obtained using the Python module snscrape[6], which is a library that allows tweets to be scraped through Twitter's API without any restrictions or request limits. In addition to the contents of tweets, the numbers of likes, retweets and quote retweets were also gathered for each tweet. The sample contains all tweets mentioning the phrase 'crypto' with more than 50 likes or 10 retweets over the period 2020–2021. The minimum like/retweet thresholds were chosen to keep the sample size manageable. The period 2020–2021 was primarily chosen because it was a very eventful period in the history of cryptocurrency, in which many narratives were born and popularised. The sample size was also a factor in the choice to not examine a broader period.

### *3.1.2 Financial data*

Closing price and daily volume were collected on a daily interval for Bitcoin ($BTC) and Ethereum ($ETH). CoinMarketCap was chosen as the source for this data because the cryptocurrency prices provided by CoinMarketCap are a weighted average of a large selection of exchanges, making it a more reliable price oracle than any single exchange. The scraped daily closing prices are then used to determine daily price returns ($P_R$). Price returns are used as an independent variable for this study as they are less likely to exhibit autocorrelation issues than a time series of regular prices (Kraaijeveld & De Smedt, 2020). The ETH/BTC ratio is determined by dividing the daily closing price of Ethereum by that of Bitcoin. Similarly, daily price returns are used for ETH/BTC. Figure 1 presents a plot of Bitcoin price and overall tweet engagement. Both series can be seen to rise over the sample period, with large peaks in engagement seemingly coinciding with sharp increases or drops in price.

[6] https://github.com/JustAnotherArchivist/snscrape

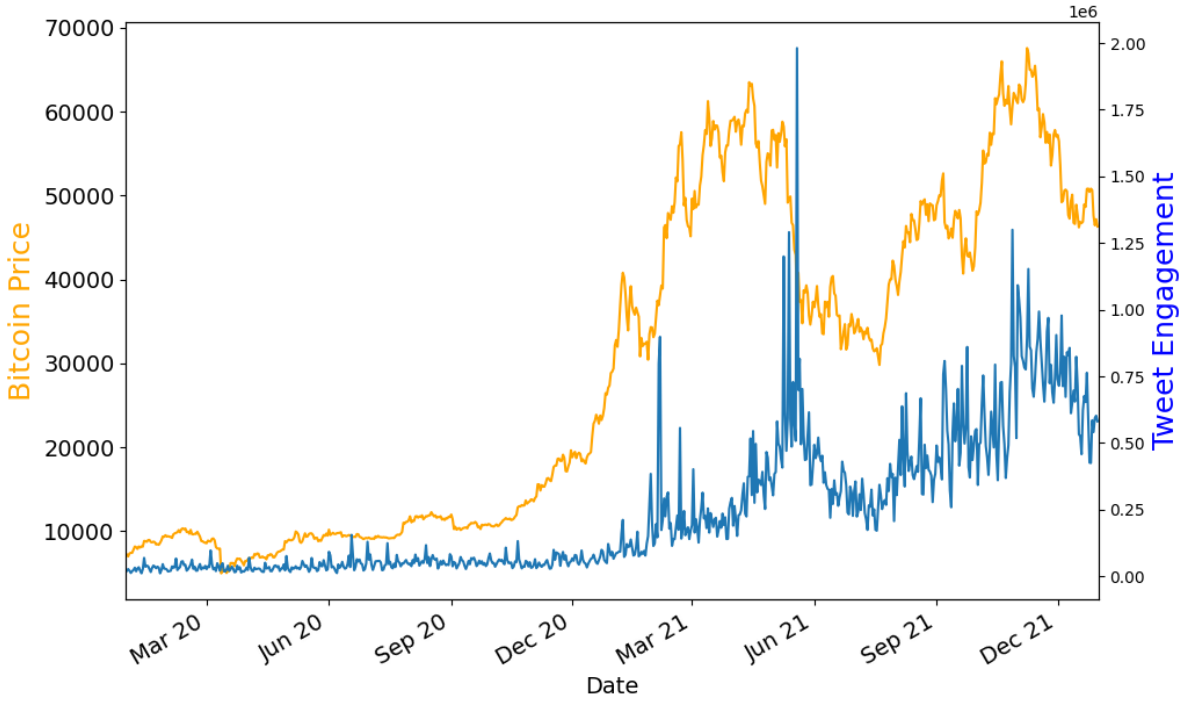


FIGURE 1: BITCOIN PRICE AND OVERALL TWEET ENGAGEMENT

## 3.2 Topic modelling

### *3.2.1 Latent Dirichlet Allocation*

A machine learning algorithm called latent Dirichlet allocation (LDA) was used to determine the leading narratives concerning cryptocurrency over the period 2020–2021. LDA is a generative probabilistic model of a corpus. Each document is represented as a random mixture over latent topics, while topics are characterised by a distribution over words (Blei et al., 2003). These distributions are obtained by maximising the probability of each word appearing in each article (tweet in our case) given the total number of topics (Azqueta-Gavaldón, 2020). Each topic found through the use of this algorithm thus comes in the form of a collection of words which are frequently mentioned together in all of the tweets. It is then up to our own interpretation to label the topics appropriately where possible.

### *3.2.2 Data pre-processing*

Before the LDA algorithm was run, the text data scraped from Twitter were extensively filtered and cleaned to remove useless data. The data were initially filtered by removing stop

words (such as 'a', 'the', 'and', 'in'). This was done using the list of stop words provided in the Python nltk.corpus package[7]. All characters were converted to lowercase, and URLs and symbols such as '@' were removed from the tweet content. The next step was to lemmatise the words and convert them into their root forms using the natural language processing tool Spacy[8]. To give an example of the transformations made in this process, the string 'Python is the greatest language in the world' would be converted to the following output: ['python', 'be', 'the', 'great', 'language', 'in', 'the', 'world']. Finally, rather than modelling the topics based on the occurrence of single words, bigrams and trigrams were also included using tools provided by the python library Gensim[9].

### *3.2.3 Narrative identification*

After the pre-processing had concluded and the algorithm had run, ten topics were revealed. Four of the ten topics brought forward by the model were identified as topics which clearly represented narratives in cryptocurrency. These topics are presented in Figure 4-7 of the Appendix. Each topic identified as a narrative was labelled accordingly. A summary of the labelled topics is displayed in **Error! Not a valid bookmark self-reference.**.

TABLE 2. TOP WORDS IN CRYPTOCURRENCY NARRATIVES

| *Topic label* | *% of tokens* | *Top words uncovered by LDA* | *Final words list after additional manual search* |
|---|---|---|---|
| Decentralised Finance | 13.7 | blockchain, exchange, new, defi, user, world, wallet, project, build, platform | defi, wallet, dex, yield, yield_farming, yield_farm, collateral, lend, lending, decentralized_exchange |
| Gaming | 7.4 | community, game, blockchain, new, thank, team, video, play, join, gaming | gaming, play_to_earn, playtoearn, game, video_game, play, gameplay |
| Memecoin | 6 | dogecoin, doge, crypto, talk, live, cryptocurrency, defi , blockchain, safemoon, interview | doge, dogecoin, dogearmy, doge_army, safemoon, shib, shiba, shiba_inu, shibainu, shiba_army, shibarmy, memecoin |
| NFT | 5 | nft, art, nftcollector, nftcommunity, cryptoart, artist, nftart, drop, new, blockchain | nft, art, artist, nftart, nftartist, nft_artist, cryptoart, nftcollector, nftcommunity, mint |

[7] https://www.nltk.org/api/nltk.corpus.html

[8] https://spacy.io/

[9] https://radimrehurek.com/gensim/

The first narrative that was identified is labelled as Decentralised Finance. What is often referred to as DeFi by crypto-natives can be broadly defined as anything using blockchain technology to provide financial products or services traditionally provided by institutions such as banks. This includes protocols which offer borrowing/lending services, as well as decentralised exchanges (DEXes). The first topic was identified by the presence of the words 'defi', 'wallet', 'platform', 'blockchain' and 'exchange', all of which appear to have a connection to Decentralised Finance. The second narrative is labelled Gaming. Blockchain gaming projects with 'play to earn' incentive structures have increased considerably in popularity in the past few years. The words 'game', 'video', 'play', 'gaming' and 'blockchain' all seem to indicate that that is what this topic concerns. The third narrative is labelled Memecoin. The phenomenon of so-called memecoins also saw a tremendous ascent, with dogecoin ($DOGE) reaching a peak market capitalisation of $89 billion and the Ethereum-based equivalent Shiba Inu ($SHIB) reaching $41 billion[10]. This topic was labelled because of the presence of the words 'doge', 'dogecoin' and 'safemoon'. The fourth and final identified narrative is labelled as NFT. NFTs, or non-fungible tokens, are cryptographic assets on a blockchain used to represent ownership of unique items. The final narrative appears to be the clearest match yet with this topic, with the words 'nft', 'art', 'nftcollector', 'nftcommunity', 'cryptoart', 'artist' and 'nftart' all having the theme in common.

After all of the topics were successfully identified, a manual search through the list of the most frequently appearing words was performed to find other words that could be considered relevant to each topic. These were included in the final lists of words for each narrative, which was used to distinguish which tweets mentioned one or more narratives and, at a later stage, to construct independent variables for each narrative. The engagements for each tweet determined to be a part of one of the four narratives are plotted in Figure 2. The figure shows us that, while eventually becoming the two most dominant narratives, the Memecoin and NFT narratives only truly emerged somewhere early in 2021. The gaming and NFT narratives also saw a great rise in engagement through 2021 but were already present in 2020. In the case of

[10] https://coinmarketcap.com/

DeFi, this is in line with expectations since the summer of 2020 is remembered by many crypto natives as the period during which the concept of Decentralised Finance started taking off, colloquially referred to as 'DeFi summer'.

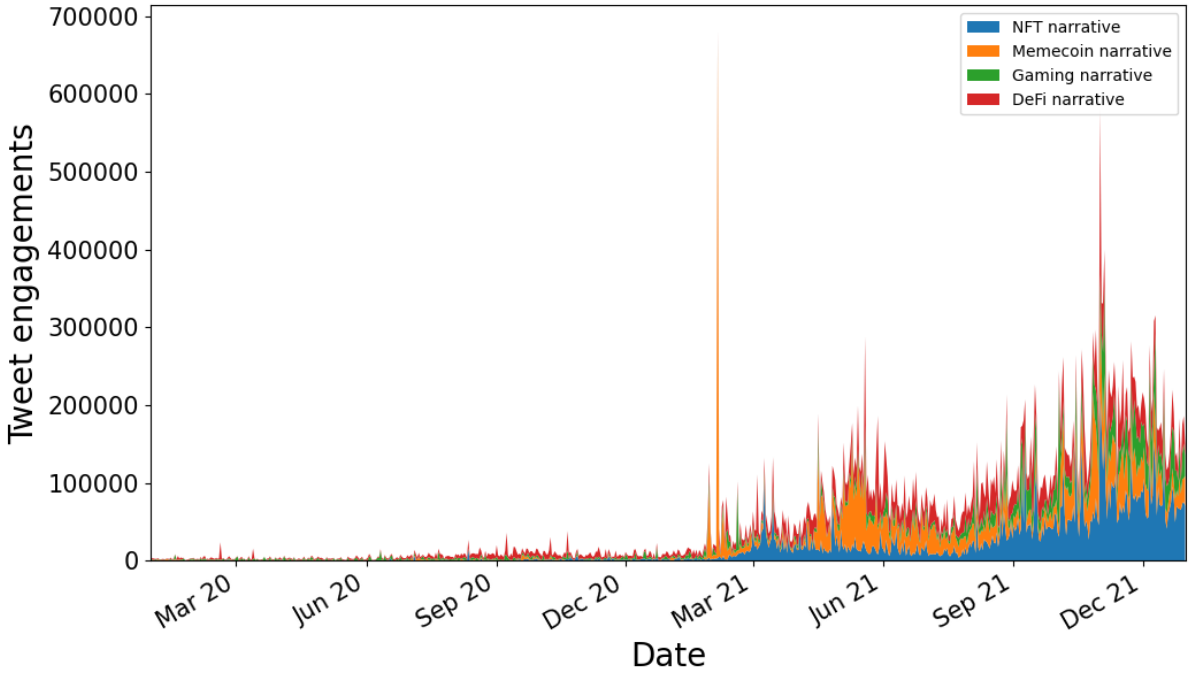


FIGURE 2. TWEET ENGAGEMENTS PER NARRATIVE

### 3.3 Sentiment analysis

The sentiment analysis was performed using the Valence Aware Dictionary and Sentiment Reasoner (VADER) model. VADER is a rule-based model for general sentiment analysis developed by Hutto & Gilbert (2014) which classifies a tweet with a normalised weighted composite compound score ranging from -1 to 1 based on the level of positivity or negativity of its contents. After a score was attributed to each tweet, the tweet was assigned to one of three categories based on its score. Scores $\geq 0.05$ were categorised as positive, $\leq -0.05$ were

categorised as negative, and everything in between was considered neutral. Following the approach set out by Antweiler & Frank (2004), the sentiment measures were aggregated into a single measure of 'bullishness', defined as

$$B \equiv \ln[\frac{1 + M^{BUY}}{1 + M^{SELL}}]$$

where $M^{BUY}$ and $M^{SELL}$ represent buy and sell messages, respectively, which in our case are tweets that were classified as being positive or negative. Figure 3 presents a scatterplot the bullishness per narrative. The figure illustrates that spreads got much tighter from early 2021 onwards, which coincides with the period during which tweet engagements increased dramatically, as seen in Figure 2. Prior to this period certain topics such as NFTs and memecoins had not yet broken into the mainstream, which is seen to lead to days with very high bullishness values, as well as days with a value of zero where there were no tweets collected for the respective narrative.

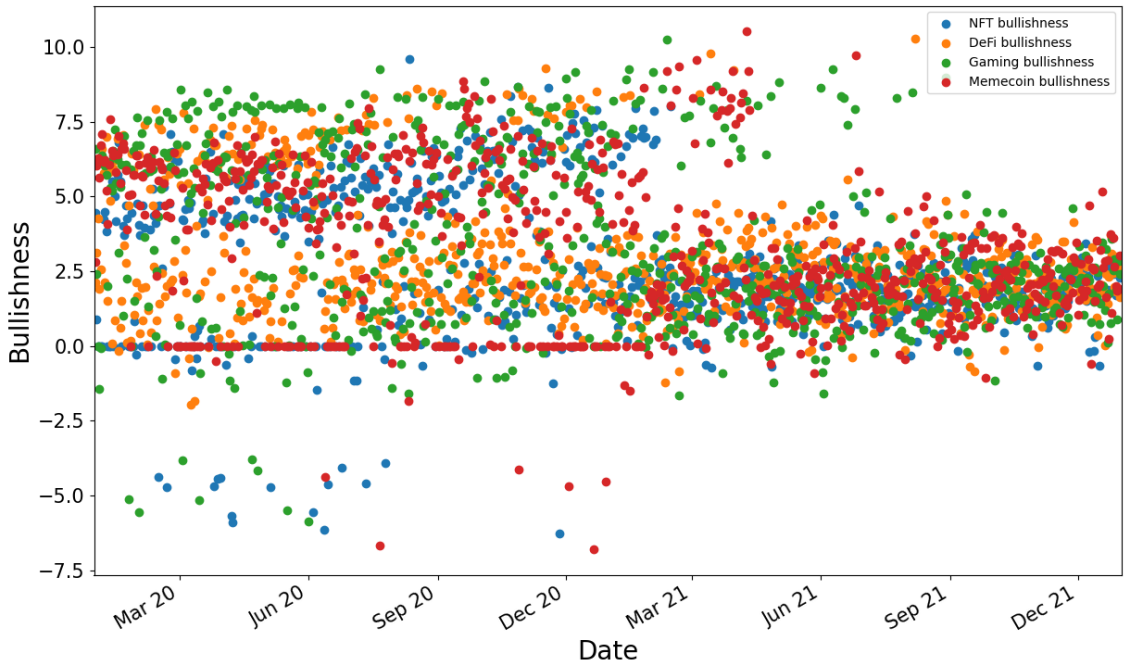


FIGURE 3. BULLISHNESS PER NARRATIVE

### 3.4 Metrics and variables

This study explores two dependent variables: price returns $P_r$ and daily trading volume $V$. While price returns are the primary focus of this study, daily trading volumes provide an additional metric by which the impact of social media narratives can be measured, as prior studies have found a statistically significant predictive effect of Twitter sentiment on daily trading volume (Kraaijeveld & De Smedt, 2020). When it comes to Twitter data, rather than opting for a strictly volume-based approach such as those employed by many of the aforementioned cryptocurrency sentiment analysis studies, where total tweet volume is used as an independent variable and the engagement of the tweets is not taken into consideration, this study aims to construct a metric to more accurately assess the overall influence of individuals tweets. Wang et al. (2017) argue that higher numbers of upvotes on social media posts signal that they are of higher quality. This higher perceived quality could lead to a tweet having a stronger influence on the proliferation of narratives and the effect these narratives could have on prices and trading volumes. Aside from impacting the perceived quality of a tweet, higher engagement in the form of likes and retweets is also a proxy for how many times a tweet has been viewed. Since the latter is not something that Twitter publicly discloses per tweet (only the owner of a Twitter account can view the impression metrics of a given tweet), it was not possible to be scraped directly. Both arguments strongly support the inclusion of tweet engagement as a part of the influence metric. Following the approach set out by El Alaoui et al. (2018), tweet engagements (likes + retweets + quote retweets) are used as an independent variable rather than tweet volume.

Overall engagement levels $E_{total}$ and a measure of weighted bullishness: $B^W_{total}$ serve as independent variables in a first set of regressions. The weighted bullishness measure is weighted by engagement rather than tweet volume, which means that it is determined using the following equation: $B^W \equiv \ln[\frac{1+E^{POS}}{1+E^{NEG}}]$. Aside from total tweet volume and bullishness, these metrics can also be broken down into separate variables representing tweets about the narratives which were identified earlier: $E_{DeFi}$, $B^W_{DeFi}$, $E_{Gaming}$, $B^W_{Gaming}$, $E_{Memecoin}$, $B^W_{Memecoin}$,

$E_{NFT}$ and $B_{NFT}^{W}$. These separate metrics serve as independent variables in a second set of regressions. Various control variables are included in the regression models, the first of which is daily S&P500 price returns ($P_r^{SP500}$), for which the data is sourced from Yahoo Finance[11]. Interest rates ($r$) are also taken into account through the inclusion of the 3-month treasury bill rate, obtained from the federal reserve's website for economic data[12]. Lastly, a time control ($\tau$) is included in the equation in the form of an incrementally increasing variable assigned to each month in the two-year period.

## 3.5 Regression models and robustness

### *3.5.1 OLS regression*

The first hypothesis is initially tested using a multivariate linear regression model, with separate regression equations adopted for analysis of the five separate dependent variables. These dependent variables are price returns for Bitcoin, Ethereum and the ETH/BTC pair, as well as Bitcoin and Ethereum trading volumes. Lastly, each model also includes the error term ($u_i$). Equations take the following shape, where DV is one of the respective dependent variables ($P_r^{BTC}$, $P_r^{ETH}$, $P_r^{ETH/BTC}$, $V^{BTC}$ and $V^{ETH}$):

$$DV = \beta_0 + \beta_1 E_{total} + \beta_2 B_{total}^{W} + \beta_3 P_r^{SP500} + \beta_4 r + \beta_5 \tau + u_i$$

For the next set of OLS regressions, all of the narrative-level independent variables were included, rather than the variables for total engagement and bullishness. As such, the equations were again estimated for each of the five dependent variables as follows:

$$DV = \beta_0 + \beta_1 E_{NFT} + \beta_2 B_{NFT}^{W} + \beta_3 E_{DeFi} + \beta_4 B_{DeFi}^{W} + \beta_5 E_{Gaming} + \beta_6 B_{Gaming}^{W} + \beta_7 E_{Memecoin} + \beta_8 B_{Memecoin}^{W} + \beta_9 P_r^{SP500} + \beta_{10} r + \beta_{11} \tau + e_t$$

For the purpose of robustness, these sets of regressions were also performed using regular price variables rather than price returns.

### *3.5.2 Granger causality*

[11] https://finance.yahoo.com/quote/%5EGSPC/history/

[12] https://fred.stlouisfed.org/series/DTB3#

Aside from exploring whether the above-named independent variables are associated with price and trading volume, we also wish to examine whether it is social media metrics that drive prices and volume or vice versa. A variable, x, is defined as having Granger causality on variable y if the given previous information on y, as well as the past values of x, enable forecasting of the current value of y (Luu Duc Huynh, 2019). In our case, we are interested in whether the independent variables for tweet engagement and bullishness Granger-cause the dependent variables price returns and trading volume or vice versa. The first step in testing for Granger causality is to describe the vector autoregressive models (VAR). Based on the analysis of lag-order selection statistics, a lag order of four is deemed appropriate. The first set of Granger-causality tests run in this study are bivariate. With five dependent variables ($P_r^{BTC}$, $P_r^{ETH}$, $P_r^{ETH/BTC}$, $V_{trad}^{BTC}$ and $V_{trad}^{ETH}$) and ten independent variables ($E_{total}$, $B_{total}^{W}$, $E_{DeFi}$, $B_{DeFi}^{W}$, $E_{Gaming}$, $B_{Gaming}^{W}$, $E_{Memecoin}$, $B_{Memecoin}^{W}$, $E_{NFT}$ and $B_{NFT}^{W}$) that results in fifty separate sets of two equations. As an illustration of the form they take, the equations for the VAR model of the variables $P_r^{BTC}$ and $E_{total}$ are displayed below:

$$P_r^{BTC}{}_t = \beta_{10} + \beta_{11} P_r^{BTC}{}_{t-1} + \cdots + \beta_{14} P_r^{BTC}{}_{t-4} + \gamma_{11} E_{total_{t-1}} + \cdots + \gamma_{14} E_{total_{t-4}} + u_{1t}$$

$$E_{total_t} = \beta_{20} + \beta_{11} E_{total_{t-1}} + \cdots + \beta_{14} E_{total_{t-4}} + \gamma_{11} P_r^{BTC}{}_{t-1} + \cdots + \gamma_{14} P_r^{BTC}{}_{t-4} + u_{2t}$$

Next, in addition to bivariate tests, a set of multivariate Granger causality tests was performed, including all narrative-specific independent variables, as well as the controls used in the OLS model. The classical estimation of Granger causality requires stationarity of signals (Hesse et al., 2003). Augmented Dickey-Fuller tests are used to test whether our variables meet this criterium. Price returns, trading volumes, tweet engagement and bullishness variables are all found to be stationary.

# 4 Results

### 4.1.1 *Granger causality tests*

The results of 50 separate bivariate Granger causality tests, one combination for each of the five dependent variables and ten independent variables, are presented in Table 3. Granger

causality is found to be present for total engagement ( $E_{total}$ ) on Bitcoin, Ethereum and ETH/BTC price returns. At the same time, Ethereum and ETH/BTC price returns are found to Granger-cause tweet engagement. This indicates that while tweet engagement is found to have predictive power over price returns, it is evidently a complex relationship, as it is also simultaneously influenced by them. Thus far, support has been found for the first hypothesis, while the second hypothesis is unsupported by the findings. Memecoin engagement is another variable found to Granger-cause multiple price returns, namely that of Bitcoin and Ethereum, while again itself being Granger-caused by the Ethereum and ETH/BTC price. Memecoin engagement is unique in that it has the same two-way relationship with Ethereum trading volume, which is a variable over which no other independent variable appears to have predictive power based on the bivariate tests. The bivariate tests find no support for the third and fourth hypotheses.

The results of the multivariate Granger causality tests are presented in Table 6–Table 10 in the Appendix. In these tests, total engagement is also found to be significant in the prediction of all three price pair returns, while total bullishness is significant for BTC and ETH price returns. The multivariate tests show no evidence of an inverse relationship with regard to total engagement and show only the ETH price returns Granger-causing total bullishness. These findings suggest a degree of congruence in bivariate and multivariate Granger causality tests. These findings support the first hypothesis but do not support the second hypothesis. Finally, memecoin engagement is found to have a significant two-way predictive relationship with both BTC and ETH volume and with ETH/BTC price returns. These findings provide some support for the third hypothesis. DeFi engagement is also found to have a two-way Granger-causal relationship with BTC and ETH price returns, as well as with ETH volume. The fourth hypothesis does not find any convincing support. Overall, tweet engagement levels appear to hold stronger predictive power over prices than bullishness metrics. While tweet engagement is not a metric that has been heavily researched in prior studies, the bullishness metric has been previously researched and our findings of weak predictive power go against the findings in most prior literature. This is likely due to differences in the sample period or sentiment analysis methodology employed.

### *4.1.2 OLS regressions*

The results of the OLS regressions are presented in Table 4. Each column represents a different specification, with columns (1) through (5) presenting the results of the first set of regressions, where total engagement $E_{total}$ and bullishness ($B^{W}_{total}$) were used as independent variables. Columns (6) through (10) present the results of the second set of regressions, which included the narrative-specific independent variables. Total engagement is found to be significant in the first two specifications and, interestingly, has a negative coefficient. These findings stand in contrast to what some might expect, as they indicate that higher engagement is associated with lower BTC and ETH prices. Total engagement is also significant in the specifications with BTC and ETH trading volumes as dependent variables and does exhibit a positive coefficient. The total bullishness variable is only found to be significant in the specifications with BTC and ETH trading volume as dependent variables, and here again, the coefficient is found to be negative.

With respect to the narrative-specific variables, the results of the OLS regressions do not support the second hypothesis. None of the narrative-specific variables is significant in the specification with ETH/BTC price returns as the dependent variable. We do find various narrative-specific variables which are significant in the explanation of BTC and ETH trading volume. The same is true for the time control, which is not surprising considering the growth in trading volumes over our sample period. One interesting result is that memecoin engagement is significant at the 1% level for both the BTC and ETH trading volume specifications and has a positive coefficient, while the same is true for NFT engagement at the 10% and 1% level with a negative coefficient. While it is an unexpected result to see different coefficients for these two variables here, looking back at Figure 2, which displays tweet engagements per narrative, we do see some difference in the periods during which NFT and memecoin engagements rose, which does indicate that differences in coefficients could to some degree have been foreseen.

The adjusted R-squared values for all of the specifications are relatively low, with some specifications even reaching low single digits. This was to be expected, considering the

complexity involved in predicting price returns and trading volumes, as well as the fact that price returns as dependent variables are stationary series. The results show that a larger part of the variance is explained for Bitcoin and Ethereum price returns than for ETH/BTC returns and that the variance in Ethereum trading volume is best explained by the independent variables out of all the specifications. While the adjusted R-squared values found through the regressions using price levels rather than price returns are significantly higher, as shown in Table 5 in the Appendix, they also come paired with multiple highly significant control variables in almost every specification. Because the control variables capture such a large portion of the explanatory power of these models, we will mainly focus on the specifications using price returns (presented in Table 4) to derive insights.

The control variable S&P500 returns is found to be significant at the 1% level with a positive coefficient in all of the specifications with price returns as dependent variables. The fact that this relationship was so strong and persistent through all specifications makes it clear that S&P500 returns are a key factor in determining cryptocurrency prices. These findings raise the question of whether Bitcoin and other cryptocurrencies are still a suitable form of diversification from stock markets, which has been found to be the case in prior research (Gkillas & Longin, 2019). One thing to note is that there are only 505 observations, despite the time series sample spanning 731 days. The reason for this is that, unlike returns on cryptocurrencies, the markets for which are open 24/7, S&P500 returns can only be calculated for days when the stock markets are open.

TABLE 3. BIVARIATE GRANGER-CAUSALITY TEST RESULTS

| | $Relation$ | $E_{total}$ | $B^W_{total}$ | $E_{NFT}$ | $B^W_{NFT}$ | $E_{DeFi}$ | $B^W_{DeFi}$ | $E_{Gaming}$ | $B^W_{Gaming}$ | $E_{memecoin}$ | $B^W_{memecoin}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_r^{btc}$ | ← | **0.021**** | 0.491 | 0.245 | 0.754 | 0.221 | 0.951 | 0.624 | 0.800 | **0.004***** | 0.581 |
| $P_r^{btc}$ | → | 0.621 | 0.474 | 0.574 | **0.042**** | 0.436 | 0.540 | 0.587 | **0.622** | 0.500 | 0.490 |
| $P_r^{eth}$ | ← | **0.001***** | 0.207 | 0.193 | 0.506 | 0.390 | 0.781 | 0.604 | 0.948 | **0.046**** | 0.890 |
| $P_r^{eth}$ | → | **0.092*** | 0.181 | 0.141 | 0.202 | 0.229 | 0.579 | 0.277 | **0.053*** | **0.047**** | 0.582 |
| $P_r^{eth/btc}$ | ← | **0.018**** | 0.396 | 0.835 | 0.263 | 0.877 | 0.736 | 0.246 | 0.553 | 0.241 | 0.749 |
| $P_r^{eth/btc}$ | → | **0.044**** | 0.211 | 0.175 | 0.526 | 0.625 | 0.449 | 0.284 | **0.012**** | **0.044**** | 0.393 |
| $Vol^{btc}$ | ← | 0.679 | 0.137 | 0.885 | 0.923 | 0.397 | 0.153 | 0.840 | 0.135 | 0.236 | 0.946 |
| $Vol^{btc}$ | → | 0.534 | . | 0.992 | . | 0.697 | . | 0.839 | . | 0.705 | . |
| $Vol^{eth}$ | ← | 0.714 | 0.647 | 0.403 | 0.650 | 0.818 | 0.242 | 0.207 | 0.139 | **0.081*** | 0.610 |
| $Vol^{eth}$ | → | 0.686 | . | 0.641 | . | **0.027**** | . | 0.310 | . | **0.000***** | . |

*Notes:* P-values are reported. Test performed with four lags. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively.

TABLE 4. OLS REGRESSION RESULTS: PRICE RETURNS

| | (1) $P_r^{BTC}$ | (2) $P_r^{ETH}$ | (3) $P_r^{ETH/BTC}$ | (4) $V^{BTC}$ | (5) $V^{ETH}$ | (6) $P_r^{BTC}$ | (7) $P_r^{ETH}$ | (8) $P_r^{ETH/BTC}$ | (9) $V^{BTC}$ | (10) $V^{ETH}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Total engagement | **-2.28e-08*** | **-3.01e-08*** | -1.08e-08 | **11925.95*** | **14583.11***** | | | | | |
| | **(1.22e-08)** | **(1.55e-08)** | (9.20e-09) | **(6383.41)** | **(3005.64)** | | | | | |
| Total bullishness | .004813 | .005718 | .001256 | **-8.28e+09***** | **-5.08e+09***** | | | | | |
| | (.003441) | (.0043692) | (.002588) | **(1.79e+09)** | **(8.45e+08)** | | | | | |
| NFT engagement | | | | | | 8.43e-08 | 6.43e-08 | -1.04e-08 | **-99531.15*** | **-130143.9***** |
| | | | | | | (1.01e-07) | (1.29e-07) | (7.63e-08) | **(53025.22)** | **(24386.47)** |
| NFT bullishness | | | | | | .000352 | .000491 | .000178 | -6.51e+08 | **-3.14e+08*** |
| | | | | | | (.000773) | (.000986) | (.000584) | (4.06e+08) | **(1.87e+08)** |
| DeFi engagement | | | | | | **-2.25e-07*** | -2.58e-07 | -5.94e-08 | 8568.205 | **67802.11**** |
| | | | | | | **(1.35e-07)** | (1.72e-07) | (1.02e-07) | (70839.51) | **(32579.33)** |
| DeFi bullishness | | | | | | .000999 | .000177 | -.000722 | -2.68e+08 | -3.24e+08 |
| | | | | | | (.000942) | (.001201) | (.000711) | (4.94e+08) | (2.27e+08) |
| Gaming engagement | | | | | | -3.22e-08 | -4.68e-08 | -2.75e-08 | **-184735.5*** | -60693.46 |
| | | | | | | (1.84e-07) | (2.35e-07) | (1.39e-07) | **(96632.33)** | (44441.53) |
| Gaming bullishness | | | | | | .000668 | .000061 | -.000533 | -4.54e+07 | -1.65e+08 |
| | | | | | | (.000703) | (.000896) | (.000531) | (3.69e+08) | (1.70e+08) |
| Memecoin engagement | | | | | | -5.21e-08 | -6.71e-08 | -1.53e-08 | **82174.08***** | **71773.33***** |
| | | | | | | (5.51e-08) | (7.02e-08) | (4.16e-08) | **(28879.72)** | **(13281.88)** |
| Memecoin bullishness | | | | | | **.001464*** | **.001674*** | .000276 | -3.28e+08 | -2.80e+08 |
| | | | | | | **(.000764)** | **(.000974)** | (.000577) | (4.01e+08) | (1.84e+08) |
| S&P500 returns | **.917046***** | **1.17217***** | **.27369***** | -4.97e+08 | -1.12e+10 | **.915429***** | **1.18254***** | **.285598***** | -2.75e+10 | -2.46e+10 |
| | **(.112782)** | **(.143217)** | **(.084824)** | (5.88e+10) | (2.77e+10) | **(.112914)** | **(.144014)** | **(.085278)** | (5.92e+10) | (2.72e+10) |
| Time control | .00032 | .0007991 | .000632 | -3.03e+08 | 1.78e+07 | .000142 | .000403 | .000349 | **6.19e+08**** | **8.13e+08***** |
| | (.000558) | (.0007084) | (.00042) | (2.91e+08) | (1.37e+08) | (.000549) | (.000700) | (.000414) | **(2.88e+08)** | **(1.32e+08)** |
| Interest rate | -.000186 | .0078284 | **.008478*** | **-6.38e+09**** | -2.32e+09 | -.002724 | .004425 | .007270 | -2.10e+09 | 1.66e+09 |
| | (.005871) | (.007455) | **(.004415)** | **(3.06e+09)** | (1.44e+09) | (.005955) | (.007595) | (.004492) | (3.12e+09) | (1.44e+09) |
| Observations | 505 | 505 | 505 | 505 | 505 | 505 | 505 | 505 | 505 | 505 |
| R-squared | 0.1372 | 0.1351 | 0.0288 | 0.0725 | 0.2507 | 0.1466 | 0.1371 | 0.0314 | 0.0722 | 0.2849 |
| Adjusted R-squared | 0.1285 | 0.1265 | 0.0191 | 0.0632 | 0.2432 | 0.1275 | 0.1178 | 0.0098 | 0.0515 | 0.2689 |

*Notes:* Coefficients are reported with standard errors in parentheses. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively.

# 5 Discussion

One limitation which could potentially affect the outcome of this study is that the sample period only spans from 2020 to 2021, which is a period during which cryptocurrencies experienced a bull market. While this period was limited to these years for practical reasons relating to the collection of Twitter data, it would be interesting to see a future study apply the methodology to a longer timeframe, especially if that timeframe included both bull and bear market periods. Another potential limitation concerns the identification of narratives. In categorising tweets into one of the identified narratives, a manual search through the list of most frequent words was performed to determine which words belong to which of the popular narratives. This approach leaves room for type 1 and type 2 errors in the classification of tweets, but there is no reason to believe that one of the two error types would be present to an excessive degree.

The presence of Twitter bot accounts is another point of discussion. While the imposed minimum numbers of likes and retweets filtered out a large proportion of bot replies, it is possible that some bot replies were still included in the sample of tweets. It is possible that some of the effect attributed to narrative-specific tweets could be a result of bot tweets rather than real human interaction, since certain spam tweets are known to include many cryptocurrency-related hashtags which could have caused them to be categorised among several of the identified narratives. An interesting avenue for future research would be to apply a rigorous filtering process which separates bot comments from humans based on various heuristics. Successfully filtering bot comments would allow them to be used as a separate independent variable in similar studies to gain insights into the predictive power of such spam. Finally, a potential limitation to the sentiment analysis part of this study is that the lexicon-based tool which was used might not have attributed appropriate sentiment values to certain specific cryptocurrency-related slang. Phrases such as ‘wagmi’ (An abbreviation for ‘we’re all going to make it’) are commonly found in tweets discussing cryptocurrency, and a sentiment analysis model would need to be manually trained to recognise all of such phrases accurately. In this area, future research could improve on the approach taken in this study.

# 6 Conclusion

This study explores concepts introduced by Robert Shiller in his work on narrative economics and incorporates them into social media sentiment analysis on the topic of cryptocurrency. Four main cryptocurrency narratives are uncovered over the period of 2020–2021 by applying a machine learning algorithm to text data acquired from the social media platform Twitter. The narratives brought forward by this topic modelling approach are Decentralised Finance, Non-Fungible Tokens, Gaming and Memecoins. The Twitter data were categorised according to these four narratives, and sentiment analysis was performed. Through empirical analysis consisting of bivariate and multivariate Granger causality tests as well as OLS regressions, it was found that Twitter engagement can be used to predict price returns for Bitcoin, Ethereum and ETH/BTC trading pairs as well as for Bitcoin and Ethereum trading volumes. There is support for the first hypothesis, which suggests that tweet engagement is a predictor of BTC and ETH price returns as well as trading volumes, while the second hypothesis concerning tweet bullishness as a predictor finds no support and is thus rejected. Despite these findings, the relationship between tweet engagements and cryptocurrency prices remains complex, as there is evidence of a two-way relationship, with price returns also Granger-causing Twitter engagement.

Of the narratives identified, the ‘memecoin’ narrative appears to hold the most predictive power over cryptocurrency prices and trading volumes. This is a narrative which saw a tremendous rise in popularity in early 2021 with the rise of dogecoin, which coincides with the period during which Elon Musk was actively tweeting about the cryptocurrency. Despite these findings, there is a lack of evidence that narrative-specific tweet engagements and bullishness are predictive of ETH/BTC price returns, which was hypothesised. Because of this lack of substantial support, the third and fourth hypotheses are rejected. Finally, the OLS regression results highlight the strong association between the S&P500 stock market index and cryptocurrency prices. This appears to fly in the face of a widely held belief that cryptocurrency can function as a form of ‘digital gold’ and a hedge against stock market prices. Another result from the OLS regressions which stands out is the negative coefficient for tweet engagement when price returns are used as dependent variable. While these findings are not intended to be

the basis of any trading strategy, when considered alongside the findings that tweet engagement is a better overall predictor of price returns than bullishness, one piece of trading wisdom which can be drawn from this would be to disregard daily sentiment swings and be wary when social media volumes increase by a great amount.

A suggestion for future research would be to study how to accurately identify tweets created by bots. Successfully doing so would subsequently allow for the effects of these bot-generated tweets on cryptocurrency price returns and trading volumes to be studied. Another interesting avenue for exploration in future studies would be the use of sentiment analysis models which are more specifically tailored to social media texts concerning financial markets and cryptocurrencies. A model trained using specific cryptocurrency-related slang could potentially more accurately judge sentiment in the messages than we were able to do in this study. Overall, the results of this study underscore the complexity of predicting cryptocurrency prices and trading volumes. We can also conclude that contagious stories do contribute to economic outcomes and that taking narrative economics into consideration may help demystify a piece of the puzzle.

# REFERENCES


Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, *1*(3), 1.

Akar, S., & Akar, E. (2020). Is it a new tulip mania age?: a comprehensive literature review beyond cryptocurrencies, bitcoin, and blockchain technology. *Journal of Information Technology Research (JITR)*, *13*(1), 44-67.

Al-Yahyaee, K. H., Mensi, W., & Yoon, S.-M. (2018). Efficiency, multifractality, and the long-memory property of the Bitcoin market: A comparative analysis with stock, currency, and gold markets. *Finance Research Letters*, *27*, 228-234.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of finance*, *59*(3), 1259-1294.

Azqueta-Gavaldón, A. (2020). Causal inference between cryptocurrency narratives and prices: Evidence from a complex dynamic ecosystem. *Physica A: Statistical Mechanics and its Applications*, *537*, 122574.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

Colianni, S., Rosales, S., & Signorotti, M. (2015). Algorithmic trading of cryptocurrency based on Twitter sentiment analysis. *CS229 Project*, *1*(5), 1-4.

El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A. (2018). A novel adaptable approach for sentiment analysis on big social data. *Journal of Big Data*, *5*(1), 1-18.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383-417.

Gkillas, K., & Longin, F. (2019). Is Bitcoin the new digital gold? Evidence from extreme price movements in financial markets. *Evidence From Extreme Price Movements in Financial Markets (January 18, 2019)*.

Greaves, A., & Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin. *No data*, *8*, 416-443.

Grobys, K., Ahmed, S., & Sapkota, N. (2020). Technical trading rules in the cryptocurrency market. *Finance Research Letters*, *32*, 101396.

Hesse, W., Möller, E., Arnold, M., & Schack, B. (2003). The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of neuroscience methods*, *124*(1), 27-44.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the international AAAI conference on web and social media,

Kaminski, J. (2014). Nowcasting the bitcoin market with twitter signals. *arXiv preprint arXiv:1406.7577*.

Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, *65*, 101188.

Linton, M., Teo, E. G. S., Bommes, E., Chen, C., & Härdle, W. K. (2017). Dynamic topic modelling for cryptocurrency community forums. In *Applied quantitative finance* (pp. 355-372). Springer.

Luu Duc Huynh, T. (2019). Spillover risks on cryptocurrency markets: A look from VAR-SVAR granger causality and student'st copulas. *Journal of Risk and Financial Management*, *12*(2), 52.

Madan, I., Saluja, S., & Zhao, A. (2015). Automated bitcoin trading via machine learning algorithms. *URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan*, *20*.

McNally, S. (2016). *Predicting the price of Bitcoin using Machine Learning* Dublin, National College of Ireland].

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, 21260.

Nguyen, T. H., & Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., & Smeaton, A. F. (2009). Topic-dependent sentiment analysis of financial blogs. Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion,

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 2016 international conference on signal processing, communication, power and embedded system (SCOPES),

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*(1), 25-33.

Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis.

Sha, H., Hasan, M. A., Mohler, G., & Brantingham, P. J. (2020). Dynamic topic modeling of the COVID-19 Twitter narrative among US governors and cabinet executives. *arXiv preprint arXiv:2004.11692*.

Shiller, R. J. (2017). Narrative economics. *American economic review*, *107*(4), 967-1004.

Shiller, R. J. (2020). *Narrative economics: How stories go viral and drive major economic events*. Princeton University Press.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),

Taskinsoy, J. (2019). Bitcoin: The Longest Running Mania–Tulips of the 21st Century? *Available at SSRN 3505953*.

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, *21*(6), 589.

Wang, T., Chen, P., & Li, B. (2017). Predicting the quality of short narratives from social media. *arXiv preprint arXiv:1707.02499*.

Zhang, W., Wang, P., Li, X., & Shen, D. (2018). The inefficiency of cryptocurrency and its cross-correlation with Dow Jones Industrial Average. *Physica A: Statistical Mechanics and its Applications*, *510*, 658-670.

# 7 Appendix

FIGURE 4: LDA TOPIC DECENTRALIZED FINANCED

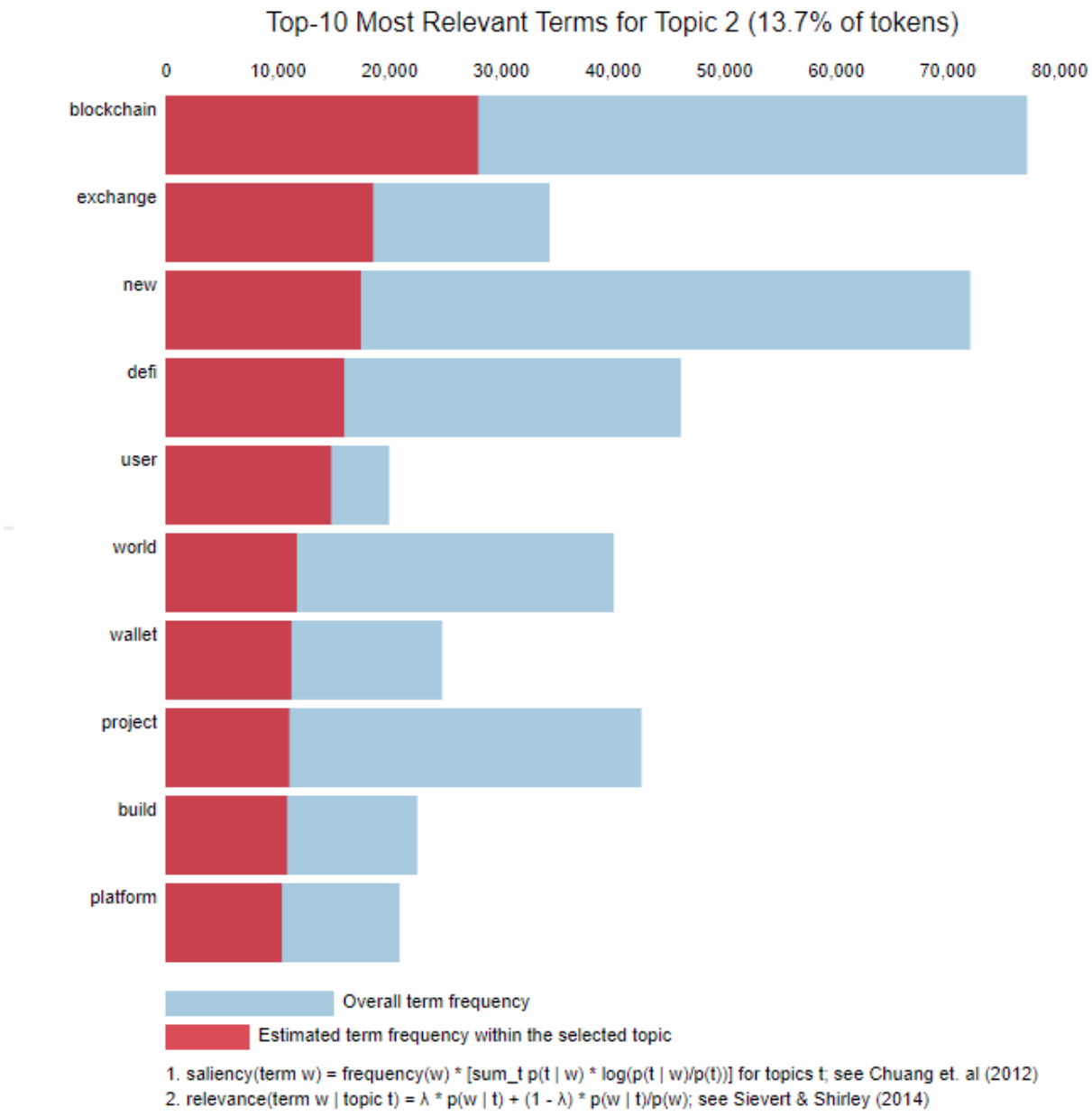

FIGURE 5: LDA TOPIC GAMING

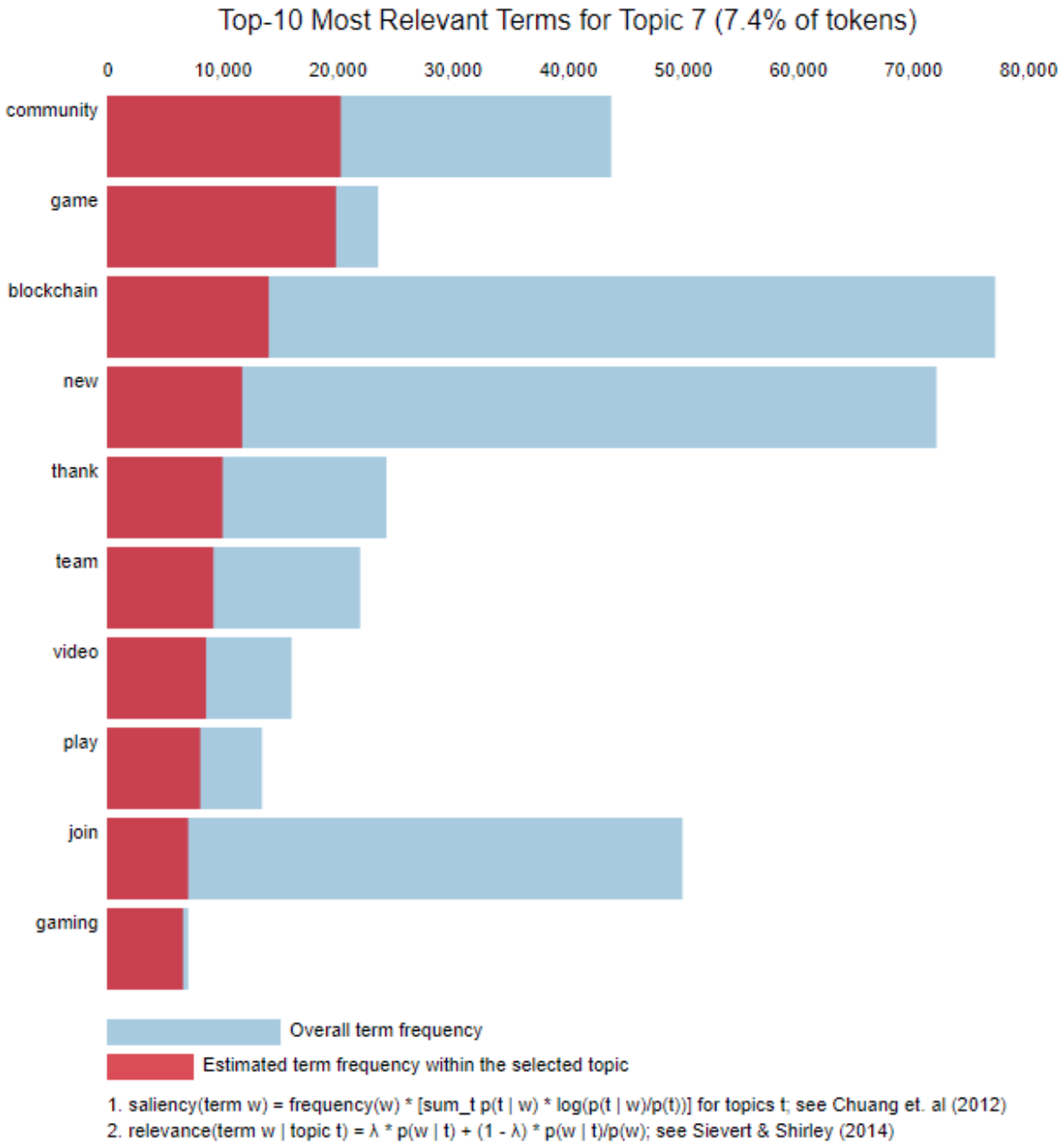

FIGURE 6: LDA TOPIC MEMECOINS

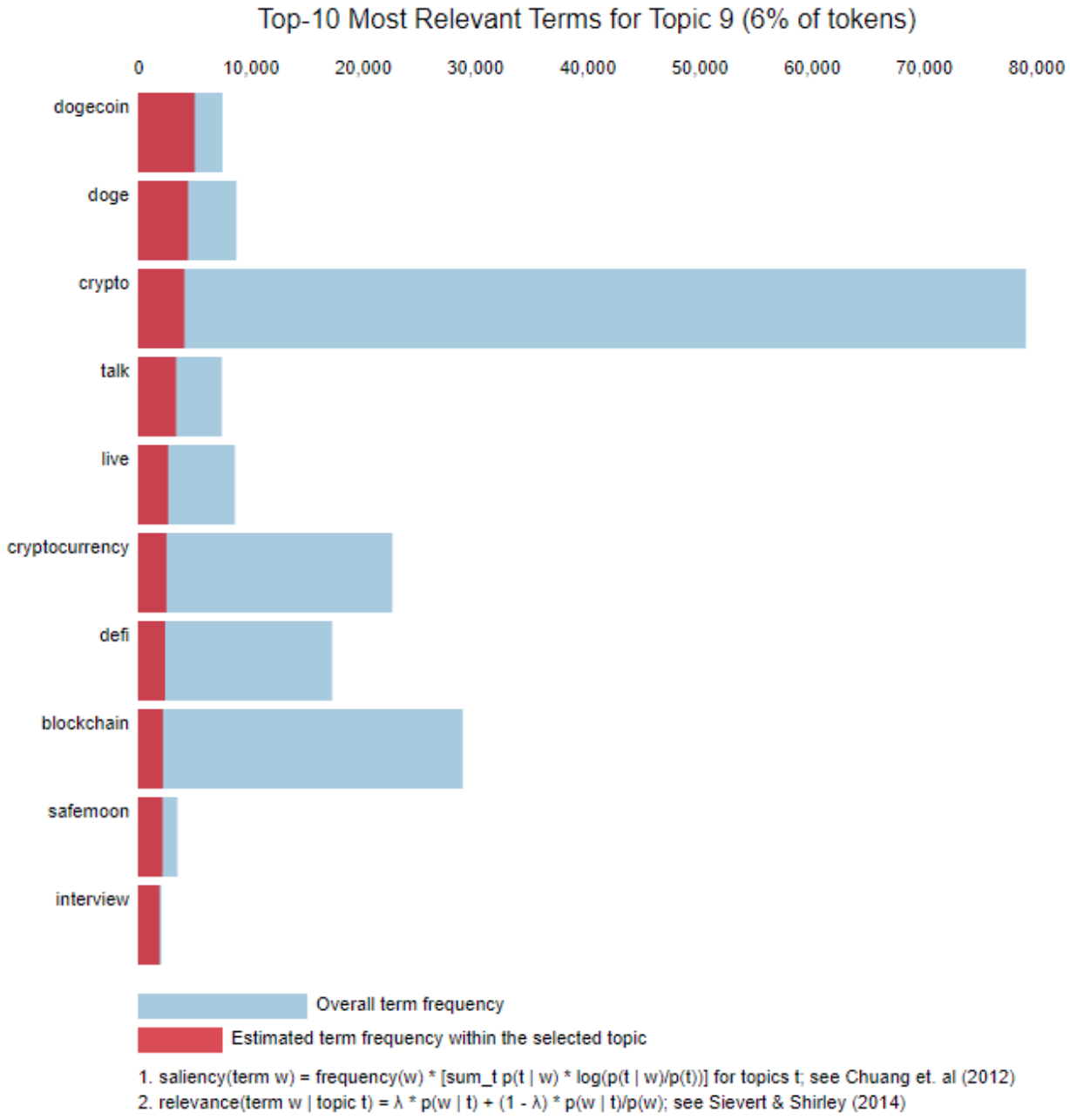

FIGURE 7: LDA TOPIC NFTS

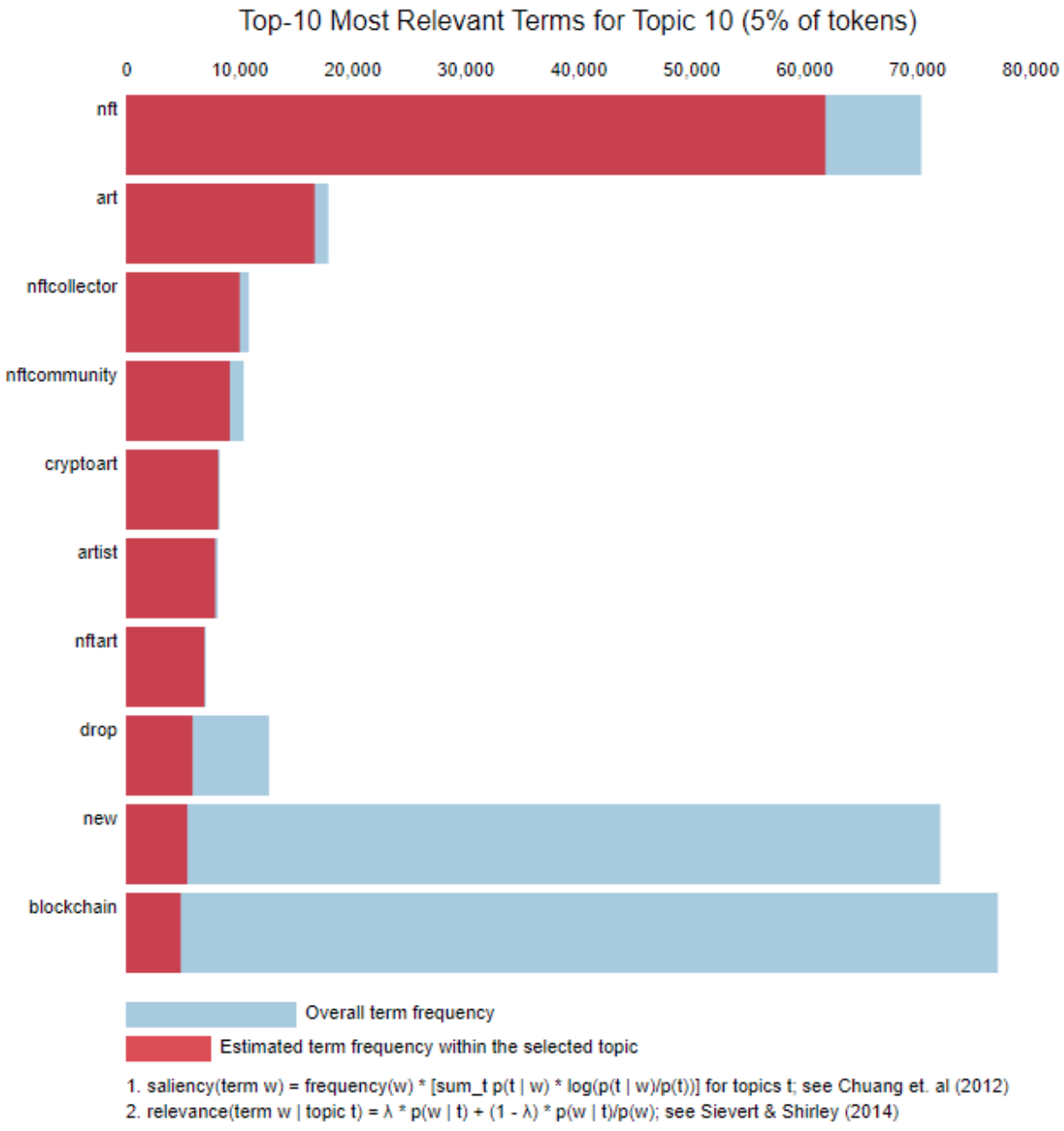

TABLE 5: OLS REGRESSION RESULTS: PRICE INSTEAD OF PRICE RETURNS

| | (1 ) $P^{BTC}$ | (2 ) $P^{ETH}$ | (3 ) $P^{ETH/BTC}$ | (4 ) $V^{BTC}$ | (5) $V^{ETH}$ | (6) $P^{BTC}$ | (7) $P^{ETH}$ | (8) $P^{ETH/BTC}$ | (9) $V^{BTC}$ | (10) $V^{ETH}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Total engagement | **.010728*** ** | **.001950*** ** | **2.13e-08*** ** | **18260.53*** ** | **19707.41*** ** | | | | | |
| | **(.002227)** | **(.000092)** | **(1.86e-09)** | **(5041.624)** | **(2394.292)** | | | | | |
| Total bullishness | **-2125.03*** ** | 19.3808 | .000716 | **-7.36e+09*** ** | **-4.59e+09*** ** | | | | | |
| | **(603.461)** | (24.9636) | (.000504) | **(1.37e+09)** | **(6.49e+08)** | | | | | |
| NFT engagement | | | | | | **.097308*** ** | **.007070*** ** | **-2.83e-08*** | **-97225.82** ** | **-108297.6*** ** |
| | | | | | | **(.019127)** | **(.000861)** | **(1.66e-08)** | **(44550.95)** | **(21267.45)** |
| NFT bullishness | | | | | | **-607.712*** ** | **-24.6362*** ** | -.000031 | -3.50e+08 | -2.06e+08 |
| | | | | | | **(134.993)** | **(6.08174)** | (.000117) | (3.14e+08) | (1.50e+08) |
| DeFi engagement | | | | | | -.0284841 | **.004326*** ** | **1.29e-07*** ** | 34035.09 | **79973.06*** ** |
| | | | | | | (.0260053) | **(.001172)** | **(2.25e-08)** | (60571.88) | **(28915.42)** |
| DeFi bullishness | | | | | | -27.0226 | .3263505 | -.000111 | -3.36e+08 | -2.75e+08 |
| | | | | | | (158.998) | (7.163216) | (.000138) | (3.70e+08) | (1.77e+08) |
| Gaming engagement | | | | | | **-.085130*** ** | **.004989*** ** | **1.15e-07*** ** | **-144128.1** ** | **-60289.15*** |
| | | | | | | **(.028639)** | **(.001290)** | **(2.48e-08)** | **(66706.9)** | **(31844.12)** |
| Gaming bullishness | | | | | | -131.8975 | **-11.6632** ** | -.000156 | -1286779 | -1.53e+08 |
| | | | | | | (118.1997) | **(5.32515)** | (.000102) | (2.75e+08) | (1.31e+08) |
| Memecoin engagement | | | | | | **.038936*** ** | **.0035984*** ** | **3.22e-08*** ** | **98828.84*** ** | **84749.24*** ** |
| | | | | | | **(.011075)** | **(.000499)** | **(9.59e-09)** | **(25796.74)** | **(12314.68)** |
| Memecoin bullishness | | | | | | **315.992** ** | -6.934337 | **-.000388*** ** | -1.48e+08 | **-2.63e+08*** |
| | | | | | | **(129.8858)** | (5.85163) | **(.000112)** | (3.03e+08) | **(1.44e+08)** |
| S&P500 price | **15.41741*** ** | **-.417415*** ** | **-6.12e-06** ** | **1.64e+07** ** | **2.24e+07*** ** | **18.0650*** ** | -.2068678 | **-8.87e-06*** ** | 4291876 | **1.10e+07*** ** |
| | **(3.00285)** | **(.12422)** | **(2.51e-06)** | **(6798984)** | **(3228871)** | **(3.13111)** | (.141063) | **(2.71e-06)** | (7293006) | **(3481489)** |
| Time control | **791.0454** ** | **190.5075*** ** | **.002801*** ** | **-2.13e+09*** ** | **-2.39e+09*** ** | **657.388*** | **183.553*** ** | **.003319*** ** | 1.65e+07 | -4.62e+08 |
| | **(323.8963)** | **(13.3987)** | **(.000271)** | **(7.33e+08)** | **(3.48e+08)** | **(342.164)** | **(15.4152)** | **(.000296)** | (7.97e+08) | (3.80e+08) |
| Interest rate | -2143.87 | **735.8252*** ** | **.009028*** ** | **-1.48e+10*** ** | **-1.35e+10*** ** | **-3558.69*** | **666.095*** ** | **.011711*** ** | -4.73e+09 | **-4.23e+09** ** |
| | (1776.20) | **(73.4769)** | **(.00148)** | **(4.02e+09)** | **(1.91e+09)** | **(1872.17)** | **(84.3450)** | **(.001621)** | (4.36e+09) | **(2.08e+09)** |
| Observations | 731 | 731 | 731 | 731 | 731 | 731 | 731 | 731 | 731 | 731 |
| R-squared | 0.8053 | 0.9374 | 0.8714 | 0.0868 | 0.2982 | 0.8123 | 0.9284 | 0.8669 | 0.0683 | 0.2765 |
| Adjusted R-squared | 0.8039 | 0.9370 | 0.8705 | 0.0805 | 0.2934 | 0.8094 | 0.9273 | 0.8649 | 0.0540 | 0.2654 |

*Notes:* Coefficients are reported with standard errors in parentheses. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively.

TABLE 6. MULTIVARIATE GRANGER-CAUSALITY RESULTS: BITCOIN PRICE RETURNS

| | $Relation$ | $E_{total}$ | $B^W_{total}$ | $E_{NFT}$ | $B^W_{NFT}$ | $E_{DeFi}$ | $B^W_{DeFi}$ | $E_{Gaming}$ | $B^W_{Gaming}$ | $E_{memecoin}$ | $B^W_{memecoin}$ | $P_r^{SP500}$ | $r$ | $\tau$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_r^{BTC}$ | ← | **0.017**** | **0.008***** | | | | | | | | | **0.025**** | 0.608 | **0.004***** |
| $P_r^{BTC}$ | → | 0.932 | 0.561 | | | | | | | | | 0.649 | 0.213 | 0.557 |
| $P_r^{BTC}$ | ← | | | **0.095*** | 0.413 | | | | | | | 0.414 | 0.649 | 0.202 |
| $P_r^{BTC}$ | → | | | 0.634 | 0.572 | | | | | | | 0.797 | 0.786 | 0.594 |
| $P_r^{BTC}$ | ← | | | | | **0.000***** | 0.509 | | | | | 0.354 | 0.623 | 0.066 |
| $P_r^{BTC}$ | → | | | | | **0.084***** | **0.087*** | | | | | 0.441 | 0.535 | 0.764 |
| $P_r^{BTC}$ | ← | | | | | | | 0.516 | 0.446 | | | 0.228 | 0.969 | 0.732 |
| $P_r^{BTC}$ | → | | | | | | | 0.394 | 0.944 | | | 0.449 | 0.135 | 0.613 |
| $P_r^{BTC}$ | ← | | | | | | | | | 0.204 | 0.626 | 0.154 | 0.847 | 0.203 |
| $P_r^{BTC}$ | → | | | | | | | | | 0.507 | 0.823 | 0.382 | 0.246 | 0.238 |

*Notes:* P-values are reported. Test performed with 4 lags. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively.

TABLE 7. MULTIVARIATE GRANGER-CAUSALITY RESULTS: ETHEREUM PRICE RETURNS

| | $Relation$ | $E_{total}$ | $B^W_{total}$ | $E_{NFT}$ | $B^W_{NFT}$ | $E_{DeFi}$ | $B^W_{DeFi}$ | $E_{Gaming}$ | $B^W_{Gaming}$ | $E_{memecoin}$ | $B^W_{memecoin}$ | $P_r^{SP500}$ | $r$ | $\tau$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_r^{ETH}$ | ← | **0.001***** | **0.030**** | | | | | | | | | 0.706 | 0.403 | **0.034**** |
| $P_r^{ETH}$ | → | 0.258 | **0.075*** | | | | | | | | | 0.623 | 0.317 | 0.998 |
| $P_r^{ETH}$ | ← | | | **0.056*** | 0.579 | | | | | | | 0.954 | 0.434 | 0.284 |
| $P_r^{ETH}$ | → | | | 0.522 | 0.420 | | | | | | | 0.522 | 0.974 | 0.976 |
| $P_r^{ETH}$ | ← | | | | | **0.004***** | 0.478 | | | | | 0.100 | 0.458 | 0.598 |
| $P_r^{ETH}$ | → | | | | | **0.000***** | **0.026***** | | | | | 0.832 | 0.659 | 0.963 |
| $P_r^{ETH}$ | ← | | | | | | | 0.637 | 0.692 | | | 0.880 | 0.634 | 0.669 |
| $P_r^{ETH}$ | → | | | | | | | 0.604 | **0.022*** | | | 0.527 | 0.581 | 0.981 |
| $P_r^{ETH}$ | ← | | | | | | | | | 0.836 | 0.768 | 0.890 | 0.868 | 0.417 |
| $P_r^{ETH}$ | → | | | | | | | | | 0.712 | 0.407 | 0.211 | 0.510 | 0.839 |

*Notes:* P-values are reported. Test performed with 4 lags. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively.

TABLE 8. MULTIVARIATE GRANGER-CAUSALITY RESULTS: ETH/BTC PRICE RETURNS

| | $Relation$ | $E_{total}$ | $B^W_{total}$ | $E_{NFT}$ | $B^W_{NFT}$ | $E_{DeFi}$ | $B^W_{DeFi}$ | $E_{Gaming}$ | $B^W_{Gaming}$ | $E_{memecoin}$ | $B^W_{memecoin}$ | $P_r^{SP500}$ | $r$ | $\tau$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_r^{ETH/BTC}$ | ← | **0.001*** ** | 0.211 | | | | | | | | | 0.115 | 0.563 | 0.082 |
| $P_r^{ETH/BTC}$ | → | 0.118 | 0.508 | | | | | | | | | **0.081*** | **0.002*** ** | **0.001*** ** |
| $P_r^{ETH/BTC}$ | ← | | | **0.010*** ** | 0.764 | | | | | | | 0.190 | 0.181 | 0.291 |
| $P_r^{ETH/BTC}$ | → | | | 0.258 | 0.685 | | | | | | | 0.045 | 0.728 | 0.787 |
| $P_r^{ETH/BTC}$ | ← | | | | | 0.560 | 0.916 | | | | | 0.121 | 0.326 | 0.360 |
| $P_r^{ETH/BTC}$ | → | | | | | **0.001*** ** | 0.777 | | | | | 0.339 | 0.472 | 0.615 |
| $P_r^{ETH/BTC}$ | ← | | | | | | | 0.437 | 0.832 | | | 0.156 | 0.177 | 0.556 |
| $P_r^{ETH/BTC}$ | → | | | | | | | 0.312 | **0.000*** ** | | | 0.270 | 0.822 | 0.485 |
| $P_r^{ETH/BTC}$ | ← | | | | | | | | | **0.007*** ** | 0.173 | **0.087*** | 0.759 | **0.006*** ** |
| $P_r^{ETH/BTC}$ | → | | | | | | | | | **0.058** ** | **0.005*** ** | 0.819 | 0.628 | 0.127 |

*Notes:* P-values are reported. Test performed with 4 lags. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively.

TABLE 9. MULTIVARIATE GRANGER-CAUSALITY RESULTS: BITCOIN VOLUME

| | $Relation$ | $E_{total}$ | $B^W_{total}$ | $E_{NFT}$ | $B^W_{NFT}$ | $E_{DeFi}$ | $B^W_{DeFi}$ | $E_{Gaming}$ | $B^W_{Gaming}$ | $E_{memecoin}$ | $B^W_{memecoin}$ | $P_r^{SP500}$ | $r$ | $\tau$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Vol^{BTC}$ | ← | **0.025** ** | 0.212 | | | | | | | | | **0.025** ** | 0.208 | 0.125 |
| $Vol^{BTC}$ | → | 0.490 | . | | | | | | | | | . | . | . |
| $Vol^{BTC}$ | ← | | | 0.984 | 0.875 | | | | | | | **0.056*** | 0.230 | 0.438 |
| $Vol^{BTC}$ | → | | | 0.409 | . | | | | | | | . | . | . |
| $Vol^{BTC}$ | ← | | | | | 0.155 | **0.100*** | | | | | 0.021 | **0.044** ** | 0.529 |
| $Vol^{BTC}$ | → | | | | | 0.419 | . | | | | | . | . | . |
| $Vol^{BTC}$ | ← | | | | | | | 0.622 | 0.195 | | | **0.013** ** | 0.052 | 0.136 |
| $Vol^{BTC}$ | → | | | | | | | 0.768 | . | | | . | . | . |
| $Vol^{BTC}$ | ← | | | | | | | | | **0.098*** | 0.535 | **0.030** ** | **0.092*** | 0.156 |
| $Vol^{BTC}$ | → | | | | | | | | | **0.079*** | . | . | . | . |

*Notes:* P-values are reported. Test performed with 4 lags. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively.

TABLE 10. MULTIVARIATE GRANGER-CAUSALITY RESULTS: ETHEREUM VOLUME

| | $Relation$ | $E_{total}$ | $B^W_{total}$ | $E_{NFT}$ | $B^W_{NFT}$ | $E_{DeFi}$ | $B^W_{DeFi}$ | $E_{Gaming}$ | $B^W_{Gaming}$ | $E_{memecoin}$ | $B^W_{memecoin}$ | $P_r^{SP500}$ | $r$ | $\tau$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Vol^{ETH}$ | ← | 0.223 | 0.294 | | | | | | | | | 0.972 | 0.109 | **0.036**** |
| $Vol^{ETH}$ | → | **0.033**** | **.** | | | | | | | | | . | . | . |
| $Vol^{ETH}$ | ← | | | 0.416 | 0.658 | | | | | | | 0.983 | 0.169 | **0.028***** |
| $Vol^{ETH}$ | → | | | **0.009***** | . | | | | | | | . | . | . |
| $Vol^{ETH}$ | ← | | | | | **0.096*** | 0.886 | | | | | 0.995 | 0.178 | **0.015***** |
| $Vol^{ETH}$ | → | | | | | **0.012**** | . | | | | | . | . | . |
| $Vol^{ETH}$ | ← | | | | | | | 0.486 | 0.836 | | | 0.987 | 0.297 | 0.319 |
| $Vol^{ETH}$ | → | | | | | | | 0.527 | . | | | . | . | . |
| $Vol^{ETH}$ | ← | | | | | | | | | **0.010***** | **0.100*** | 0.963 | 0.115 | **0.010***** |
| $Vol^{ETH}$ | → | | | | | | | | | **0.079*** | . | . | . | . |

*Notes:* P-values are reported. Test performed with 4 lags. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively.