Can YouTube Sentiment Information Forecast Bitcoin Returns?

Pierre Fay*, David Bourghelle*, Fredj Jawadi*

*IAE Lille University School of Management
pierre.fay.etu@univ-lille.fr, david.bourghelle@univ-lille.fr, fredj.jawadi@univ-lille.fr

January 29, 2023

Abstract

This study aims at investigating whether the sentiment information collected using YouTube could help to improve the forecast of Bitcoin returns. To this end, we collect daily data over the period 2017-2022, which is relevant as our sample includes calm period (before the coronavirus crisis), as well as turbulent times (COVID-19 outbreak and post-covid phase) capturing therefore different types and episodes of emotions. Further, unlike previous literature, we rely on YouTube videos to propose two sentiment proxies: the investor attention on YouTube (Daily number of views of YouTube videos) and the sentiment of investor on YouTube (the average daily sentiment of these videos). Econometrically, we assess for lead-lag effects and we set up a linear Vector Autoregressive (VAR) model to specify the interdependence dynamics between sentiment and bitcoin returns. We also propose to evaluate the forecasting performance of bitcoin returns with YouTube sentiment using state-of-the-art deep learning LSTM model. Our study provides two interesting results. First, we find that the attention and the sentiment toward specific subjects on YouTube (Hacks, Tutorials and videos about "crypto personalities") are relevant, which helps to explain the Granger causality of attention and sentiment toward bitcoin returns. Second, we show that the consideration of sentiment information help to provide better forecasts than does a benchmark Buy and Hold Strategy.

Keywords: YouTube Sentiment, bitcoin returns, VAR Model, LSTM Model, Forecast.

JEL: C2, F10, G10.

1 Introduction

Following the informational Efficient Market Hypothesis (EMH) of Fama (1965, 1970), the price of a financial asset follows a random walk process and the returns of this financial asset are independent and they even show the properties of a white noise. Leaving the framework of EMH, in particular when considering its failure to explain several financial crises on the stock market: 1636 tulip mania, 1987 stock crash, 2000 dotcom bubble, 2007 subprime crises, among others, suggests a further dependence structure for these returns and it becomes necessary to explain and specify this dependence structure for the return's dynamics. For a standard financial asset, it does happen always that economic fundamentals are used to try to explain this dependence across returns given that it is admitted that a financial asset should have a fundamental value and that at the equilibrium its price should converge toward its fundamental or fair value (Samuelson, 1965). Other approaches also such as chartist techniques can be used to characterize and reproduce return's dependence dynamics. For cryptomonnaies, the fundamental analysis is less credible as cryptocurrencies do not have an explicit or specific fundamental value. Rather, different alternative factors (news, behavioral factors, etc.) appear to drive the pricing of cryptomonnaies yielding different episodes of market up and down and raising the question of forecasting of crypto's returns in order to better apprehend changes in cryptocurrencie's prices.

Basically, news and investor's sentiments or emotions appear among these drivers, which is in line with the behavioral finance theory founded by Amos Tversky, Daniel Kahneman, Richard Thaler, and developed later thanks to Robert Shiller and Richard Taffler through the irrational exuberance hypothesis (Shiller, 2015), animal spirits hypothesis (Akerlof and Shiller, 2009), Narrative economy (Shiller, 2019) and emotional finance hypothesis (Taffler, 2018), among others. Indeed, behavioral finance theory suggests that investor's behavior and psychology play an important role and that they might directly or indirectly impact the investment and funding decisions and choices and therefore the prices of financial assets. Indeed, unlike conventional finance theory, the behavioral economists consider investors as normal and even irrational. Indeed, investors always make cognitive errors and wrong decisions caused by their own biases because they have limits to their self-control. For example, when investors show further evidence of overconfidence, or over-reaction or over-representation, they can be guided more by these biases and their feeling than by rational reasoning or fundamentals. This behavior and the power of these psychological factors are high and more likely to appear when the financial market is so volatile and therefore open for high investor's appetite for trading and risk as for cryptomonnaies. Accordingly, the excess of volatility for cryptomonnaies in general and bitcoin in particular has attracted the attention of media, investors and regulators over the last years.

In the literature, several studies have been conducted in order to analyze the dynamics of cryptomonnaie's returns but the related results do not provide unanimous conclusions. For example, the weak form of efficiency hypothesis was tested for the bitcoin by several authors. Tran and Leirvik (2019) showed that bitcoin market alternates between inefficiency (dependence) and efficiency (independence). Other papers showed that bitcoin has become more efficient at the end of their study period (Urquhart (2016), Tran and

Leirvik (2019)), while in practice several serious corrections have characterized the bitcoin price (i.e. 2018 and 2021 bubbles) and this yields several questions about bitcoin market efficiency and the drivers of this cryptomonney. With reference to behavioral finance analysis of cryptomonnaie's dynamics and through the analysis of active engagement by investors on social mediaCorbet, B. Lucey, et al. (2019) found that the bitcoin market is particularly sensitive to behavioral factors and sentiments, which can play a key role to better explain changes in bitcoin price.

In this context, Barber and Odean (2008) showed that investors, because of their limited attention ability, invest mainly in stocks that have first caught their attention. The author pointed to a positive correlation between investor's attention and an increase in stock prices. Some authors reached the same conclusion for cryptocurrencies when approximating investor's attention using the google search intensity, but there is still no consensus on the direction of this relationship. Indeed, Garcia et al. (2014) showed evidence that a spike in google searches preceedes a sharp price decline of bitcoin. Urquhart (2018) showed that an increase in Google search intensity is associated with an amplification of an upward or downward trend. Bouoiyour and Selmi (2015), Philippas (2019),Nasir et al. (2019), Liu and Tsyvinski (2021) found that Google search is positively and significantly correlated with bitcoin returns. In the same context, (Ciaian, Rajcaniova, and Kancs (2016)) studied the impact of investor attention measured by the intensity of discussion on internet forums and reached the same conclusion.

Besides, investor's sentiment seems to play a key role when taking investment decisions (Bollen, Mao, and Zeng 2011, Nofsinger 2005). Accordingly, Shleifer and Summers (1990) showed that investor sentiment can drive the price of financial asset and that it provides a key information to forecast the dynamics of financial asset's prices (Tetlock (2007)). Bourghelle, Jawadi, and Rozin (2022) studied the impact of investor sentiment using the Fear and Greed index [1] on Bitcoin volatility and showed that investors sentiment has a time-varying effect of Bitcoin volatility. Further, investor's sentiment might help to forecast bitcoin volatility.

That is, while the related previous literature suggests the usefulness of investor's sentiment and attention, the conclusion varies with the measure of these two variables, which is still challenging. Indeed, investor's information sources exceed the newspapers and rely more on the web 2.0, internet forums and social medias, which have became the main drivers of their sentiments and attention. For example, Twitter has become a popular source of sentiment mining. In fact, Bollen, Mao, and Zeng (2011) used Twitter's news to forecast the price of the Dow Jones Industrial Average. Philippas (2019) highlighted that Twitter's news might impact the prices of cryptomonnaies in particular in a context of uncertainty. Mai, Bai, and Shan (2015) showed that filtering tweets on users with the most followers reveals a significant relationship between the sentiment of the tweets and the return of bitcoin. Reddit is also a popular source of data in the financial literature. In fact, Long, B. M. Lucey, and Yarovaya (2021) showed that investor's sentiment informaiton recorded in Reddit has had a key role to explain abrupt changes of GameStop price in 2021.

[1] The Crypto Fear and Greed index aggregate various source of data traducing the sentiment of investors.

Empirically, different proxies were used to measure investor's sentiment: through a survey using using a proxy variable (internet searches, technical indicators, put/call ratios, etc.) as in Baker and Wurgler (2003), an algorithm on a media item (press articles, messages in social networks, internet forums, etc.) as in Tetlock (2007). Also, other sentiment analysis methods have been applied. For example, VADER proposed by Hutto and Gilbert (2015), is a rule-based and lexicon-based approach to sentiment analysis, which has been widely used for sentiment analysis on social media, but it is not specialized for Financial texts. Another popular sentiment analysis method is FinBERT, proposed by Araci (2019), which is a transformer based deep learning technique used to perform sentiment analysis on financial text.

Some other studies tested the impact of sentiment on cryptocurrenices through a topic modelling approach, which is is text mining technique that extracts topics and associated keywords of a corpus. Topic modeling was also applied to assess the impact of some particular topics in the medias and social networks on cryptos. For example, Corbet, Larkin, et al. (2020) showed that macroeconomic news are relevant to forecast bitcoin returns. Phillips and Gorse (2018), Uras, Vacca, and Destefanis (2020) and Ortu et al. (2022) showed that the occurrence of certain types of topics can help predict certain types of price movements. Loginova et al. (2021) combined the use of VADER and topic modeling technique with various source of information (Google Trends, Reddit, Cryptocompare, forums and news) to improve bitcoin forecasting.

This paper aims at investigating the impact of investor's and sentiment attention on the formation and forecasting of bitcoin returns. Unlike previous related studies, we rely on YouTube, which is the second most popular social network [2], to assess for sentiment and attention news. This is relevant because not only YouTube news is playing a key role in opinion formation, (Susarla, Oh, and Tan 2012), but also, to our knowledge, this is the first study that studies uses YouTube videos to asses the relationship between bitcoin returns and sentiment/attention news. Accordingly, this study attempts to address this gap in the academic literature on Bitcoin by studying the impact of attention and sentiments of YouTube videos on Bitcoin returns. Our second contribution is related to the application of Deep Learning Models and in particular the class of Long Short-Term Memory (LSTM) model to forecast the bitcoin returns. In fact, unlike usual time series models (ARMA, VAR model, etc.), a LSTM model enables us to reproduce further long-term dependence induced by persistence in investor sentiment and emotions (Wang, Shen, and Li (2022)). For more applicaiton of LSTM model on stock markets and Gold market, see Nelson, Pereira, and Oliveira 2017 and Livieris, E. Pintelas, and P. Pintelas 2020 respectively. Our findings show two interesting results. We propose an original measure of investor attention obtained when assessing the number of views of YouTube videos. Further, we proxy investor's sentiment using the sentiment of YouTube videos. These two proxies are particularly relevant as they capture investor's attention and the personal opinion and feeling of the publisher who uses YouTube. Second, we study the effect of attention and sentiment of investors provided by YouTube on bitcoin returns. Interestingly, both an overall impact of sentiment as well as the effect of attention and sentiment of specific subjects separately are investigated. In order to better explain the

[2] According to Statista, 2021. Full study : https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

contribution of sentiment/attention YouTube news, we carry out and provide out-of-sample forecasts using LSTM with and without these news. Our main results shows that a LSTM forecast with YouTube sentiments news outperforms the forecast of a buy and hold strategy model. The remainder of this paper is organized into four sections. Section 2 presents our data and explains our methodology. Section 3 discusses the empirical results. Section 4 concludes.

2 Data and Methodology

Our study uses daily data and covers the period: 17 August 2017 - 11 November 2022 accoungting for 1912 observations, which includes several bitcoin market overreaction and crashes. The OHLC and Volume data were collected from Binance API as Binance is the leading cryptocurrency exchange platform by volume. Using daily bitcoin prices, we computed the bitcoin returns as a first difference of bitcoin prices in logarithm. As for investor's attention and sentiment data, We used the API of YouTube to search for videos related to "Bitcoin" keyword. The total number of videos gathered is 63 862. For each video, we extracted the number of views received by the video under consideration and its respective title. We present hereafter the process to extract a sentiment from the video, the classification of videos per subject and the calculation of our daily variables from these data to proxy investor's attention and sentiment.

2.1 Sentiment Analysis

With the increasing amount of user generated content, the interest in automatic sentiment analysis has increased in the last years. Various methodology have been used to analyze the sentiments of financial texts from online newspapers or social networks like Twitter or Reddit. Unlike previous studies, we prefer the extraction of investor's sentiment through analysis of YouTube videos while investigating the sentiment of YouTube videos using their titles. To this end, we use state-of-the-art deep learning technique FinBERT proposed by Araci (2019) for sentiment analysis. FinBERT (Financial Bidirectional Encoder Representations from Transformers) is a transformer based deep learning technique based on BERT model published by Devlin et al. (2019) from Google AI Language. It is pre-trained on 1.8M news articles from Reuters TRC2 dataset, published between 2008 and 2010.

Our sentiment analysis approach is original and innovative. Indeed, in practice, the traditional word embedding technique builds a global vocabulary using a unique representation for each word in the documents. Accordingly, with these methods, a word can have only one representation. FinBERT, that we apply in this study, uses rather a contextual embedding, which aims to learn multiple representation for each word in the documents and therefore allows a word to have different representations depending on the context. Also, A lot of previous papers used VADER, a lexicon based approach to sentiment analysis, unlike these studies, we preferred FinBERT as it performs better. In the same context, Leow, Nguyen, and Chua (2021) and Mishev et al. (2020) showed also that the improved efficiency of transformers over lexicon for financial sentiment analysis of FinBERT is not rejected when considering Twitter.

In practice, With FinBERT, we classified our videos as positive, negative or neutral using their titles as input to the model, as example, we report some results of this classification in table 1.

Table 1 Sentiment analysis on YouTube video titles using FinBERT.

Title of the video	Sentiment
Why Alexandria green new deal is bullish...	positive
3 trends show ethereum is on track for strong growth...	positive
4 things you need to know about four tokens	neutral
5 altcoins to look out for this summer	neutral
4 reasons why bitcoin price continues to crash	negative
$31 million in Ethereum liquidates in past 12 hours	negative

Note: This table shows some examples of FinBERT sentiment analysis on YouTube video titles from our dataset.

In our database, the sentiment field is encoded as 1 for positive, 0 for neutral and -1 for negative to be able to calculate the daily mean sentiment.

2.2 Classification of the videos by subjects

Next, we classify each YouTube video of the dataset by subjects using their titles. First, we manually define a list of keywords for each subject. This list of keywords is used by the classification algorithm to classify videos among the corresponding subjects. An example of words associated to subject is reported in table 2.

Table 2 List of subjects and example of associated keywords

Subject	Associated keywords example
Hacks	scam, phished, hack, pirate, attack, steal
Network activities	mining, addresses, miner, farm, pools, network
Bitcoin adoption	partnership, adoption, accepted
Institutional and Central banks	institutional, bank, cdbc
Nft and Metaverse	nft, metaverse, opensea, axies, sport
Personality	ceo, burry, musk, butterin
Ico	ico, funding, participate, venture, capital
Trading robot	bot, robot
Regulation	ban, regulation, watchdog, lawsuit, authority
Price predictions	breakout, predict, analysis, high, resistance
Tutorials	explained, how, understanding

Note: This table shows the subjects and some examples of associated keywords used to classify our YouTube videos.

That is, we cleaned up the text and applied several preprocessing steps before moving to the classification algorithm. For example, we we remove stopwords, lower the text and delete special characters. We also lemmatize the text and filter it to get only names and adjectives using the Python NLTK library. Then, the classification algorithm looks at each video title: If a title contains one keyword associated to a subject, it assigns the subject to the video. If the title does not contain any of the keywords in the lists, it assigns the category "not classified" to the video.

2.3 Calculation of the independent variables

Let's recall that our dataset is composed of a list of 63 862 videos. For each video, we have the number of views it received, the date of publication of the video, the sentiment of the video and the subject of the video. From these data, we compute the daily independent variables used in the empirical analysis.

The first variable is the total number of views ($V_{d,s}$) of the day d and the subject s,

$$V_{d,s} = \sum_{i=0}^{n} v_{d,s,i} \tag{1}$$

where $v_{d,s,i}$ is the number of views of the video i of the day d and the subject s, and n the total number of videos of the day d. We also compute $V_{d,All}$ the total number of views of the day d for all subjects.

The second variable is the mean sentiment E of the day d and the subject s.

$$E_{d,s} = \frac{1}{n} \sum_{i=0}^{n} e_{d,s,i} \tag{2}$$

where $e_{d,s,i}$ is the sentiment of the video i of the day d and the subject s, and n the total number of videos of the day d. We also compute $E_{d,All}$ the daily mean sentiment of all the subjects.

3 Empirical Analysis

3.1 Measuring the impact of YouTube attention on Bitcoin returns

We reported all subjects in Table 3, the number of videos by subject and the total number of views by subject along with the respective percentages its represent.

Table 3 Subject distribution in our dataset

Subject (s)	Number of videos (N_s)	Number of views (V_s)
Hacks	894 (1.4)	4 907 500 (0.65)
Network activities	1 405 (2.2)	19 935 880 (2.62)
Bitcoin Adoption	223 (0.35)	4 383 895 (0.58)
Institutional and Central banks	662 (1.04)	7 464 115 (0.98)
NFT Metaverse	271 (0.42)	3 901 665 (0.51)
Personality	2 122 (3.32)	39 268 858 (5.17)
ICO	246 (0.39)	1 980 986 (0.26)
Bot	70 (0.11)	796 764 (0.1)
Regulation	679 (1.06)	8 927 921 (1.17)
Price predictions	46 891 (73.43)	335 207 613 (44.1)
Tutorials	3 185 (4.99)	104 040 005 (13.69)
Not classified	7 214 (11.3)	229 264 196 (30.16)

Note: This table shows the numbers of videos by subject in our dataset. Values in (.) denote the value in percentage.

We note that all the videos except the not classified and the "tutorials" concern the state of the market (price, regulation, hacks, network activities, bitcoin adoption, NFT / metaverse, ICO). Among these videos, price prediction represent 73.43% of the videos in our dataset, which shows the importance of narrative technical analysis among bitcoin traders on YouTube. This finding is not unexpected as unlike other social platforms where information is transmitted through a text, YouTube allows always the creation of videos and a video is the ideal media to show a graph and to comment it with technical analysis. Further, even though the price prediction subject represents 73.43% of the videos published on YouTube, it only accounts for 44.1% of the total views received. Interestingly, we plot the rolling 30 days mean of the number of views received by the subject "price predictions" on the figure 1. We can see a peak of activity during the 2018 and 2021 bitcoin bubbles, suggesting further evidence of linkage between bitcoin price movements and the attention of the investors on such videos.

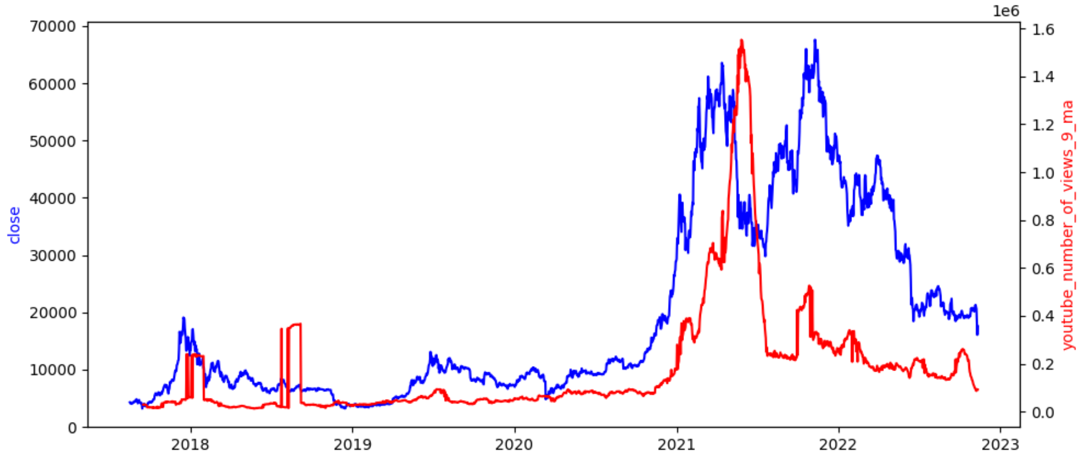

Figure 1: Number of views of the "price predictions" videos

We also note a significant number of "tutorials" videos, which are the kind of videos explaining how to create an account on an exchange and how to buy or sell the first Bitcoin. These videos are interesting as they help investors to take an action. We note that the "tutorials" subject account for only 4.99% of the videos about Bitcoin on YouTube, but this video receives 13.69% of the total views. Also, we can see a rolling 30 days mean of the number of views received by the subject "tutorials" on the figure 2. We note that these videos receive more views in period of bull and bear market. Interestingly, we see a first important price move and a spike in attention on tutorials videos in the 2018 bubble, and that the phenomenon increased for both variables during the bubble of 2021. Perhaps, this phenomenon was intensified because of the cryptocurrency increasing adoption by the public between these two periods.

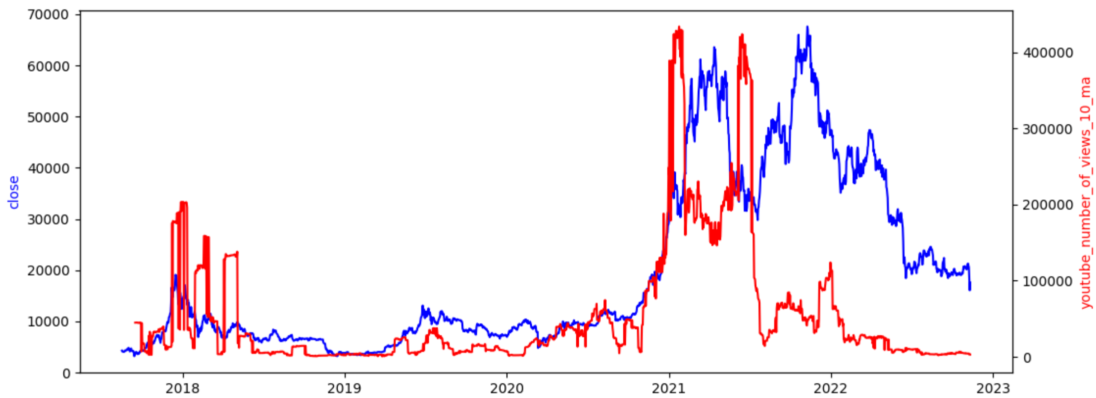

Figure 2: Number of views of the "tutorials" videos

Unlike other social media platforms, YouTube allows a direct measure of the investor attention by publishing the numbers of times a video has been watched by a user. Accordingly, we focus on the relationship between the number of views on YouTube on "Bitcoin" related videos and bitcoin returns. To do this, first, we check the stationarity of our variables and we reported the main results in table 4.

Table 4 Stationarity test results

	ADF statistic	p-value
V_{Hacks}	-3.795	0.003
$V_{NetworkActivities}$	-43.199	0.000
$V_{BitcoinAdoption}$	-12.677	0.000
$V_{InstitutionalAndCentralbanks}$	-6.645	0.000
$V_{NftandMetaverse}$	-5.538	0.000
$V_{Personality}$	-3.921	0.002
V_{Ico}	-14.410	0.000
$V_{Tradingrobot}$	-43.746	0.000
$V_{Regulation}$	-5.410	0.000
$V_{Pricepredictions}$	-3.557	0.007
$V_{Tutorials}$	-5.757	0.000
V_{All}	-3.261	0.017
r	-13.043	0.000

Note: This table shows the results of the augmented Dickey–Fuller test (ADF) for our "number of views" variables.

Next, we analyze the unconditional correlations between the bitcoin returns and the V_s proxy. We reported the main results in table 5.

Table 5 Correlation of V_s with returns

	Correlation
V_{Hacks}	-0.039
$V_{NetworkActivities}$	-0.018
$V_{BitcoinAdoption}$	0.025
$V_{InstitutionalAndCentralbanks}$	-0.015
$V_{NftandMetaverse}$	-0.023
$V_{Personality}$	-0.045
V_{Ico}	-0.011
$V_{Tradingrobot}$	-0.023
$V_{Regulation}$	-0.033
$V_{Pricepredictions}$	0.024
$V_{Tutorials}$	-0.012
V_{All}	0.012

Note: This table shows the unconditional correlations between our variables V_s and the returns of Bitcoin.

Accordingly, we note that subject number of views are weakly and insignificantly correlated with bitcoin returns. Further, only two subjects: "bitcoin adoption" and "price prediction" enter positively with bitcoin returns, sugegsting that an increase of investor's attention on bitcoin forecast and adoption might increase bitcoin returns and vice versa. The other subjects are negatively correlated with Bitcoin returns, which suggest, that when the attention of investors on these videos is increasing, bitcoin returns are decreasing and vice-versa. To go further in the analysis of linkages between these variables, we propose to check for causality relationships between these variables using Granger Causality test. We reported the main results in table 6 and obtain different findings.

Table 6 Granger causality test between V_s and returns

Null hypotheses	F-statistic	p-value
V_{Hacks} does not granger cause r	30.447	0.000
r does not granger cause V_{Hacks}	1.940	0.379
$V_{NetworkActivities}$ does not granger cause r	0.042	0.837
r does not granger cause $V_{NetworkActivities}$	0.219	0.640
$V_{BitcoinAdoption}$ does not granger cause r	1.387	0.239
r does not granger cause $V_{BitcoinAdoption}$	0.595	0.441
$V_{InstitutionalAndCentralbanks}$ does not granger cause r	0.503	0.478
r does not granger cause $V_{InstitutionalAndCentralbanks}$	1.431	0.232
$V_{NftandMetaverse}$ does not granger cause r	1.477	0.224
r does not granger cause $V_{NftandMetaverse}$	1.108	0.293
$V_{Personality}$ does not granger cause r	8.076	0.045
r does not granger cause $V_{Personality}$	16.134	0.001
V_{Ico} does not granger cause r	0.871	0.351
r does not granger cause V_{Ico}	2.938	0.087
$V_{Tradingrobot}$ does not granger cause r	1.498	0.221
r does not granger cause $V_{Tradingrobot}$	0.337	0.562
$V_{Regulation}$ does not granger cause r	1.045	0.593
r does not granger cause $V_{Regulation}$	0.252	0.616
$V_{Pricepredictions}$ does not granger cause r	1.926	0.382
r does not granger cause $V_{Pricepredictions}$	11.860	0.037
$V_{Tutorials}$ does not granger cause r	6.380	0.012
r does not granger cause $V_{Tutorials}$	0.006	0.940
V_{All} does not granger cause r	2.667	0.264
r does not granger cause V_{All}	4.728	0.316

Note: F-Statistic denotes the statistic of Fisher test and p-value denotes the p-value of this test.

First, we find no significant causality relationship between between the overall number

of views and Bitcoin returns, suggesting that an overall investor's attention does not cause bitcoin returns. However, when considering disaggregated data in particular when looking at the classification of our videos by subject, we can see some causal relationships. This result confirms the importance of decomposing investor's attention by subject when looking at the relationship between YouTube investor's attention and Bitcoin returns. In particular, we find no causality relationships between bitcoin returns and $V_{NetworkActivities}$, $V_{BitcoinAdoption}$,
$V_{InstitutionalAndCentralbanks}$, $V_{NftandMetaverse}$, $V_{Tradingrobot}$ or $V_{Regulation}$.

Second and interestingly, the hypothesis of Granger causality between bitcoin returns (r) and $V_{Pricepredictions}$ and between r and V_{Ico} suggesting further evidence of lead-lag effects. These causality relationships are unidirectional suggesting that the change in number of views of those videos is Granger caused by the returns of Bitcoin.. This does mean that bitcoin market performance might be attractive and activate more investor's attention. While price predictions represent more than 70% of the content about Bitcoin on YouTube and represents more than 40% of a total views on this media, we cannot however use the number of views of such videos to forecast future bitcoin price moves.

Third, the results of Granger causality test between bitcoin returns and V_{Hacks} showed a significant lead lag effect, even still unidirectional. Indeed, the change in investor's attention on these videos seems to Granger cause the bitcoin returns. This means that the information provided by the attention of investor on such subject can provide useful information to forecast the future bitcoin return. Indeed, in the cryptocurrency space, Hacks are common events and are an important risk for crypto investors. When there is a big hack, investors can feel insecure and sell their assets, or they can expect future price crash due to this event. The example of LUNA in 2022 which value has crashed in two hours after a hack, letting millions of investors with massive losses is an illustration of this risk. We also find a unidirectional relationship between $V_{Tutorials}$ and bitcoin returns, sugesting that these attention on kind of videos, considered as "call to actions" videos, could help to forecast bitcoin returns. Finally, we find a bilateral causality relationship between bitcoin returns and $V_{Personality}$. This is an interesting result, which is in line with the result of Huynh (2022) on Twitter that explains that the tone of the world's wealthiest person can drive Bitcoin returns. Here we can see that the attention on the videos about these people are driving bitcoin returns too, and it can trigger feedback loops explaining an apparent irrational investor's behaviors. In order to better assess these causality relationships, we run hereafter a linear VAR models allowing us to model the relationships between these variables within a 2 equation system for which each equation includes the lagged bitcoin return and a lagged value of the YouTube attention proxy. Taking the results of the Granger causality test, we consider only Youtube subject video that has a lead-lag effect with the bitcoin return.

Formally, we set up, for example, a bilateral VAR specification with one lag between the bitcoin return r and the Number of views on "ICO" V_{Ico} as:

$$\begin{cases} r_d = c_1 + a_1 r_{d-1} + a_2 V_{Ico,d-1} + e_1 \\ V_{Ico,t} = c_2 + a_3 V_{Ico,d-1} + a_4 r_{d-1} + e_2. \end{cases} \tag{3}$$

Where c is a constant and e the error terms.

In practice, for the following VAR models, we choose the number of lags by using the Bayesian Information Criteria (BIC).

For $V_{Pricepredictions}$, V_{Ico} the VAR models results are reported respectively in table 7 and 8. They confirm the unidirectional lead lag relationship. Bitcoin returns lead the change in attention on these videos, but the inverse relationship is not true.

For price predictions, the relationship is positive which means that a positive (resp. negative) change in bitcoin returns leads to a positive (resp. negative) change in attention on such videos. To explain this positive relationship, we can suppose that when bitcoin price is rising investors are searching for information confirming the current trend, but they are less prone to search for such price prediction when the market is falling due to their confirmation bias.

For ICO, the relationship is negative, which mean a positive (negative) change in returns will lead to a negative (positive) change in attention on such videos. To explain this negative relationship, we can suppose that when bitcoin returns are falling, people are searching for new positions to allocate their capital and are much in search for videos about ICO, whereas they are less in search of new investments opportunities when prices are rising because they will maintain their current positions.

Table 7 Result of a linear var model using price predictions videos

	$V_{PricePredictions}$		r	
C	-0.000	[-0.024]	0.001	[0.607]
$V_{PricePredictions,d-1}$	0.189***	[8.287]	0.000	[0.196]
r_{d-1}	0.295	[0.880]	0.094***	[4.073]
$V_{PricePredictions,d-2}$	0.149***	[6.493]	-0.001	[-0.896]
r_{d-2}	0.643*	[1.905]	0.008	[0.337]
$V_{PricePredictions,d-3}$	0.149***	[6.474]	0.001	[0.560]
r_{d-3}	-0.438	[-1.295]	-0.053**	[-2.285]
$V_{PricePredictions,d-4}$	0.145***	[6.339]	-0.002	[-1.345]
r_{d-4}	-0.265	[-0.783]	-0.059**	[-2.556]
$V_{PricePredictions,d-5}$	0.119***	[5.212]	-0.000	[-0.014]
r_{d-5}	0.837**	[2.482]	-0.021	[-0.920]
Number of observations		1906.000		
Log likelihood:		369.249		
BIC		-5.976		
AIC		-6.040		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

Table 8 Result of a linear var model using ICO videos

	V_{Ico}		r	
C	0.001	[0.027]	0.001	[0.522]
$V_{Ico,d-1}$	0.242***	[10.905]	-0.001	[-0.932]
r_{d-1}	-0.655*	[-1.713]	0.098***	[4.297]
Number of observations		1910.000		
Log likelihood:		83.006		
BIC		-5.739		
AIC		-5.756		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

For V_{Hacks}, $V_{Tutorials}$ the VAR models results are reported respectively in table 10 and 9. For $V_{Tutorials}$, results confirms the lead lag relationship too. Investors attention on tutorials positively impact Bitcoin returns which means that when attention on tutorials videos increase (decrease), Bitcoin returns increase (decrease).

Table 9 Result of a linear var model using tutorials videos

	$V_{Tutorials}$		r	
C	0.000	[0.005]	0.001	[0.522]
$V_{Tutorials,d-1}$	0.044*	[1.913]	0.003**	[2.524]
r_{d-1}	-0.030	[-0.076]	0.099***	[4.344]
Number of observations		1910.000		
Log likelihood:		28.438		
BIC		-5.682		
AIC		-5.699		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

For V_{Hacks}, results confirms the lead lag relationship, the impact of the attention on such videos is highly significant, and the relationship is negative which mean when there is more (less) attention on those videos, Bitcoin returns are decreasing (increasing). We note a potential bidirectional effect, but it has a poor significance. To explain this negative relationship we can suppose that such videos are transmitting fears of hacks to investors when the attention of investors on such subjects increase, they tend to sell their assets to reallocate their money. Our chart tends to confirm this hypothesis, we plot the number of views of "hacks" videos in figure 3. The highest peak is the 10 May 2021, it is the day of the hack of Terra (LUNA) blockchain.

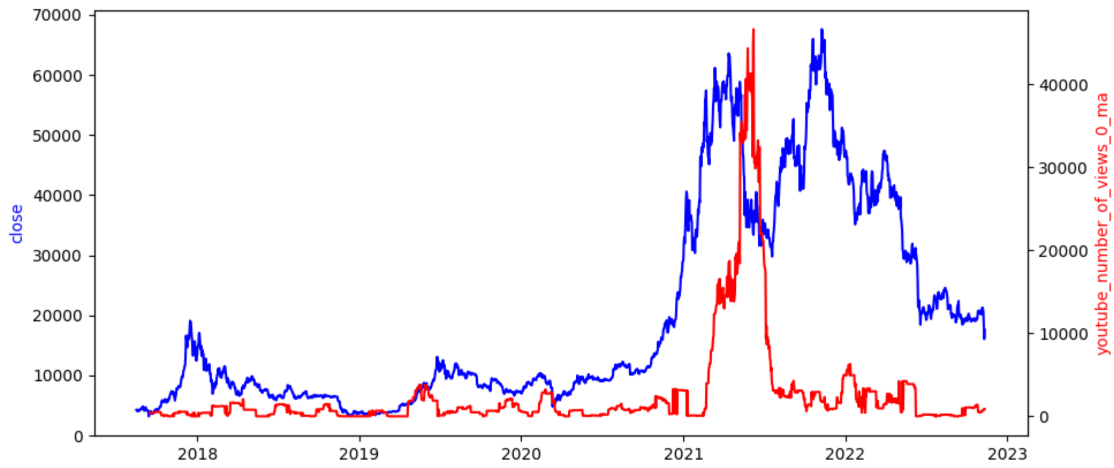

Figure 3: Number of views of the "hacks" videos

Table 10 Result of a linear var model using hacks videos

	V_{Hacks}		r	
C	-0.001	[-0.027]	0.001	[0.691]
$V_{Hacks,d-1}$	0.065***	[2.889]	-0.000	[-0.298]
r_{d-1}	0.088	[0.234]	0.095***	[4.105]
$V_{Hacks,d-2}$	0.012	[0.537]	0.000	[0.048]
r_{d-2}	0.345	[0.907]	0.011	[0.465]
$V_{Hacks,d-3}$	0.094***	[4.185]	-0.004***	[-3.268]
r_{d-3}	-0.058	[-0.153]	-0.063***	[-2.715]
$V_{Hacks,d-4}$	0.038*	[1.673]	0.001	[0.854]
r_{d-4}	0.265	[0.697]	-0.063***	[-2.708]
$V_{Hacks,d-5}$	0.006	[0.269]	0.001	[0.507]
r_{d-5}	0.210	[0.551]	-0.019	[-0.835]
$V_{Hacks,d-6}$	0.161***	[7.210]	-0.002	[-1.167]
r_{d-6}	0.156	[0.409]	-0.004	[-0.161]
$V_{Hacks,d-7}$	0.010	[0.462]	0.001	[0.673]
r_{d-7}	-0.266	[-0.702]	-0.014	[-0.584]
$V_{Hacks,d-8}$	0.040*	[1.757]	0.002	[1.571]
r_{d-8}	0.753**	[1.993]	-0.062***	[-2.676]
$V_{Hacks,d-9}$	0.009	[0.402]	-0.004***	[-2.972]
r_{d-9}	-0.411	[-1.087]	-0.053**	[-2.310]
$V_{Hacks,d-10}$	0.225***	[9.953]	0.001	[0.457]
r_{d-10}	0.165	[0.438]	0.023	[1.012]
Number of observations		1901.000		
Log likelihood:		191.708		
BIC		-5.711		
AIC		-5.833		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

For $V_{Personality}$ the VAR models results are reported in table 11. Result confirms the lead lag bilateral relationship. To better understand the dynamics of these variables we plot the impulse response function in Figure 4. Impulse response function allows us to visualize the evolution of a variable in reaction of an upward shock in another variable of the model.

We note that an upward shock in investors attention on "personality" videos create a positive returns for the first day followed by a deep correction the next 2 days. An upward shock in returns creates an increase in investors attention on this subject for the first two days followed by a deep correction the next 2 days. When a personality express his opinion on Bitcoin, the first two days see an increase in attention on this subject which create an overreaction of investors buying bitcoin. We can suppose a lot of these buyers just want to benefit from the event and sell when the rally is finished that's why the next days they sell their bitcoin creating a market correction. This result is interesting because it shows the power of "crypto personalities" over Bitcoin returns.

Table 11 Result of a linear var model using personality videos

	$V_{Personality}$		r	
C	0.001	[0.033]	0.001	[0.588]
$V_{Personality,d-1}$	0.097***	[4.238]	0.002*	[1.682]
r_{d-1}	0.881**	[2.469]	0.095***	[4.138]
$V_{Personality,d-2}$	0.153***	[6.965]	0.001	[0.482]
r_{d-2}	0.130	[0.362]	0.009	[0.395]
$V_{Personality,d-3}$	0.274***	[12.460]	-0.003**	[-2.274]
r_{d-3}	-1.247***	[-3.462]	-0.055**	[-2.388]
$V_{Personality,d-4}$	0.111***	[4.864]	-0.002*	[-1.671]
r_{d-4}	0.182	[0.505]	-0.059**	[-2.567]
Number of observations		1907.000		
Log likelihood:		251.421		
BIC		-5.868		
AIC		-5.921		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

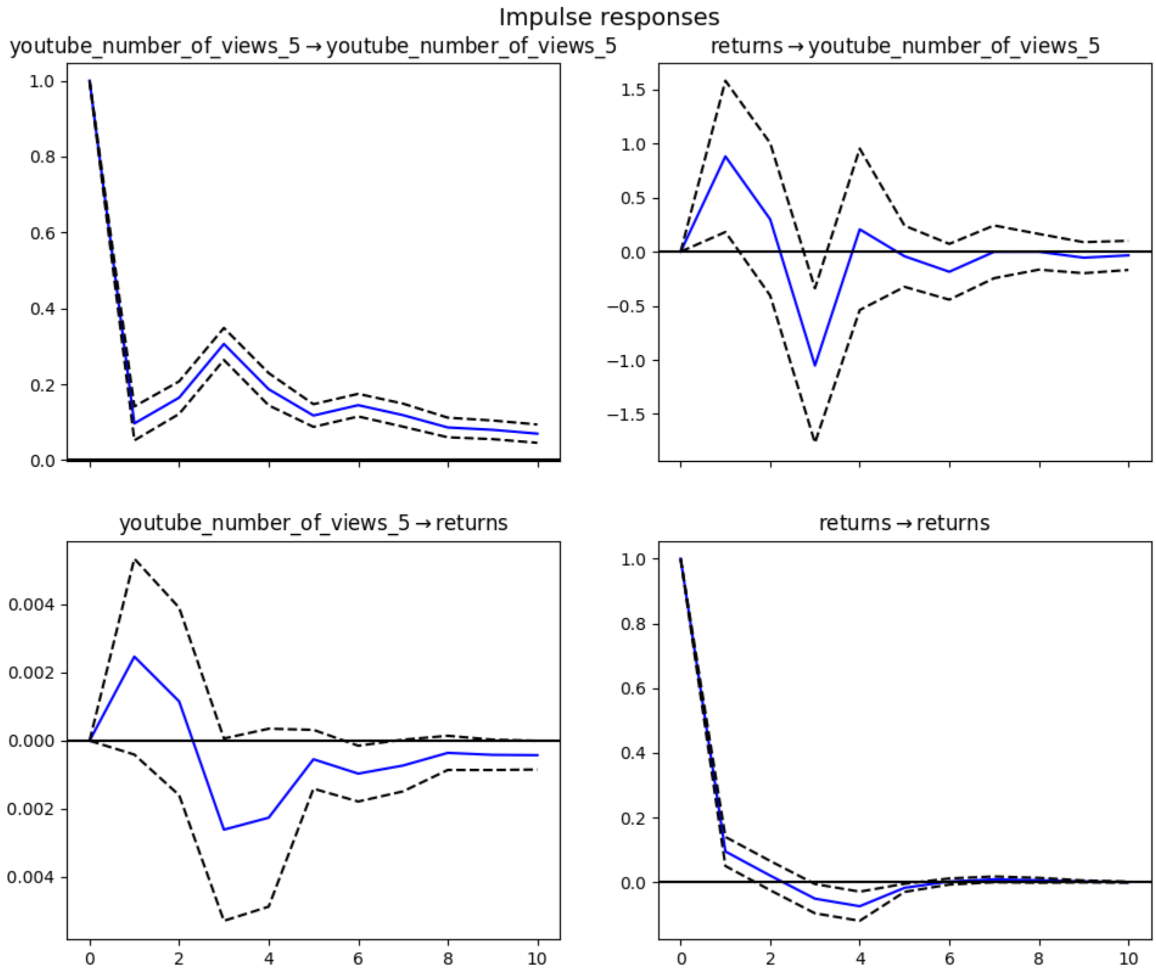

Figure 4: Impulse response function of investors attention on "personality" videos and returns

3.2 Part II : Impact of YouTube sentiment on Bitcoin returns

We report in table 12 the mean sentiment of videos for subjects on YouTube. We can see the overall mean sentiment on videos is slightly positive (0.023). Furthermore, we can note the videos about regulation subject tends to be associated with negative sentiment which indicate the mistrust in regulation by crypto investors. Recently the FTX crash highlighted the importance of implementing regulation in the cryptocurrency sector to protect investors, we can suppose this sentiment about regulation could evolve in the future.

Table 12 Mean sentiment by subject in YouTube

Subject	Mean sentiment
Hacks	-0.001
Network activities	0.026
adoption	0.103
Institutional and Central banks	0.026
Nft Metaverse	0.055
Personality	-0.002
Ico	0.057
Bot	0.057
Regulation	-0.100
Price predictions	0.023
Tutorials	0.011
Not classified	0.023
Mean	0.023

Note: This table shows the average sentiment of the YouTube videos in our dataset. This metric range from -1 for very negative to 1 for very positive.

Because the timeserie is very noisy, we plot the rolling 30 days mean of the daily overall sentiment of Bitcoin YouTube videos to visualize it more clearly in figure 12. We note that the sentiment increase during the creation of the bubble and sharp decline during explosion of bubble for both 2018 and 2021. We can suppose there is a relationship between these two series.

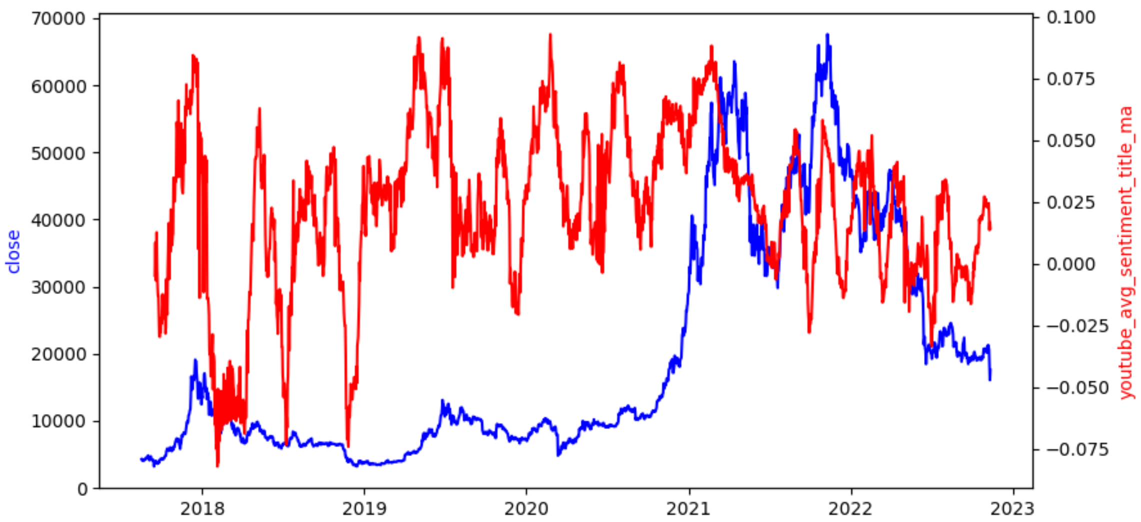

Figure 5: Number of views of the "tutorials" videos

First we check the stationarity of our sentiment variables, results are reported in table

13.

Table 13 Stationarity test results

	ADF statistic (p-value)
E_{Hacks}	-44.506 (0.000)
$E_{NetworkActivities}$	-42.251 (0.000)
$E_{BitcoinAdoption}$	-26.150 (0.000)
$E_{InstitutionalAndCentralbanks}$	-42.718 (0.000)
$E_{NftandMetaverse}$	-8.053 (0.000)
$E_{Personality}$	-44.407 (0.000)
E_{Ico}	20.936 (0.000)
$E_{Tradingrobot}$	-43.751 (0.000)
$E_{Regulation}$	-29.212 (0.000)
$E_{Pricepredictions}$	-11.897 (0.000)
$E_{Tutorials}$	-44.081 (0.000)
E_{All}	-6.385 (0.000)
r	-13.043 (0.000)

Note: This table shows the results of the augmented Dickey–Fuller test (ADF) for our "sentiment" variables.

Because we suspect a relationship between these variables, we'll check the correlation between our series E_s and Bitcoin returns r. We reported results in table 14. Result shows that the overall sentiment is the sentiment the most correlated with returns. Institutional and central banks and price predictions have the strongest correlation with returns among our variables.

Table 14 Correlation of E_s with returns

	Correlation
E_{Hacks}	0.000
$E_{NetworkActivities}$	0.049
$E_{BitcoinAdoption}$	0.041
$E_{InstitutionalAndCentralbanks}$	0.105
$E_{NftandMetaverse}$	0.044
$E_{Personality}$	0.039
E_{Ico}	0.010
$E_{Tradingrobot}$	-0.001
$E_{Regulation}$	0.024
$E_{Pricepredictions}$	0.156
$E_{Tutorials}$	0.031
E_{All}	0.182

Note: This table shows the correlations between our variables E_s and the returns of Bitcoin.

We know look at causality relationship between our sentiment variables and bitcoin returns using Granger Causality test, results are reported in table 15.

Table 15 Granger causality test between E_s and returns

Null hypotheses	F-statistic	p-value
E_{Hacks} does not granger cause r	1.001	0.317
r does not granger cause E_{Hacks}	2.157	0.142
$E_{NetworkActivities}$ does not granger cause r	1.081	0.299
r does not granger cause $E_{NetworkActivities}$	4.847	0.028
$E_{BitcoinAdoption}$ does not granger cause r	1.321	0.250
r does not granger cause $E_{BitcoinAdoption}$	0.066	0.797
$E_{InstitutionalAndCentralbanks}$ does not granger cause r	0.926	0.336
r does not granger cause $E_{InstitutionalAndCentralbanks}$	0.238	0.626
$E_{NftandMetaverse}$ does not granger cause r	0.003	0.954
r does not granger cause $E_{NftandMetaverse}$	2.935	0.087
$E_{Personality}$ does not granger cause r	0.017	0.897
r does not granger cause $E_{Personality}$	0.027	0.871
E_{Ico} does not granger cause r	0.110	0.740
r does not granger cause E_{Ico}	1.787	0.181
$E_{Tradingrobot}$ does not granger cause r	0.104	0.747
r does not granger cause $E_{Tradingrobot}$	0.011	0.917
$E_{Regulation}$ does not granger cause r	4.378	0.036
r does not granger cause $E_{Regulation}$	3.928	0.048
$E_{Pricepredictions}$ does not granger cause r	18.988	0.000
r does not granger cause $E_{Pricepredictions}$	3.617	0.057
$E_{Tutorials}$ does not granger cause r	0.030	0.863
r does not granger cause $E_{Tutorials}$	3.966	0.046
E_{All} does not granger cause r	21.406	0.000
r does not granger cause E_{All}	7.542	0.006

Note: F-Statistic is the statistic of the Fisher test and p-value denotes the p-value of the test.

Result shows sentiment of E_{Hacks}, $E_{BitcoinAdoption}$, $E_{InstitutionalAndCentralbanks}$, $E_{Personality}$, E_{Ico}, $E_{Tradingrobot}$ videos have no direct causality relationship with Bitcoin returns, so we won't go further on analysis on such variables. $E_{NetworkActivities}$, $E_{NftandMetaverse}$ and $E_{Tutorials}$ have a unidirectional lead lag relationship with returns. Sentiment of such subjects seems to be caused by Bitcoin returns but have no direct causality relationship on it.

VAR results are reported in tables 16, 17 and 18 are confirming the positive unidirectional relationship between the sentiment of such videos and bitcoin returns.

Table 16 Result of a linear var model for networks videos

	$E_{NetworkActivities}$		r	
C	-0.001	[-0.029]	0.001	[0.521]
$E_{NetworkActivities,d-1}$	0.031	[1.343]	-0.001	[-1.039]
r_{d-1}	0.867**	[2.200]	0.099***	[4.354]
Number of observations		1910.000		
Log likelihood:		29.170		
BIC		-5.683		
AIC		-5.700		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

Table 17 Result of a linear var model for nft videos

	$E_{NftAndMetaverse}$		r	
C	-0.001	[-0.024]	0.001	[0.522]
$E_{NftAndMetaverse,d-1}$	0.045**	[1.965]	-0.000	[-0.058]
r_{d-1}	0.674*	[1.712]	0.098***	[4.305]
Number of observations		1910.000		
Log likelihood:		28.275		
BIC		-5.682		
AIC		-5.699		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

Table 18 Result of a linear var model for tutorials videos

	$E_{Tutorials}$		r	
C	-0.001	[-0.026]	0.001	[0.522]
$E_{Tutorials,d-1}$	-0.011	[-0.460]	-0.000	[-0.172]
r_{d-1}	0.784**	[1.990]	0.098***	[4.310]
Number of observations		1910.000		
Log likelihood:		25.971		
BIC		-5.679		
AIC		-5.697		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

The overall sentiment of the market E_{All}, $E_{Regulation}$ and $E_{Pricepredictions}$ have a bidirectional lead lag relationship with Bitcoin returns. To clarify the causality relationships between these sentiment variables and bitcoin returns, we'll run a linear VAR models allowing us to model the relationships within a 2 equation system. For each model, we choose the number of lags by using the Bayesian Information Criteria (BIC).

First, the VAR model using E_{All} and r results are reported in table 19. There is a positive bilateral relationship of the overall YouTube sentiment on bitcoin videos and bitcoin returns which means when the sentiment of YouTube videos about bitcoin increase (decrease), the returns tends to increase (decrease). This relationship is bilateral which means that it can conduct to feedback loops: because returns are increasing, videos sentiment is increasing which in turn increase returns etc... This result is interesting because it shows that YouTube can play a role in an apparent irrationality of investors behavior.

Table 19 Result of a linear var model for all videos

	E_{All}		r	
C	-0.001	[-0.047]	0.001	[0.536]
$E_{All,d-1}$	0.134***	[5.824]	0.006***	[4.623]
r_{d-1}	1.086***	[2.744]	0.079***	[3.417]
Number of observations		1910.000		
Log likelihood:		83.931		
BIC		-5.740		
AIC		-5.757		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

Second, VAR results reported in table 20 confirms the bilateral relationship of $E_{Regulation}$ and Bitcoin returns. The relation is positive which means when the sentiment of videos about regulation is increasing (decreasing) bitcoin returns tends to increase (decrease). Cryptocurrencies are relatively new assets and the surrounding regulation is still a work in progress by financial authorities around the world. While investors are playing on such market, they have to follow the regulators rules of the games which have strong probabilities to change in the next future. For each change, the investors has to decide if it's worth staying on the game or not. If new rules are favorable to this new ecosystems it can lead investment to more returns or more safety but if rules are restricting the development of the ecosystem it can lead to future losses for investors. Investors will logically sell their assets if his probability of earning decrease and buying if his probability increase or if he fell more secure.

Table 20 Result of a linear var model for regulation videos

	$E_{Regulation}$		r	
C	-0.001	[-0.028]	0.001	[0.523]
$E_{Regulation,d-1}$	0.024	[1.030]	0.003**	[2.091]
r_{d-1}	0.780**	[1.980]	0.097***	[4.261]
Number of observations		1910.000		
Log likelihood:		28.246		
BIC		-5.682		
AIC		-5.699		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

Third, VAR results in table 22 confirms the positive bilateral relationship between the sentiment of price prediction videos and bitcoin returns. Which means that when the sentiment of such video increase (decrease), returns on bitcoin tends to increase (decrease) which in turn increase the sentiment of these videos etc... These results show the importance of the self-realization phenomenon of such video on Bitcoin prices.

Table 21 Result of a linear var model

	$E_{PricePredictions}$		r	
C	0.000	[0.007]	0.001	[0.571]
$E_{PricePredictions,d-1}$	0.126***	[5.491]	0.005***	[3.749]
r_{d-1}	0.566	[1.442]	0.078***	[3.406]
$E_{PricePredictions,d-2}$	0.108***	[4.674]	0.002*	[1.660]
r_{d-2}	-0.236	[-0.597]	-0.007	[-0.321]
$E_{PricePredictions,d-3}$	0.078***	[3.376]	0.004***	[2.854]
r_{d-3}	0.639	[1.624]	-0.073***	[-3.142]
Number of observations		1908.000		
Log likelihood:		109.498		
BIC		-5.735		
AIC		-5.776		

Note: (***), (**) and (*) denote significance at 1%, 5% and 10% statistical level respectively. Values in [.] denote the t-ratios. C denotes the constant.

3.3 Part III : Forecasting Bitcoin returns using YouTube

In the previous sections we have seen that $V_{Tutorials}$, V_{Hacks}, $V_{Personality}$, E_{All}, $E_{Regulation}$ and $E_{Pricepredictions}$ have a significant causality relationship with Bitcoin returns. We will now use them to predict Bitcoin returns using LSTM deep learning model.

Because the training of such models can lead to different results due to the stochastic nature of its initialization, we'll use Bootstrap Aggregation (also called as "bagging") strategy to reduce the forecast variance. The principle of bagging technique is to train

K models and to aggregate their predictions by taking the K forecasts average as final forecast. In this study we use an arbitrary value of K=100. The K models are trained using a 70 units LSTM layer and a Dropout Layer. The model is compiled using Adam optimizer with a learning rate of 0.0001. The Loss function used is MSE.

We separate our data into three sets. 80% of the dataset is used to train the model[1]. 10% of the dataset is used as validation set for cross validation [2] and 10% is used as test set to make an out of sample forecast. [3]

To ensure our variables improve our model, we will compare our model with past returns and YouTube variables as input to our model (named "YouTube model") with a model using only past returns as input (named "base model").

The metrics used to compare the two models forecasts are the Mean Squared Error (MSE) and the Mean Absolute Error (MAE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4}$$

$$MAE = \sum_{i=1}^{D}|x_i - y_i| \tag{5}$$

Table 22 MSE and MAE of our two models

Model	MSE	MAE
Base model	1.726E03	2.932E02
YouTube model	1.779E03	2.935E02

Note: This table shows the results of the forecast errors measured by the mean squared error (MSE) and the mean absolute error (MAE).

We plot the results of the out of sample forecast with and without our YouTube variables in figure 6.

[1]Training set : from to 2017-08-17 to 2021-10-24
[2]Validation set : from 2021-10-25 to 2022-05-03
[3]Test set : 2022-05-04 to 2022-11-09

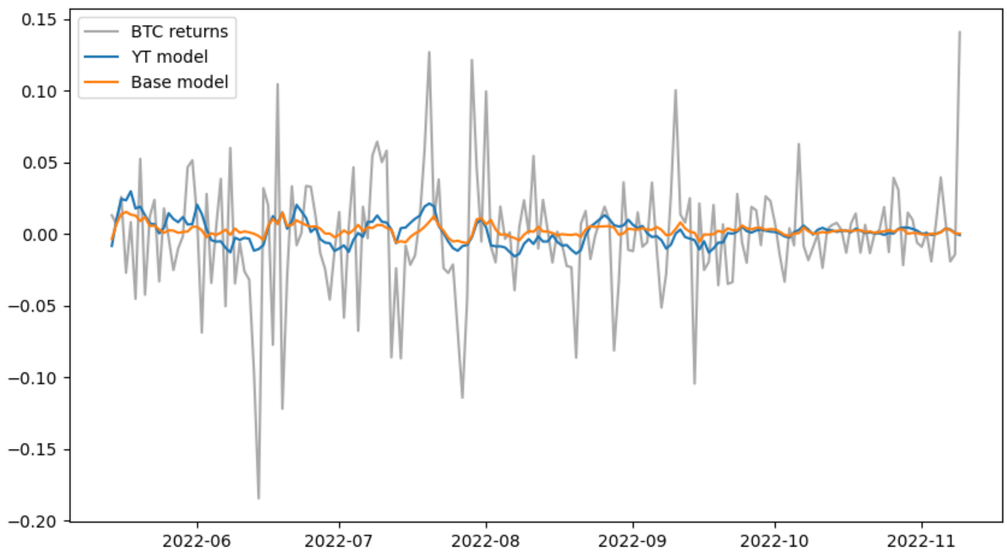

Figure 6: Bitcoin returns vs predicted returns

To compare the two forecast we use the Diebold-Mariano statistical test, which test the null hypothesis that the two forecast has the same accuracy. The results indicate a DM statistic of -1.272 and a p-value of 0.103. The null hypothesis cannot be rejected. Even if those data add values to the model it's not enough to make a statistical difference in the squared error. When looking at figure 6 we can see the forecast of YouTube model can better capture the up and down of Bitcoin returns than the base model during the test periods but compared to the volatility of the returns, the two forecast are still far from the reality. However, we can suppose the forecast give a good indication on whether the market go up or down. If so, we can use it in a simple strategy. To verify this assumption, we compare the results of 3 trading strategies : the buy and hold strategy, the strategy using LSTM model without YouTube variables and the strategy using LSTM model with our YouTube Variables. We do not take into account the costs of the trade (trading fees of the exchange, slippage). For each day of the test dataset, we forecast the value of the next day. If the model predict a value superior to zero, we buy bitcoin for the day. The return of a strategy on the test period is calculated as follows:

$$RS = -1 + \prod_{d=d_1}^{d=d_n} X \tag{6}$$

Where :

$$X = \begin{cases} (1 + r_d) & P_d > 0 \\ 1 & P_d < 0 \end{cases} \tag{7}$$

r_d is the return of Bitcoin for the day d, P_d is the prediction of the model for the day d. n is the last date of the test dataset. d_1 is the first day of the test dataset, d_n is the last day of the dataset.

We plot the cumulative returns of the strategies in figure 7.

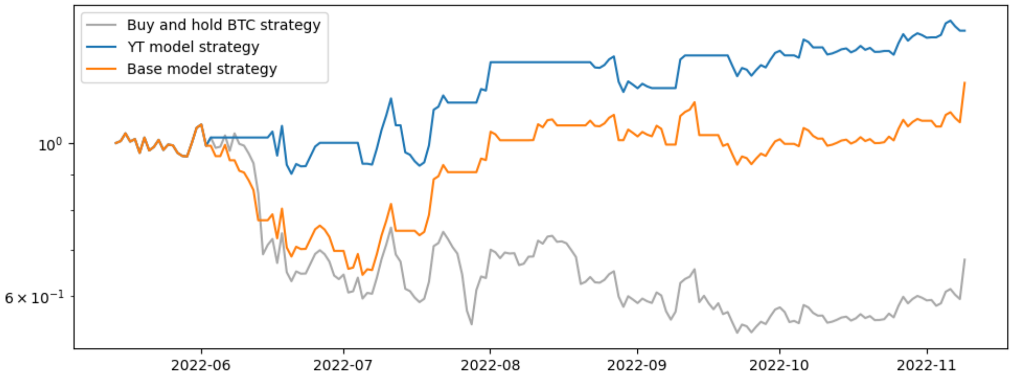

Figure 7: Cumulative returns of the strategies

BTC buy and hold strategy has a total returns of -32.3% on the test period. The strategy using LSTM with only past prices as input has 22.27% of returns. Our model with past prices and our YouTube variables achieve 45.61% returns on the test period. The LSTM models has been able to anticipate some big price fall and to avoid them successfully. We can see the LSTM achieve a better performance when it includes our YouTube variables.

4 Conclusion

This paper study whether data from investors behavior on YouTube could be useful to anticipate Bitcoin returns. In particular, we investigate the role played by investor attention and investor sentiment on YouTube on Bitcoin returns. Those analyses lead us to interesting findings which helps to better understand the dynamics between bitcoin returns and the sentiment of investors and between the attention of investors and bitcoin returns. In a first part, we investigated the relationship between investors attention and bitcoin returns, our results show the importance to classify our videos by subject when looking at these variable. We show that the attention on YouTube videos about "Hacks" has a significant negative causality relationship with Bitcoin returns. We also identify that attention on "tutorials" videos has a significant positive causality relationship with Bitcoin returns. Furthermore, we also identified a bidirectional causality relationship between videos about crypto "personalities" and Bitcoin returns. In a second part, we investigate the relationship between bitcoin returns and investors sentiment. We identified that the overall sentiment of YouTube videos, the sentiment about regulation and the sentiment about Price predictions have bidirectional positive causality relationships with Bitcoin returns. In the third part, we used these six variables to forecast future returns with a state-of-the-art LSTM model and found it was not enough to make a statistical improvement in the mean squared error compared to a LSTM model using only past prices, but we show that the forecast with YouTube variables can be used in a strategy to improve returns on the test. A future extension of the present study would be to add other variables to the model to improve the forecast and to test the ability YouTube to forecast intraday returns.

5 References

Araci, Dogu (Aug. 27, 2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. arXiv: 1908.10063[cs]. URL: http://arxiv.org/abs/1908.10063 (visited on 12/04/2022).

Baker, Malcolm P. and Jeffrey A. Wurgler (2003). "Investor Sentiment and the Cross-section of Stock Returns". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.464843. URL: http://www.ssrn.com/abstract=464843 (visited on 04/08/2022).

Barber, Brad and Terrance Odean (2008). "All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors". In: *Review of Financial Studies* 21.2, pp. 785–818. URL: https://EconPapers.repec.org/RePEc:oup:rfinst:v:21:y:2008:i:2:p:785-818.

Bollen, Johan, Huina Mao, and Xiao-Jun Zeng (Mar. 2011). "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1, pp. 1–8. ISSN: 18777503. DOI: 10.1016/j.jocs.2010.12.007. arXiv: 1010.3003. URL: http://arxiv.org/abs/1010.3003 (visited on 05/06/2020).

Bouoiyour, Jamal and Refk Selmi (2015). "What Does Bitcoin Look Like?" In: p. 39.

Bourghelle, David, Fredj Jawadi, and Philippe Rozin (Apr. 2022). "Do collective emotions drive bitcoin volatility? A triple regime-switching vector approach". In: *Journal of Economic Behavior & Organization* 196, pp. 294–306. ISSN: 01672681. DOI: 10.1016/j.jebo.2022.01.026. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167268122000397 (visited on 04/05/2022).

Ciaian, Pavel, Miroslava Rajcaniova, and d'Artis Kancs (Apr. 20, 2016). "The economics of BitCoin price formation". In: *Applied Economics* 48.19, pp. 1799–1815. ISSN: 0003-6846, 1466-4283. DOI: 10.1080/00036846.2015.1109038. URL: http://www.tandfonline.com/doi/full/10.1080/00036846.2015.1109038 (visited on 05/06/2020).

Corbet, Shaen, Charles Larkin, et al. (Sept. 21, 2020). "The impact of macroeconomic news on Bitcoin returns". In: *The European Journal of Finance* 26.14, pp. 1396–1416. ISSN: 1351-847X, 1466-4364. DOI: 10.1080/1351847X.2020.1737168. URL: https://www.tandfonline.com/doi/full/10.1080/1351847X.2020.1737168 (visited on 01/28/2022).

Corbet, Shaen, Brian Lucey, et al. (Mar. 2019). "Cryptocurrencies as a financial asset: A systematic analysis". In: *International Review of Financial Analysis* 62, pp. 182–199. ISSN: 10575219. DOI: 10.1016/j.irfa.2018.09.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1057521918305271 (visited on 04/08/2022).

Devlin, Jacob et al. (May 24, 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805[cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 12/08/2022).

Garcia, David et al. (Aug. 7, 2014). "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy". In: *arXiv:1408.1494 [nlin,* *physics:physics, q-fin]*. DOI: 10.1098/?rsif.2014.0623. arXiv: 1408.1494. URL: http://arxiv.org/abs/1408.1494 (visited on 05/06/2020).

Hutto, C.J. and Eric Gilbert (Jan. 1, 2015). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Journal Abbreviation: Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014 Publication Title: Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

Huynh, Toan Luu Duc (Jan. 13, 2022). "When Elon Musk Changes his Tone, Does Bitcoin Adjust Its Tune?" In: *Computational Economics*. ISSN: 0927-7099, 1572-9974. DOI: 10.1007/s10614-021-10230-6. URL: https://link.springer.com/10.1007/s10614-021-10230-6 (visited on 12/04/2022).

Leow, Edmund Kwong Wei, Binh P. Nguyen, and Matthew Chin Heng Chua (Oct. 2021). "Robo-advisor using genetic algorithm and BERT sentiments from tweets for hybrid portfolio optimisation". In: *Expert Systems with Applications* 179, p. 115060. ISSN: 09574174. DOI: 10.1016/j.eswa.2021.115060. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417421005017 (visited on 12/04/2022).

Liu, Yukun and Aleh Tsyvinski (May 21, 2021). "Risks and Returns of Cryptocurrency". In: *The Review of Financial Studies* 34.6. Ed. by Itay Goldstein, pp. 2689–2727. ISSN: 0893-9454, 1465-7368. DOI: 10.1093/rfs/hhaa113. URL: https://academic.oup.com/rfs/article/34/6/2689/5912024 (visited on 05/24/2021).

Livieris, Ioannis E., Emmanuel Pintelas, and Panagiotis Pintelas (Dec. 2020). "A CNN–LSTM model for gold price time-series forecasting". In: *Neural Computing and Applications* 32.23, pp. 17351–17360. ISSN: 0941-0643, 1433-3058. DOI: 10.1007/s00521-020-04867-x. URL: https://link.springer.com/10.1007/s00521-020-04867-x (visited on 12/08/2022).

Loginova, Ekaterina et al. (Nov. 18, 2021). "Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data". In: *Machine Learning*. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-021-06095-3. URL: https://link.springer.com/10.1007/s10994-021-06095-3 (visited on 01/29/2022).

Long, Cheng, Brian M. Lucey, and Larisa Yarovaya (2021). ""I Just Like the Stock" versus "Fear and Loathing on Main Street" : The Role of Reddit Sentiment in the GameStop Short Squeeze". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3822315. URL: https://www.ssrn.com/abstract=3822315 (visited on 01/29/2022).

Mai, Feng, Qing Bai, and Jay Shan (2015). "The Impacts of Social Media on Bitcoin Performance". In: *Thirty Sixth International Conference on Information Systems, Fort Worth*, p. 16.

Mishev, Kostadin et al. (2020). "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers". In: *IEEE Access* 8, pp. 131662–131682. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3009626. URL: https://ieeexplore.ieee.org/document/9142175/ (visited on 12/04/2022).

Nasir, Muhammad Ali et al. (Dec. 2019). "Forecasting cryptocurrency returns and volume using search engines". In: *Financial Innovation* 5.1, p. 2. ISSN: 2199-4730. DOI: 10.1186/s40854-018-0119-8. URL: https://jfin-swufe.springeropen.com/articles/10.1186/s40854-018-0119-8 (visited on 05/06/2020).

Nelson, David M. Q., Adriano C. M. Pereira, and Renato A. de Oliveira (May 2017). "Stock market's price movement prediction with LSTM neural networks". In: *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017 International Joint Conference on Neural Networks (IJCNN). Anchorage, AK, USA: IEEE, pp. 1419–1426. ISBN: 978-1-5090-6182-2. DOI: 10.1109/IJCNN.2017.7966019. URL: http://ieeexplore.ieee.org/document/7966019/ (visited on 12/08/2022).

Nofsinger, John R. (Sept. 2005). "Social Mood and Financial Economics". In: *Journal of Behavioral Finance* 6.3, pp. 144–160. ISSN: 1542-7560, 1542-7579. DOI: 10.1207/s15427579jpfm0603_4. URL: http://www.tandfonline.com/doi/abs/10.1207/s15427579jpfm0603_4 (visited on 04/08/2022).

Ortu, Marco et al. (Mar. 2022). "Cryptocurrency ecosystems and social media environments: An empirical analysis through Hawkes' models and natural language processing". In: *Machine Learning with Applications* 7, p. 100229. ISSN: 26668270. DOI: 10.1016/j.mlwa.2021.100229. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666827021001158 (visited on 01/28/2022).

Philippas, Dionisis (2019). "Media Attention and Bitcoin Prices". In: p. 12. DOI: https://doi.org/10.1016/j.frl.2019.03.031. URL: https://www.sciencedirect.com/science/article/abs/pii/S1544612319300558?via%3Dihub.

Phillips, Ross C. and Denise Gorse (2018). "Mutual-Excitation of Cryptocurrency Market Returns and Social Media Topics". In: *Proceedings of the 4th International Conference on Frontiers of Educational Technologies - ICFET '18*. the 4th International Conference. Moscow, Russian Federation: ACM Press, pp. 80–86. ISBN: 978-1-4503-6472-0. DOI: 10.1145/3233347.3233370. URL: http://dl.acm.org/citation.cfm?doid=3233347.3233370 (visited on 06/21/2021).

Shleifer, Andrei and Lawrence H Summers (May 1, 1990). "The Noise Trader Approach to Finance". In: *Journal of Economic Perspectives* 4.2, pp. 19–33. ISSN: 0895-3309. DOI: 10.1257/jep.4.2.19. URL: https://pubs.aeaweb.org/doi/10.1257/jep.4.2.19 (visited on 04/08/2022).

Susarla, Anjana, Jeong-Ha Oh, and Yong Tan (Mar. 2012). "Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube". In: *Information Systems Research* 23.1, pp. 23–41. ISSN: 1047-7047, 1526-5536. DOI: 10.1287/isre.1100.0339. URL: http://pubsonline.informs.org/doi/abs/10.1287/isre.1100.0339 (visited on 01/29/2022).

Tetlock, Paul C. (June 2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market". In: *The Journal of Finance* 62.3, pp. 1139–1168. ISSN: 00221082. DOI: 10.1111/j.1540-6261.2007.01232.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2007.01232.x (visited on 01/28/2022).

Tran, Vu Le and Thomas Leirvik (Nov. 2019). "Efficiency in the markets of cryptocurrencies". In: *Finance Research Letters*, p. 101382. ISSN: 15446123. DOI: 10.1016/j.frl.2019.101382. URL: https://linkinghub.elsevier.com/retrieve/pii/S1544612319310438 (visited on 05/06/2020).

Uras, Nicola, Stefano Vacca, and Giuseppe Destefanis (June 27, 2020). "Investigation of Mutual-Influence among Blockchain Development Communities and Cryptocurrency Price Changes". In: *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*. ICSE '20: 42nd International Conference on Software Engineering. Seoul Republic of Korea: ACM, pp. 779–782. ISBN: 978-1-4503-7963-2.

DOI: 10.1145/3387940.3392245. URL: https://dl.acm.org/doi/10.1145/3387940.3392245 (visited on 01/28/2022).

Urquhart, Andrew (2016). "The Inefficiency of Bitcoin". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.2828745. URL: https://www.ssrn.com/abstract=2828745 (visited on 02/04/2021).

– (May 2018). "What causes the attention of Bitcoin?" In: *Economics Letters* 166, pp. 40–44. ISSN: 01651765. DOI: 10.1016/j.econlet.2018.02.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S016517651830065X (visited on 05/06/2020).

Wang, Chen, Dehua Shen, and Youwei Li (Oct. 2022). "Aggregate Investor Attention and Bitcoin Return: The Long Short-term Memory Networks Perspective". In: *Finance Research Letters* 49, p. 103143. ISSN: 15446123. DOI: 10.1016/j.frl.2022.103143. URL: https://linkinghub.elsevier.com/retrieve/pii/S154461232200366X (visited on 12/04/2022).