Machine Learning with Applications 7 (2022) 100229

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Cryptocurrency ecosystems and social media environments: An empirical analysis through Hawkes' models and natural language processing

Marco Ortu [a,*], Stefano Vacca [b], Giuseppe Destefanis [c], Claudio Conversano [a]

[a] Department of Business and Economics Sciences, University Of Cagliari, Viale Fra Ignazio 17, Cagliari, Italy
[b] Eustema S.p.a., Via Carlo Mirabello, 7, Roma, Italy
[c] Department of Computer Science, Brunel University, Uxbridge, Middlesex UB8 3PH, London, UK

ARTICLE INFO

Keywords:
Cryptocurrencies
Social media analysis
Fundamental analysis
Forecasting price movements
Hawkes model

ABSTRACT

We analyse, using a mixture of statistical models and natural language process techniques, what happened in social media from June 2019 onwards to understand the relationships between Cryptocurrencies' prices and social media, focusing on the rise of the Bitcoin and Ethereum prices. In particular, we identify and model the relationship between the cryptocurrencies market *price changes*, and *sentiment* and *topic discussion occurrences* on social media, using Hawkes' Model. We find that some topics occurrences and rise of sentiment in social media precedes certain types of price movements. Specifically, discussions concerning governments, trading, and Ethereum cryptocurrency as an exchange currency appear to negatively affect Bitcoin and Ethereum prices. Those concerning investments, appear to explain price rises, whilst discussions related to new decentralized realities and technological applications explain price falls. Finally, we validate our model using a real case study: the already famous case of "Wallstreetbet and GameStop"[1] that took place in January 2021.

1. Introduction

Cryptocurrencies stir intense interest in the scientific and financial disciplines and within social media communities, making the analysis of their price movements one of the most discussed topics of the last few years, see Kyriazis (2019), Zheng et al. (2018). Even if the number of studies related to cryptocurrencies price forecasting increases, determinants of their price behaviours are still mostly unexplored, and the knowledge about predicting their price movements is still limited.

This paper builds on a previous work Uras et al. (2020) to ascertain possible relationships between cryptocurrency market prices and social media discussions and understand what topics have a higher potential to predict price movements. It is well known that developers' moods can affect software quality Ortu et al. (2015), and if this background knowledge is applied in the field of cryptocurrency software production and their quality metrics (Destefanis et al., 2017), this may affect cryptocurrency market prices as well. We first introduce in Section 4 the analysis of the significant case of the launch of Libra [2] cryptocurrency in 2019 and the influence of the subsequent online discussions on the cryptocurrencies' markets.

This example opens the road for a more in-depth analysis in Section 5 where we retrieved the discussions and comments from the social media platform *Reddit*, which is one of the most valuable sources of information related to cryptocurrency markets (Bartolucci et al., 2020). We model these online discussions using the Hawkes Model to understand the mutual influence of online technical discussion of the two leading cryptocurrencies (Bitcoin and Ethereum). We continue on this direction in Section 6, where we analyse the occurrence of particular topics from social media content through dynamic topic modelling, that is an extension of Latent Dirichlet Allocation (LDA), along with emotional features extracted from comments, and again we apply the Hawkes model to identify possible hidden interactions between these features and cryptocurrency market prices.

The key contributions of our study are the following:

- Deciphering an hidden connection among crypto-markets and social media;
- Identification of a semantic model of occurrences (based on topic discussion occurrences) deriving from mapping the signs in signals;
- Design of a cost-effective solution for a real-time alarm system that can be used to support investors' decisions;
- Specification of a unique mixture of natural language processing, statistical model and pre-existing tools to promote and validate the research hypothesis;

* Corresponding author.
E-mail addresses: marco.ortu@unica.it (M. Ortu), s.vacca@eustema.it (S. Vacca), giuseppe.destefanis@brunel.ac.uk (G. Destefanis), conversa@unica.it (C. Conversano).

[1] https://www.economist.com/finance-and-economics/2021/02/06/how-wallstreetbets-works.
[2] https://www.diem.com/en-us/.

https://doi.org/10.1016/j.mlwa.2021.100229
Received 6 October 2021; Received in revised form 24 November 2021; Accepted 2 December 2021
Available online 8 December 2021

- Model validation using a real case study: the ”WallstreetBets VS GameStop”. We found that our model is able to detect warning signals of imminent financial distress.

The most remarkable contribution of our analysis is the specification and implementation of an alarm system capable of highlighting warning events that can be considered precursors of financial distress in the cryptocurrencies markets.

The proposed *social media seismograph* will translate the warning signals from the Hawkes model into a direct and user-friendly format that will be able to communicate real-time information to the final user.

2. Related works

Cryptocurrency markets are in many aspects similar to stock markets, and links with social media are even more robust (Keskin & Aste, 2019), and some economists even compared the cryptocurrencies market to the gold market (Al-Yahyaee et al., 2018). Over the years, several approaches related to forecasting cryptocurrency price movements have been developed (Bartolucci et al., 2020; L. Cocco, 2019a, 2019b). McNally et al. tried to predict with the highest possible accuracy, achieving 52% and a RMSE of 8%, the directions of Bitcoin prices in USD using machine learning algorithms like LSTM (Long short-term memory) and RNN (Recurrent Neural Network) (S. McNally, 2018). Naimy and Hayek tried to forecast the Bitcoin/USD exchange rate volatility using GARCH (Generalized AutoRegressive Conditional Heteroscedasticity) models (V. Y. Naimy, 2018). Numerous studies tried to use online information (including social media topics discussions) to predict cryptocurrencies price changes. For example, Google searches for Bitcoin-related terms have been shown to have a relationship with the Bitcoin price (Kristoufek, 2013). Garcia and Schweitzer have considered the strength and polarization of opinions displayed on Twitter. They show that an increase in the polarization of sentiment (disagreement of sentiment) anticipates a rise in the price of Bitcoin (Garcia & Schweitzer, 2015). In another work, several machine learning pipelines were implemented to identify cryptocurrency market movements to prove whether Twitter data relating to cryptocurrencies can be utilized to develop promising crypto coin trading strategies (Stuart G. Colianni, 2015). R. C. Phillips (2017) monitored the activity on the social media platform Reddit to detect the epidemic-like spread of investment ideas beneficial in the prediction of cryptocurrency price bubbles.

Other studies highlighted the potential prediction power of social media features on cryptocurrencies markets (Bartolucci et al., 2020) while Phillips and Gorse (2018) showed that particular topics tend to precede certain types of price movements, for example the discussion of ‘risk and investment vs trading’ being indicative of price falls, the discussion of ‘substantial price movements’ being indicative of volatility, and the discussion of ‘fundamental cryptocurrency value’ by technical communities being indicative of price rises. Thanks to all the works that helped prove possible association relationships between the cryptocurrency price changes and social media, we can state that the discussion topics’ knowledge that affects prices seems to be a useful component of a successful trading model.

2.1. Limitation of current literature

The proposed approaches lack, in general, a global vision of the heterogeneous factors influencing cryptocurrencies markets, focusing on specific aspects such as the financial time series, social media, media websites, Google trends (Wołk, 2020). Most of the proposed approaches are also missing a practical validation in the field. We overcome these limitations by modelling most of these factors using Hawkes’ models to decipher relationships among social media, cryptocurrencies markets, media websites and financial data. We tested our model on a real case scenario, showing the practical application of the proposed approach and the designed system.

2.2. Statement of purpose

Starting with the current research, we hypothesized that social media contains sufficient information on causal relationships:

- Between topic discussion occurrences on social media and cryptocurrencies market price changes
- Among different Blockchain communities discussions which influence one with another.
- Between prices of Cryptocurrencies and sentiment arising from social media discussion’s groups.

These relationships can be aggregated into a monitoring system with the purpose of *warning system* of possible incoming financial distresses. We tested this hypothesis modelling the occurrences of rise/fall of topics concerning cryptocurrencies, the occurrences of rise/fall of sentiment and emotions in social media, and the occurrences of cryptocurrencies’ price rise/fall using Hawkes’ model. We found consistency with previous studies, in particular, sentiment and emotions occurrences in social media (Bartolucci et al., 2020) along with specific discussion topics (Phillips & Gorse, 2018) help decipher users behaviours and reactions to certain information regarding future cryptocurrencies market trends, and causing further price changes.

2.3. Practical implication: Implementation of a social warning system

We propose a *social seismograph* to help decipher users behaviours and reactions to certain information regarding future market trends and causing further price changes. In particular, Hawkes’ models could be used in a **real-time alarm system**, as causal relationships start to emerge, a digital dashboard light up imitating an alarm system, where warning lights begin to change colours or blinking, highlight possible critic events. Fig. 1 shows the conceptual architecture of such a system where the output of the Hawkes process are mapped into the real-time decision support system, which translates the information into a direct, informative and visual dashboard.

2.4. Context

The context of the present study is represented by the social media environments around the cryptocurrencies ecosystems. These two environments are constantly growing in terms of users involved and the volumes of investments. Fig. 2 summaries the main concept involved in our study.

Cryptocurrencies, in many cases, and especially the ones considered in this study, are supported by open source communities that contribute to the development and maintenance of the underlying technology. The research in software engineering on online open source communities is hence suitable for the context of the present study.

Indeed, recent studies in software engineering have been conducted considering the human aspects of software development in online open source communities, focusing the attention on a better understanding of how developers interact with each other, studying their emotions and affects.

Numerous tools to extract emotions and affects measures, specifically in the software engineering domain, have also been developed. For example, Murgia et al. (2017) demonstrated the feasibility of a machine learning classifier using emotion-driving words and technical terms to identify developers’ comments containing gratitude, joy and sadness.

We exploit these tools to extract emotional content from comments expressed by technical users in specialized social media discussion groups on cryptocurrencies.

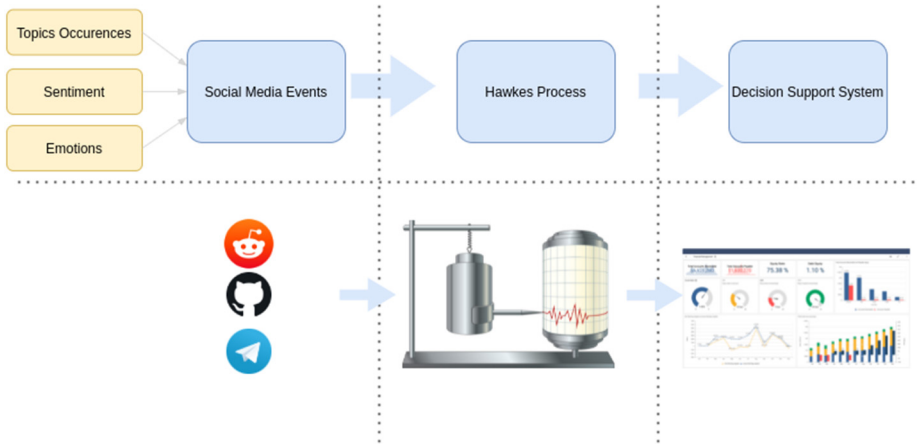


Fig. 1. High-Level conceptual architecture.

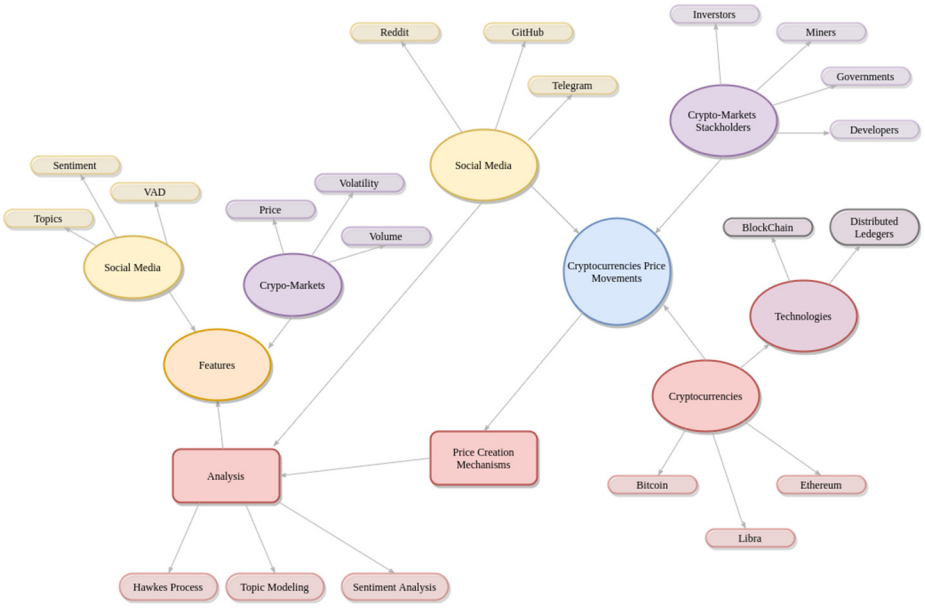


Fig. 2. Conceptual Context Diagram.

3. Background

This section introduces the basics of the two main techniques we used to analyse the relationship between cryptocurrencies market price changes and topic discussion occurrences on social media: Topic Modelling and Hawkes Process Model.

3.1. Topic modelling

A topic model is a specific statistical model used to identify the abstract topics within a collection of documents. Topic modelling is a frequently used text-mining tool for automatically identifying themes within a corpus, finding the distribution of words in each topic, and topics in each document. In this work, we use *Latent Dirichlet Allocation* (LDA) (D. Blei & Jordan, 2003), a popular unsupervised learning technique for topic modelling. This type of topic model assumes each document contains multiple topics to different extents. In the following, we briefly discuss the generative process by which LDA assumes each document originates.

- The first step is to choose, for each document, the number of words N to generate.
- The process then randomly chooses a distribution over topics. This parameter is usually labelled as θ.
- Finally, for each word to be generated in the document, the process randomly chooses a topic, Z_n, from the distribution of topics, and from that topic chooses a word, W_n, using the distribution of words in the topic.

If we consider a given document d and topic t, the variables of interest in this model are the distribution of topic t in document d and the distribution of words in topic t. These variables are latent, hidden parameters that can be estimated via inference for any specific dataset. It has been proved that the standard LDA model cannot understand both the ordering of words within a document and the ordering of documents within a corpus. For this reason, an extension of this model was developed by Blei and Lafferty (2006). This extended LDA model is known as *dynamic topic model*, and even if it still has no understanding of the order of words in a document, at least the order of documents in the corpus is taken into account.

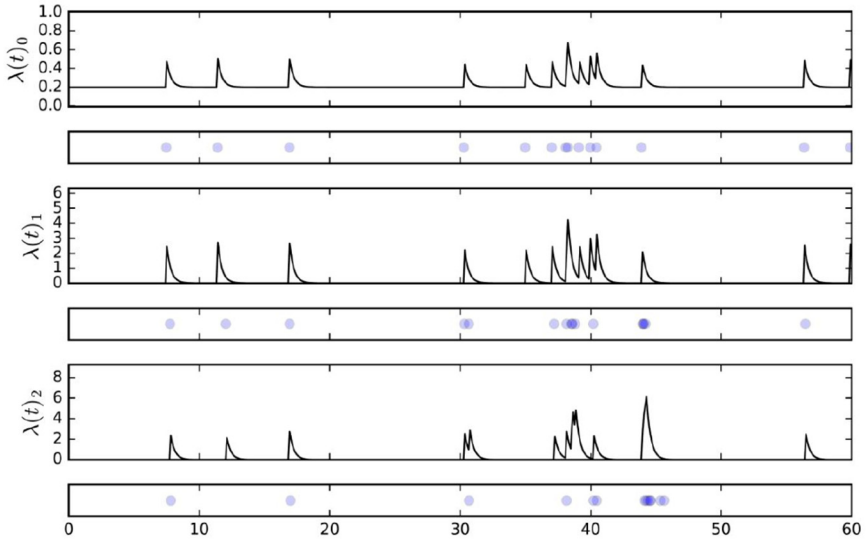


Fig. 3. Example of multivariate Hawkes process.

3.2. Hawkes model

The Hawkes process is a point process class (Neuts, 1979), also known as a self-exciting counting process, in which the impulse response function explicitly depends on past events (Hawkes, 1971). In this type of process, the observation of an event causes the increase of the process impulse function. From a mathematical point of view, a point process is a Hawkes process if the impulse function $\lambda(t|H_t)$ of the process takes the form of (1).

$$\lambda(t|H_t) = \lambda_0(t) + \sum_{i:t_i<t} \phi(t - t_i) \tag{1}$$

In Eq. (1) H_t represents the history of given past events, $\lambda_0(t)$ is a positive function that determines the basic intensity of the process, and ϕ is another positive function known as *memory kernel*, since it depends on past events occurred before time t. Hawkes models can be used to identify the dynamics of interactions between a group of K processes. The occurrence of an event on a particular process can cause an impulse response on that process (self-excitation), determining an increase of the likelihood of other events and on other processes (mutual-excitation). Thus, given a set of events occurring on several processes, a Hawkes model can be used to quantify previously hidden connections between the processes.

In this work, we apply a Hawkes model to decipher how topics are related to one another and how price changes are related to the topic occurrence and sentiment and emotion expressed in social media comments.

Fig. 3 illustrates an explanatory example of a multivariate Hawkes process with three flows of events: $\lambda(t)_0$, $\lambda(t)_1$ and $\lambda(t)_2$. In this example, the event flows have been constructed so that $\lambda(t)_0$ is not influenced by other flows, but only by events that happen on its own flow (self-exciting effect); otherwise, events in the same $\lambda(t)_0$ can have effects in $\lambda(t)_1$ and $\lambda(t)_2$ (mutual-exciting effect).

We can consider these flows of events as follows:

- $\lambda(t)_0$: as a price movement flow of events (up and down movements per time interval).
- $\lambda(t)_1$: as the occurrences of a specific topic in social media.
- $\lambda(t)_2$: as the occurrences of positive comments in social media.

This example helps to understand how the Hawkes' processes can model the many events happening simultaneously in both social media and Crypto-Markets and quantify hidden connections between the processes.

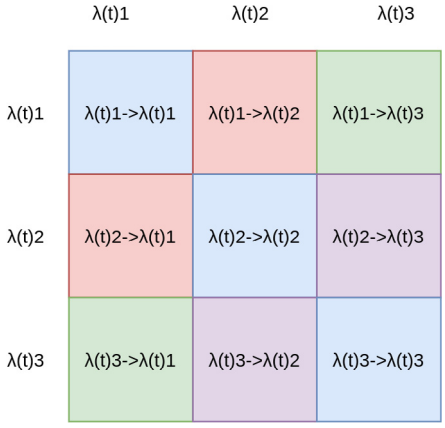


Fig. 4. Example of Hawkes Coefficient Matrix.

Once the Hawkes model is fitted on data, it will contain some weights (in a matrix-wise fashion) representing the directional strength of any interaction between processes interpreted as the expected number of events on a specific process resulting from an event on another process.

Fig. 4 represents the Hawkes coefficient matrix, fitted with the hypothetical data from the previous example. This matrix represents the coefficients of the fitted model: along the diagonal we have the coefficients for the *self-excitation* and the coefficients outside the diagonal represent the *mutual-excitation*. We used the right arrow symbol "↛" to highlight the *direction* of the relationship, i.e. $\lambda(t)_1 \nrightarrow \lambda(t)_2$ coefficient represents the strength of the relationship of $\lambda(t)_1$ events on $\lambda(t)_2$ and $\lambda(t)_2 \nrightarrow \lambda(t)_1$ vice versa.

4. Cryptocurrencies and sentiment analysis: a case study on Libra

The cryptocurrency market is volatile, and its performance is influenced by information from various sources. Since they are based on different blockchain technologies, cryptocurrencies are candidates to be complementary currencies to the current fiat currencies or, shortly, to replace them.

The Blockchain is a technology, and as such, we can study it through Rogers' innovation adoption model. According to Rogers (2010), all technologies go through a first phase in which only the "Innovators"

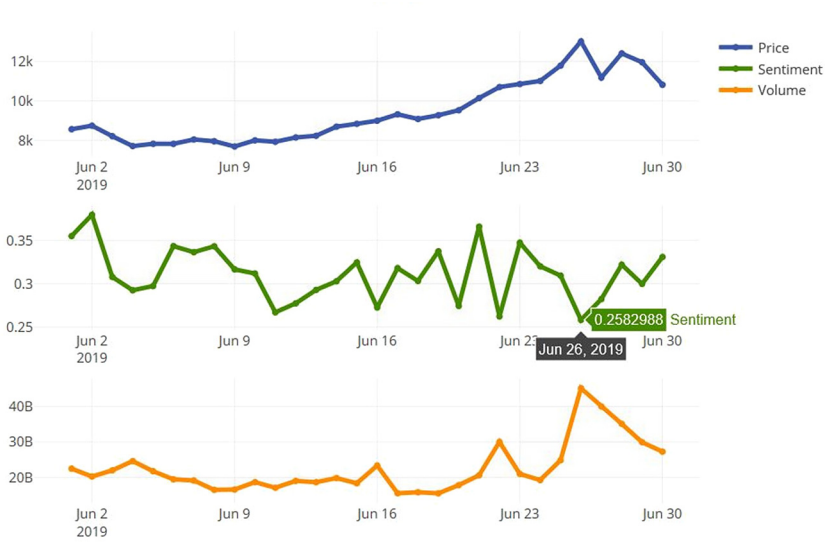


Fig. 5. Bitcoin trend — June 2019.

(about 2.5%, this is the case of the period before autumn 2017) and the "First acquirers" adhere (13.5%). After that, however, to establish itself in the market, a technology must overcome what Rogers calls "Chasm" (i.e. ravine), which identifies the phase in which the technology is adopted by a more significant market segment, called "early-adopters". Nowadays, especially on Telegram, users use cryptocurrencies as currency to trade and make money buying and selling. There are also the so-called "Pioneers", who have held Bitcoins since before the 2017 bubble and are not interested in making direct profits but believe in the project and the future inclusion in society. It is no coincidence that the major IT companies (e.g., Facebook) are creating their cryptocurrencies, thus laying the foundations for several mechanisms such as collaboration, win-to-win and competition (or conflict), on which virtual currencies will become established and those that will disappear. It is not the first time in history that more coins or currencies coexist. What seems inevitable is that Bitcoin is having more and more success and spread throughout the world, and the technology that supports it is now the future of society. Given the broad ecosystem of cryptocurrencies, we can understand if and how different blockchain's communities influence each other, especially from the software implementation and design.

In 2017 Facebook announced its intention "to reach 1.7 billion people in the world who do not yet have a bank account". Facebook, which already holds personal data of the profiles of 2.23 billion monthly active users, will seek to obtain information related to financial trends. As stated by the company, the cryptocurrency is independent but controlled by the Libra Association (an association based in Switzerland), intending to regulate and validate transactions related to social funds. In support of the association, there are several small tech companies including PayPal, eBay, Spotify, Uber and Lyft, and financial companies and venture capitals such as Andreessen Horowitz, Thrive Capital, Visa and Mastercard.

In this illustrative case, we analyse the technical discussion of two blockchain developer's communities, Bitcoin and Libra, from two different public platforms: Reddit and Telegram. We use text mining techniques to understand whether the announcement of Libra influenced Bitcoin's price variations.

4.1. Reddit

The social media platform *Reddit* is an American social news aggregation, web content rating, and discussion website that reaches about 8 billion page views per month. It is a top-rated social network in English-speaking countries, especially Canada and the United States. Almost all the messages are written in English, while the minority are Spanish, Italian, French, and German. Reddit is built over multiple subreddits, where each subreddit is dedicated to discussing a particular subject. Therefore, there are specific subreddits related to major cryptocurrency projects.

Our data collection led to the extraction of 9453 messages, 713 different discussions and 5507 unique users who talked about Libra. We carried out the research by investigating the keyword **Libra**. The term **Bitcoin** instead had much larger volumes, thanks to the widespread diffusion and reputation of the technology. The collected messages were 68,243, the discussions 2625 and the total number of users was 22,799.

We used the Vader Hutto and Gilbert (2014) tool to extract the Sentiment from these comments. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media. Fig. 5 shows Sentiment (the central trend in green), price (upper part in blue) and volume of the transactions (lower orange) for Bitcoin in June 2019. June 26th 2019 was when Bitcoin prices rose the most, indicating an excellent opportunity to sell for those who already owned Bitcoin, but a great disappointment for those who did not buy it previously, thus losing the chance to profit from a possible sale. Furthermore, the Sentiment recorded a sharp decline, bringing the lowest value of the month. This is confirmed by the fact that the volume of trade has increased disproportionately, inferring that most investors have sold the cryptocurrency on that date.

In Fig. 6, the trend of Libra's Sentiment is represented with the same type of graph shown in Fig. 5. In this case, it is essential to note that the Sentiment recorded a remarkable price rise on June 14th, the day of the disclosure of the news that would soon see the entry into the market of Facebook's cryptocurrency. This graph does not show a definite direct correlation with the price of Bitcoin, but this does not imply, however, that the news did not contribute to increasing the value of Bitcoin.

As for now, beyond Bitcoin, there are few "virtual" currencies to which investors can rely on. However, Libra could establish itself as a reliable currency because it is kept alive by a large company that cares for the image and aims to make it last over time. Despite the scandal that engulfed Facebook in May 2018, it has shown that it is robust and continuously investing in the market. These features make Libra more reliable than other currencies, born without basic programming and much more vulnerable to speculative attacks. Since Facebook's mission is to create a stable currency, many people claim that Libra will be the new Tether, a virtual currency linked to the US dollar, which has always been at the centre of accusations and scandals.

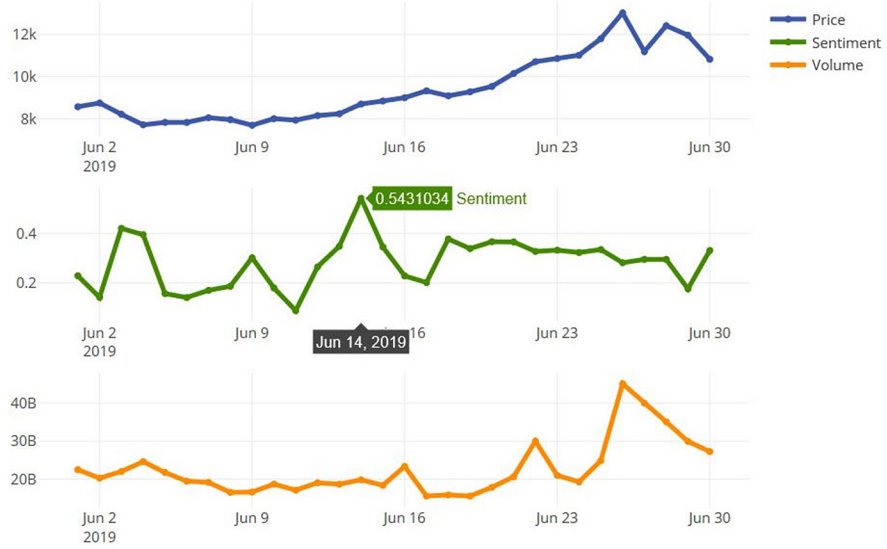
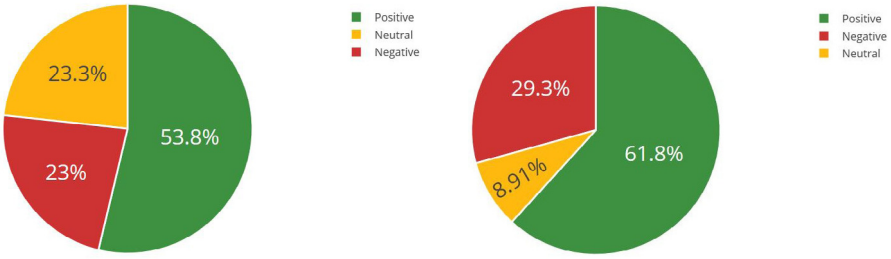


Fig. 6. Libra trend — June 2019.



(a) Bitcoin Sentiment on June 2019 (b) Libra Sentiment on June 2019

Fig. 7. Comparison of sentiment in June 2019 between Bitcoin and Libra.

Fig. 7(a) shows the sum up of the percentage of comments classified as positive, negative and neutral during June 2019 for Bitcoin. The Sentiment is positive in 53.8% of comments, while negative and neutral are balanced at around 23%.

Fig. 7(b) shows the sum up of the percentage of the comments classified as positive, negative and neutral during June 2019 for Libra. Contrarily, in this case, the Sentiment is much more unbalanced, with the neutral comments being less than 10%, while positive comments account for the 61.8% and negative ones for the 29,3%.

4.2. Telegram communities

We extracted the messages written by users of some specific *communities* on Telegram. These communities are public chat groups in which people with particular interests discuss in real-time. For this reason, compared to Reddit users, they are much more critical and sensitive to market changes and are more likely to seek higher profits. The considered Telegram communities were the following:

- Bad Crypto Podcast,
- The Coin Farm,
- Ripple Group (XRP),
- WCSE RA TALKS.

Fig. 8 shows Sentiment's performance for each of the considered groups during June 2019. The trend is mostly negative, but with positive peaks, probably due to the excellent performance of Bitcoin on the market.

We classified the groups with more significant influence and those with the trend of the common Sentiment. Fig. 9 shows the Pearson correlation matrix between the four groups and the Bitcoin closing price trend. We built the matrix by grouping the days of positive Bitcoin growth (i.e., from June 21st, until the peak of the month — June 26, 2019). A correlation coefficient close to +1 denotes a direct (or positive) correlation, so the two variables under analysis have the same type of linear trend. A correlation coefficient close to −1 denotes an indirect (or negative) correlation and indicates that the two variables have an opposite trend; finally, a correlation equal to zero indicates that the two variables do not exhibit a linear correlation.

The analysis shows that the price of Bitcoin is highly positively correlated with all the considered groups except for *Ripple Group (XRP)*, which has a robust negative correlation equal to −0.71. This data is not accidental and explains how this is the only group in our analysis that does not appreciate Bitcoin's price growth. The correlation analysis further confirms this on June 27th, when Bitcoin's price began its descent. In this case, the Sentiment is on the rise because of the decline in Bitcoin prices.

4.3. Topic analysis

The *Topic modelling analysis* for Libra communities is shown in Fig. 11, which shows three well separated and rater different *topics*. The word **Libra** appears in the first topic, representing the most important context of the discussions from the 21st to the 26th of June. This suggests that the users' concern was linked to the price of the Ripple

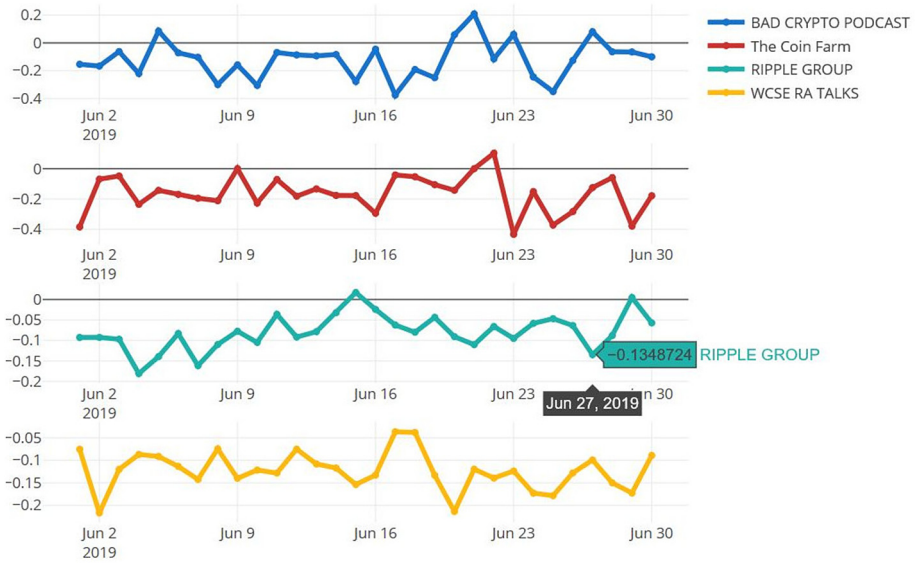


Fig. 8. Sentiment Analysis on a specific Community Telegram on June 2019.

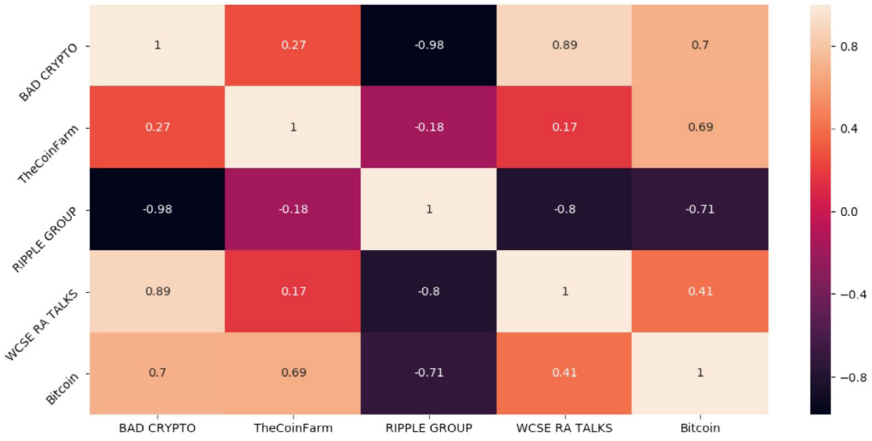


Fig. 9. Correlation Matrix between Sentiment Bitcoin price on June 2019.

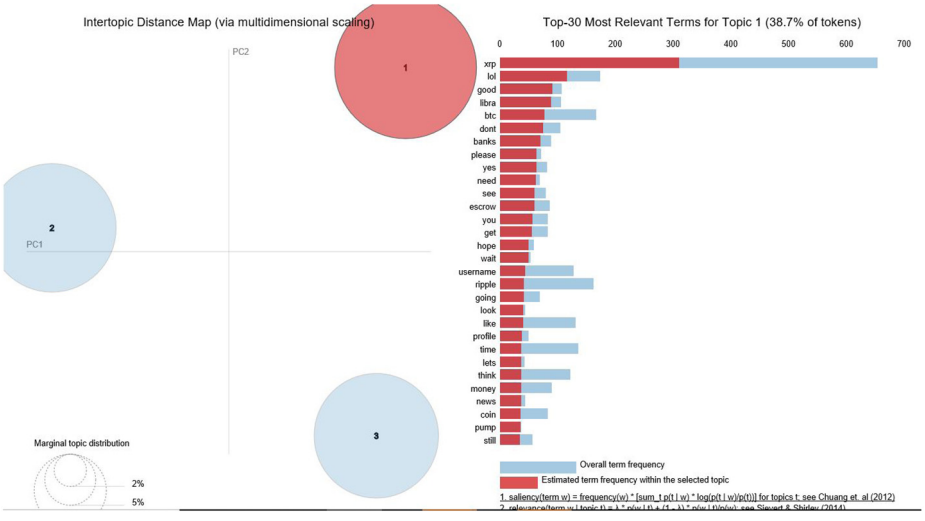


Fig. 10. Latent Dirichlet Allocation per Ripple Group XRP Telegram.

cryptocurrency and the Libra's recent information. Fig. 10 shows the *topic analysis* with words which appear most frequently in the same word corpus or document. For example, the analysis carried out for the keyword Libra, shows that words like *Facebook*, *Bitcoin*, *money*,

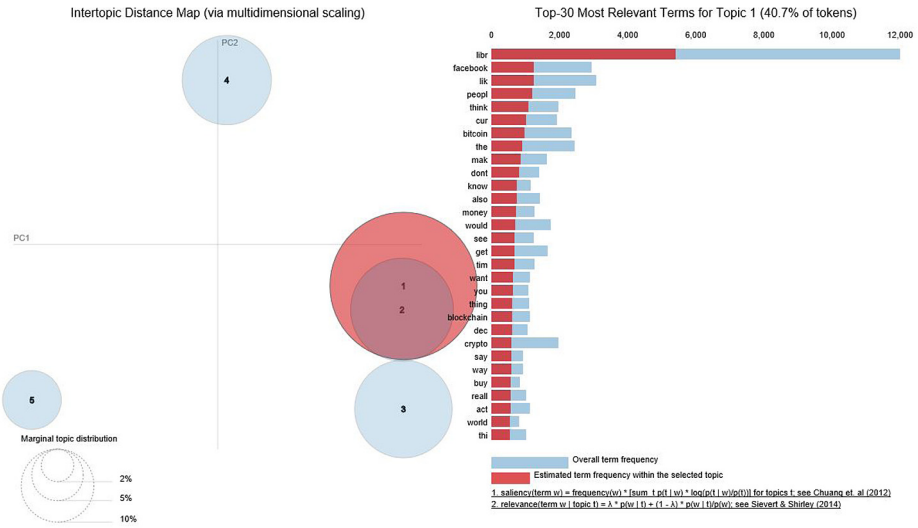


Fig. 11. Latent Dirichlet Allocation for Libra.

Blockchain and the same *Libra*, are often used together in the same context.

4.4. Discussion on Libra and Bitcoin

Our preliminary analysis showed how the price of Bitcoin is highly positively correlated with the considered discussion's groups, although limited to one month, this study shows how Blockchain development communities influence each other from Sentiment expressed in technical discussions.

Further analysis will involve more sophisticated methodologies, such as Phillips and Gorse (2018), to deeper understand how discussions in different Blockchains communities influence each other and influence the cryptocurrencies' price (Bartolucci et al., 2020). In Sections 5 and 6 we apply these methodologies.

5. Cryptocurrencies' communities mutual influence: A case study on Bitcoin and Ethereum

In this section, we analyse online discussion comments from the social media platform *Reddit* using LDA topic modelling and Hawkes models. Hawkes models are applied to these online discussions to understand the mutual influence of online technical discussion of the two leading cryptocurrencies, Bitcoin and Ethereum. We first introduce some information about the data sources and their processing achieved applying dynamic topic modelling approach (Blei & Lafferty, 2006). Secondly, we applied the Hawkes models to these discussions' topics to understand the mutual influences between online discussion groups of Bitcoin and Ethereum.

5.1. Data sources

Reddit is built over multiple subreddits, where each subreddit is dedicated to discussing a particular subject. Therefore, there are specific subreddits related to major cryptocurrency projects. For each considered cryptocurrency, two subreddits are analysed, one technical and one trading related. They are mentioned in Table 1.

For each subreddit, we fetched a given amount of comments for almost one million comments analysed. The historical prices of Bitcoin and Ethereum were extracted from the "Historical Data" section available on *Crypto Data Download* website, specifically from the *Coinbase* trading exchange. The hourly prices time series were retrieved, stored and then aggregated to the required granularity. The sample period considered in this work is one year, from January 1st 2019 to December 31st 2019. The chosen data period appears to be suitable since in September 2019, a significant fall in the Bitcoin price occurred with consequent ripple effects on Ethereum prices, allowing us to investigate the interaction between prices and social media during this considered period.

Table 1
Considered subreddits.

Cryptocurrency	*Technical Discussions*	*Trading Discussions*
Bitcoin	r/Bitcoin	r/BitcoinMarkets
Ethereum	r/Ethereum	r/EthTrader

5.2. Results

Before applying topic modelling, the corpus has been pre-processed. Therefore, topics were obtained removing stop words (such as "the"), links, special characters, and varied punctuation. We used part-of-speech (POS) tagging to categorize words into types; nouns and adjectives are maintained while other types are removed. Furthermore, we applied stemming techniques to reduce derived words to their base root. These techniques allow grouping different terms into one unique root term and simplify the number of features to increase attention to the most critical terms. After this data pre-processing step, we applied Latent Dirichlet Allocation and, in particular, the *LdaModel* method provided by the *Gensim* python library (R. Rehurek, 2010) to identify distinct topics through topic modelling technique, thus generating a time series of topic occurrences. For insight into this topics' selection process, Table 2 shows all topics selected for their coherent cryptocurrency-related content. For the sake of brevity, we only report those for the *r/Bitcoin* subreddit.

The chosen topics are then analysed in a Hawkes model, alongside market prices. We used the *HawkesConditionalLaw* method provided by the *tick* Python library (Bacry et al., 2017). Once we created the topics, the creation of the features in events and processes was performed. We aggregated data into groups composed of events happening in a time interval of sixty-minutes ($\Delta_t = 60$ min).

Two features were processed from each cryptocurrency prices, namely the $delta\_price$ feature, which is the difference between the closing price and the hourly opening price, and the log_return feature, i.e. the logarithm of the difference between higher price and hourly lower price. We then considered a total of fourteen features, five topics and two price features for each cryptocurrency. The max_lag parameter models the maximum time for which an individual event can affect

Fig. 12. Hawkes matrix for *r/Bitcoin* and *r/Ethereum* with *max_lag* 24.

Fig. 13. Hawkes matrix for *r/Bitcoin* and *r/Ethereum* with *max_lag* 48.

Table 2
Selected topics from *r/Bitcoin* subreddit.

r/Bitcoin		
#Topic	Label	Top Topic's Words
0	personal investment	'money', 'time', 'wallet', 'way', 'thing', 'work', 'coin', 'transact', 'crypto', 'point', 'actual', 'try', 'mean', 'mine', 'sure', 'person'
1	Bank	'price', 'exchange', 'thank', 'day', 'dollar', 'atm', 'wrong', 'shit', 'purchase', 'withdraw', 'satoshi', 'check', 'list', 'demand', 'name', 'hope', 'google', 'com', 'lt'
2	Bitcoin & Blockchain	'bitcoin', 'btc', 'year', 'new', 'look', 'post', 'currency', 'question', 'remove', 'address', 'node', 'network', 'change', 'free', 'month', 'internet', 'user', 'power', 'believe'
3	Government	'people', 'good', 'market', 'value', 'govern', 'right', 'world', 'account', 'reason', 'everyone', 'country', 'maybe', 'talk', 'idea', 'guy', 'last'
4	Trading	'gt', 'use', 'bank', 'pay', 'fee', 'cash', 'gold', 'trade', 'tax', 'scam', 'lol', 'need', 'word', 'coinbase', 'rate', 'kyc', 'ask', 'trust'

other events (of the same or of different type), in this work we tested different values of *max_lag*. We chose these *max_lag* values according to the period's length and the number of events associated per interval.

Before illustrating our results, it is important to clarify the interpretation of the Hawkes' coefficients matrices' values. The coefficients are obtained from the matrix representing the weights obtained from applying the Hawkes model fitted to the dataset and displayed from the vertical to the horizontal axis. They represent the average number of expected events per time interval.

The following graphs are read from the vertical to the horizontal axis, and the arrows help to understand which process is causing and which is caused. A blue colour indicates a causal relationship with a positive weight, namely when the number of events per time interval increases on the vertical axis variable, then the number of events increases in the horizontal axis variable. Red colour expresses a negative relationship, namely when the number of events per time interval increases on the vertical axis variable, then the number of events decreases in the horizontal axis variable. Intermediate weights and values very close to zero are represented with a colour that gradually approaches white (no causal relationship).

Fig. 12 shows the strength of the connections between the considered processes for Bitcoin and Ethereum technical discussions for a *max_lag* of 24. There is a general pattern of soft *self-excitement* positive relationships between all the variables, highlighted by the coefficients placed on the diagonal. Some causal relationship are established between *topic*_0 (related to discussion investments) of the Bitcoin and *topic*_3 and *topic*_4 of Ethereum (related to discussions about decentralized realities and deployment applications). The Bitcoin and Ethereum *log_return* feature is negatively affected by Bitcoin *topic*_0 feature, while positively affected by Ethereum *topic*_3 and *topic*_4.

The results obtained with a *max_lag* of 48, shown in Fig. 13, highlights that the *self-excitement* relationship are no longer present and the appearance of some causal relationship between *btc_log_return* feature and Bitcoin and Ethereum *delta_return* features. It also appears that

Fig. 14. Hawkes matrix for *r/BitcoinMarkets* and *r/EthTrader*, *max_lag* 24.

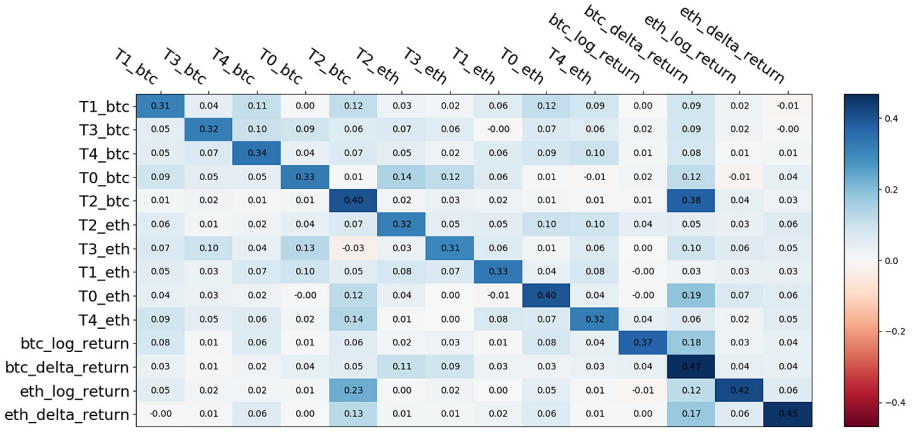


Fig. 15. Hawkes matrix for *r/BitcoinMarkets* and *r/EthTrader*, *max_lag* 48.

Bitcoin and Ethereum topics negatively affect the *delta_price* feature of both cryptocurrencies.

The Bitcoin and Ethereum *log_return* feature is negatively affected by Bitcoin *topic*_0 feature, while positively affected by Ethereum *topic*_3 and *topic*_4. The results obtained with a *max_lag* of 48, shown in Fig. 13, highlights that the *self-excitement* relationship are no longer present and the appearance of some causal relationship between *btc_log_return* feature and Bitcoin and Ethereum *delta_return* features. It also appears that Bitcoin and Ethereum topics negatively affect the *delta_price* feature of both cryptocurrencies.

Fig. 14 shows the connections strength between the considered processes for Bitcoin and Ethereum trading discussions for a *max_lag* of 24.

In this case the relationships between topics features and the variables related to the prices are almost zero. Instead, the *self-excitement* relationships between all the features are stronger than those occurred in the technical discussion case.

Fig. 15 shows the same Bitcoin and Ethereum trading discussions results obtained with a *max_lag* of 48 hours. It appears that there are no substantial differences compared to the previous case with *max_lag* equal to 24, the only remarkable observation that is worth noting is that all the negative relationships established between the topics features become positive in this case.

6. Cryptocurrencies price changes and social media: a case study on Reddit

This section focuses on the influences between online discussion groups of Bitcoin and Ethereum and their cryptocurrencies' markets price changes. We apply the same approach used in the previous section based on modelling online discussions with topic modelling and then use Hawkes models to highlight hidden relationships between social media activities and cryptocurrencies. In particular, we focus on shedding light on the relationships between social media discussions and Ethereum and Bitcoin price changes. Concerning the previous analysis in Section 5, we consider social features such as sentiment and emotions expressed in online social media comments.

Table 3
Considered subreddits.

Cryptocurrency	*Technical Discussions*	*Trading Discussions*
Bitcoin	r/Bitcoin	r/BitcoinMarkets
Ethereum	r/Ethereum	r/EthTrader

6.1. Data sources

We collected data from Reddit and selected two subreddits for each cryptocurrency considered, one technical and one trading related. Table 3 shows the considered subreddits.

The distinction between subreddits of technical discussions and subreddits with speculation (trading) topics is vital. In the first case, we expect to find users whose interest is more focused on technology (on the implementations of new products or technological systems); in the second case, in principle, we expect to find users who discuss the present, past and future trend of the cryptocurrency price (the financial aspects).

Table 4 shows the total number of messages extracted from the 1st of January till the 31st of December 2019, and the fraction of messages per user (or Redditor).

Table 4
Subreddits Statistics.

	Subreddit	# Comments	Comments/Redditor Ratio
Technical	***r/Bitcoin***	430.183	**8,21**
	r/Ethereum	80.130	**6,51**
Trading	***r/BitcoinMarkets***	212.828	**29,13**
	r/EthTrader	245.786	**16,42**

Bitcoin (the first cryptocurrency by market capitalization) is the most talked-about of the subreddits under review; by adding the two subreddits (technical and trading type), the total exceeds 630 thousand comments. However, for Ethereum, the volume of messages is higher for the trading theme than for technical discussions, and the reason might be that when it comes to Ethereum, users are more interested from the financial perspective than the technological one.

We created the dataset from data extracted from two primary sources: the market prices of Bitcoin and Ethereum and the comments extracted from the four main subreddits. Specifically, we extracted the data relating to Bitcoin and Ether prices from the "Historical Data" section available on the English exchange Coinbase.com. These data contain various information at one-hour intervals, in particular:

- the opening price.
- the closing price.
- the highest price.
- the lowest price
- the trading volume.

Prices have been modelled to express increases and decreases in volatility over time with the following two features.

- *log_returns_volatility*: it represents the difference between the highest price and the lowest price per hour, expressed in logarithmic form as represented in Eq. (2):

$$log\_returns\_volatility_i = \log(P_{High,i} - P_{Low,i}) \tag{2}$$

- *delta_return*: it represents the difference between the closing price and the opening price in one hour, as expressed in Eq. (3):

$$delta\_return_i = (P_{Close,i} - P_{Open,i}) \tag{3}$$

Along with these technical features, we added sentiment and emotional features, in particular for each comment we evaluated the following features:

- Sentiment.
- Valence Arousal and Dominance (VAD) metrics.
- Dynamic Topic Modelling (Blei & Lafferty, 2006).

The sentiment of Reddit's comments has been evaluated using a Sentiment Analysis tool called VADER (Hutto & Gilbert, 2014) as we described in Section 4.1. VAD metrics have been computed using the well known Warriner framework Warriner et al. (2013). The topic features refer to the five topics identified in the previous section (and reported in Table 2).

Starting on the 2nd of April 2019, the Bitcoin price began a phase of progressive rise, which ended on the 28th of June 2019 (followed by strong fluctuations in prices which lasted throughout the summer). In the last quarter of the year, there were two significant drops in price. The first began on the 23rd September 2019, and the second (with a gradual decrease which began on 27th October 2019 and ended on 23rd November 2019) with a loss in value of over $ 2500 (25%).

The fluctuations in the price of Ether roughly follow those of Bitcoin. Both cryptocurrencies recorded upward peaks starting from April 2nd 2019. Unlike Bitcoin, in Ethereum, the price hike's start begins in the first days of February (when the news on the protocol change started a trail of cautious optimism). There are more negative fluctuations, especially during the summer, when Ether's price collapsed on July 9th 2019. As of July 17th 2019, in just eight days, the price has dropped by over $ 100 (about 30% of the value). We used these drops and rises in prices as representative case studies in Section 6.3.

6.2. Methodology

The use of Hawkes processes models a time series of occurrences of events. Therefore, it is essential to express the events as an aggregation of information: we merged the data with a frequency of 60 min (Δ_t), as described in previous Section 5. This periodicity made it possible to count the number of related events per hour, from the 1st of January to the 31st of December 2019.

Specifically, for the sentiment, we decided to count the number of positive and negative events. Therefore, in an interval of one hour, n comments with positive polarity and m comments with negative polarity are counted. We excluded neutral events because they were considered uninformative; in fact, most of these represent questions (e.g., "Is Bitcoin a good choice?"), or comments without any impact on the market trend (e.g., "I don't know anything about trading!"). VAD metrics are composed of continuous values. We decided to divide the VAD metrics into two categories to count events:

- *Low*: if the associated VAD value is lower than the median value of the entire year 2019;
- *High*: if the associated VAD value is instead higher than the median threshold for the entire year 2019.

For example, when a comment has an Arousal value greater than the median, the *low_arousal* value is zero, while the *high_arousal* value is one. The same holds for the other two VAD metrics: Dominance and Valence. These occurrences were then aggregated to form a single event with a duration of Δ_t, which, as mentioned, is one hour. For each subreddit, the LDA model generated five topics, to which we manually assigned a label. In all four subreddits, the recurring themes are similar: they mainly talk about the world of cryptocurrencies, trading and currency appreciation and depreciation, but they also refer to Blockchain technology, decentralized realities and their implementation at the government level and politic. Each comment has been assigned to one of the five identified topics (dominant topic). In this case, if the number of comments associated to a topic, for example with *topic_0*, is equal to 10, this will be the count of the events related to *topic_0* in that time interval.

During the whole year 2019, comments were grouped to create 8760 events (24 events a day for 365 days). Table 5 shows the distribution of events during the entire period.

The least frequent subreddit is r/Ethereum, with an average of 9 comments per events per time interval. The most densely packed subreddit is r/Bitcoin, with over 48 comment events per time interval.

6.3. Results

In Fig. 16, the first subreddit analysed is r/Bitcoin. In this paragraph and for all four subreddits, it has been chosen to implement a Hawkes weight matrix using a *max_lag* equal to 24. This value's choice is linked to the length of the period considered: shorter relationships take on less importance than the effects propagated over time.

The matrix in Fig. 16 shows that the relationships between the variables express self-excitement effects; therefore, the generation of an event in a process generates an effect in the same process: this is clear by looking at the blue diagonal in the graph.

Fig. 17, on the other hand, shows the same type of chart for r/Ethereum, in which prices are related to Ether's performance. Remarkably, in this case, there are no marked self-excitement relationships as in the case seen in Bitcoin (Fig. 16), but there are strict mutual-excitement relationships. In particular, on the graph's right,

Table 5
Reddit Dataset Statistics.

Subreddit	Avg. # Comments per event	# Comments	Comments/Redditor Ratio
r/Bitcoin	48,27	430.183	8,21
r/Ethereum	9,02	80.130	6,51
r/BitcoinMarkets	23,97	212.828	29,13
r/EthTrader	27,69	245.786	16,42

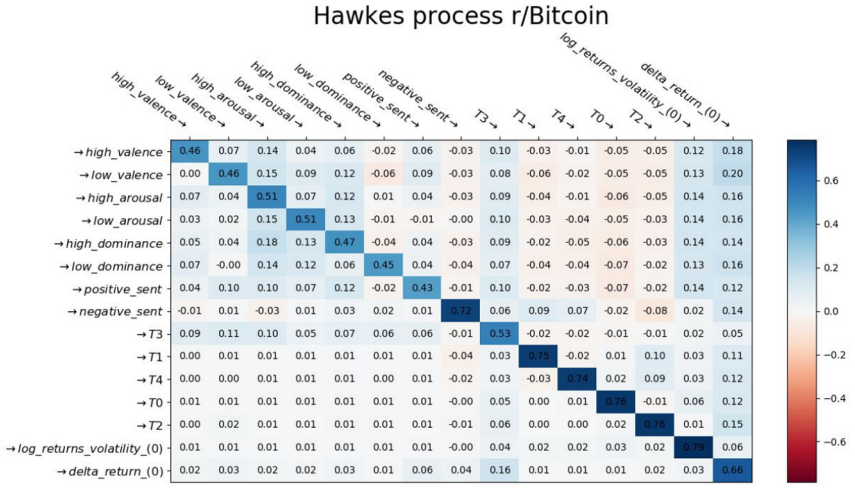


Fig. 16. Hawkes matrix for *r/Bitcoin* *max_lag* 24 during all 2019.

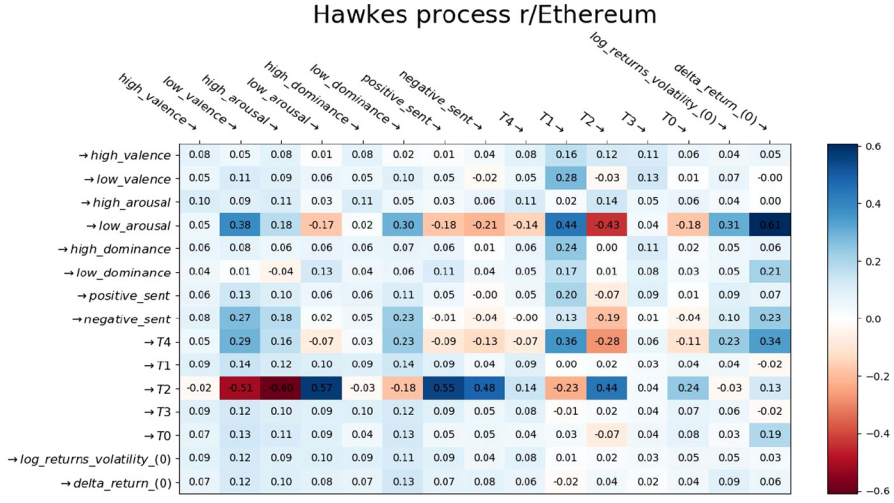


Fig. 17. Hawkes matrix for *r/Ethereum* *max_lag* 24 during all 2019.

we see that events in the *delta_return* cause low-level positive effects of *Arousal*. Hence, a price increase makes users unresponsive. This is consistent with the fact that Redditors are, in general, attracted to price devaluations so that they can take advantage of sudden depreciation and profits from the future rise in the currency.

Figs. 18 and 19 instead show the matrices whose discussions are strictly characterized by financial and trading themes, namely: r/BitcoinMarkets and r/EthTrader. As widely discussed throughout this study, the users who belong to these two subreddits are very attentive to price trends, and, for this reason, they are also very sensitive. The matrix confirms this in Fig. 18, where an increase in price volatility causes a reduction of negative comments mainly linked to the negative sentiment of the Redditors. This can be explained by the fact that investors prefer situations of stress and high volatility to speculate more on the price.

In Fig. 19, on the other hand, the effects are mainly related to the generation of high and low Arousal events. Thus, during the entire period, r/EthTrader users are not influenced by prices, nor do they influence them. On the other hand, high Arousal situations generate positive effects, therefore an increase, of the events linked to negative sentiments.

The effects just illustrated could change depending on the type of information driving the market. For example, in February, the news of the Ethereum protocol change was the main topic discussed. Conversely, users expressed different emotions in June due to the large speculative bubble that swept through prices. For this reason, periods of particular financial stress had to be investigated more closely. In the following subsections, we selected four of these particular periods of interest:

- Case 1: The Days Close To The Price Increase As Of April 2nd 2019.
- Case 2: the update of the Ethereum protocol in February.
- Case 3: the fall of the Ether on July 14th.
- Case 4: the fall in the price of Bitcoin on September 24th.

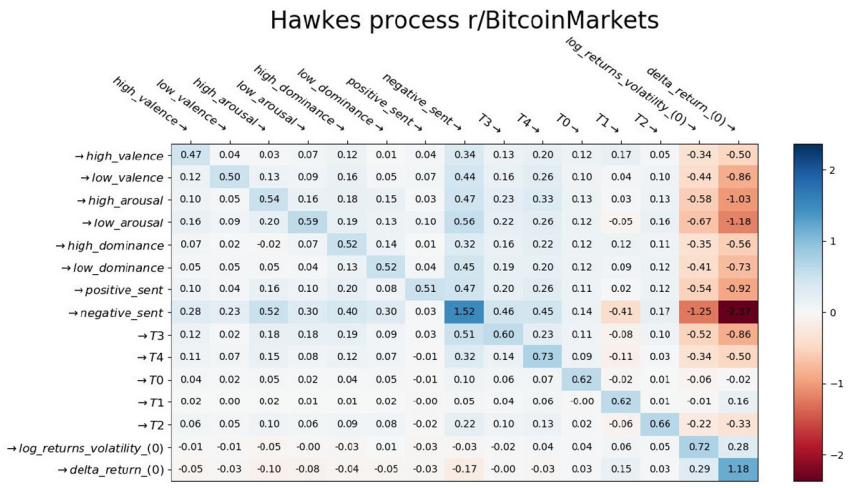


Fig. 18. Hawkes matrix for *r/BitcoinMarkets* *max_lag* 24 during all 2019.

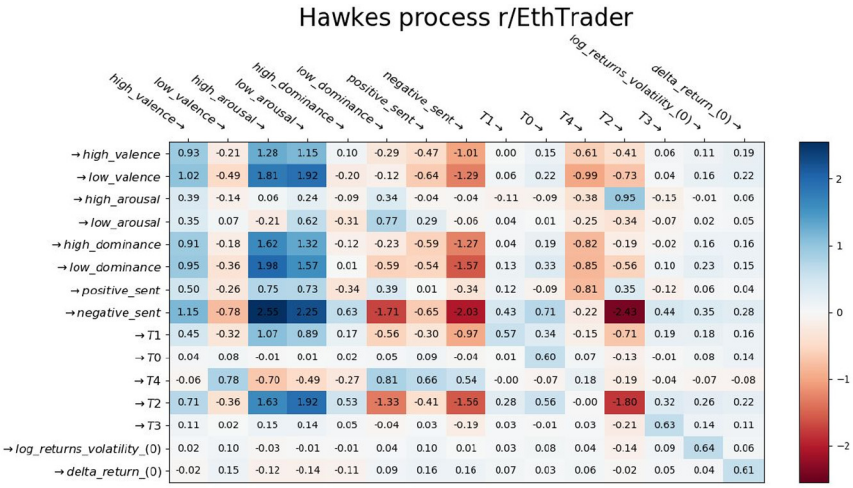


Fig. 19. Hawkes matrix for *r/EthereumTraders* *max_lag* 24 during all 2019.

6.3.1. *Case 1: The Beginning Of The Spring 2019 Rises*

Compared with the whole of 2019, the first three months of the year are characterized by general price stability in both cryptocurrencies. Figs. 20 and 21 show the coefficient matrices for trading discussions for Ethereum and Bitcoin. The analysis period chosen in this first subparagraph corresponds to the days preceding the gradual rise in prices. Between April 2 and 4, 2019, the price of Bitcoin suddenly increased by nearly $ 1000, and on the same incremental wave, Ether's value went from $ 140 to over $ 160 (+ 12.5%) in the same days. Studying relationships in trading-type subreddits has led to observable results, while this has not been the case for technical discussions. Fig. 21 shows the weight matrix relating to the r/EthTrader subreddit. In this case, the events linked to high Valence and Arousal values generate adverse effects in price reduction, while the events of high Dominance, on the contrary, generate positive effects in price fluctuations.

Fig. 22 shows the r/BitcoinMarkets subreddit. As in Fig. 21, here again, high levels of *Arousal* generate significant decreases in the price. It can be noted how high values of *high_arousal* generate positive effects on the topic T4 relating to time, taxes and the future. This topic has generated concern among users, also confirmed by the negative effect generated by the *low_arousal* events.

6.3.2. *Case 2: The update of the Ethereum Protocol in February*

The price of Ether has gone through a period of recovery since early February 2019. In this regard, the analysis was conducted on the Ethereum platform subreddits by selecting the period between January 31st and February 6th 2019 (two days before the rise). Fig. 23 shows the matrix of the coefficients obtained with Hawkes models for discussions in r/Ethereum. In this case, the different topics are mostly affecting prices. Interestingly, the most important topic is related to discussions on the fork (T_1), which causes positive effects on the price. There is also a trail of positive comments, which in turn generates price increases.

Curiously, there were no price influences from the metrics extracted from r/EthTrader. This result might suggest that the rise is due precisely to the news of the imminent change of protocol of Ethereum, and that it is not a trivial coincidence.

6.3.3. *Case 3: The Fall of the Ether on July 14th*

In the time interval between 11th and 13th July 2019, Ether's price is quite stable, all this happens before the price, on July 14th, undergoes through a deep drop of over $ 40 (−11.3%). The weight matrix in Fig. 24 (r/Ethereum) is characterized by a strong influence on prices due to low *Dominance* events (positive effects on *delta_return*, while negative effects on *log_returns_volatility*). Although the price is generally stable, users do not feel in control of the situation, which generates continuous and oscillatory movements in the price.

The r/EthTrader coefficient matrix is illustrated in Fig. 25. There are a few significant relationships here. Indeed, the prices are positively caused by the topic T_0 related to generic discussions on the Blockchain.

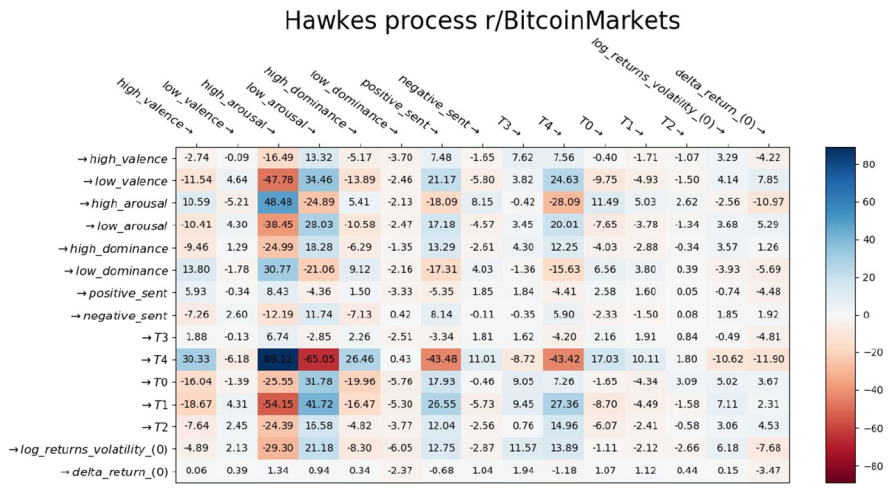


Fig. 20. Hawkes matrix for *r/BitcoinMarkets* *max_lag* 12 during 1st April 2019.

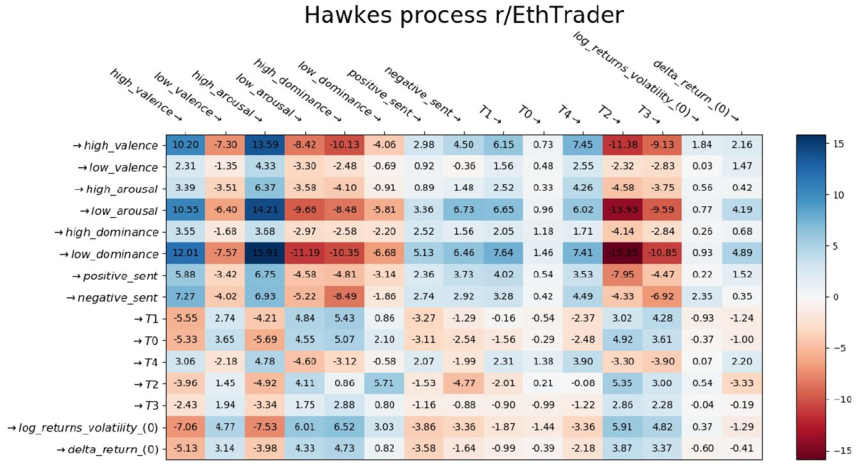


Fig. 21. Hawkes matrix for *r/EthereumTraders* *max_lag* 12 during 1st April 2019.

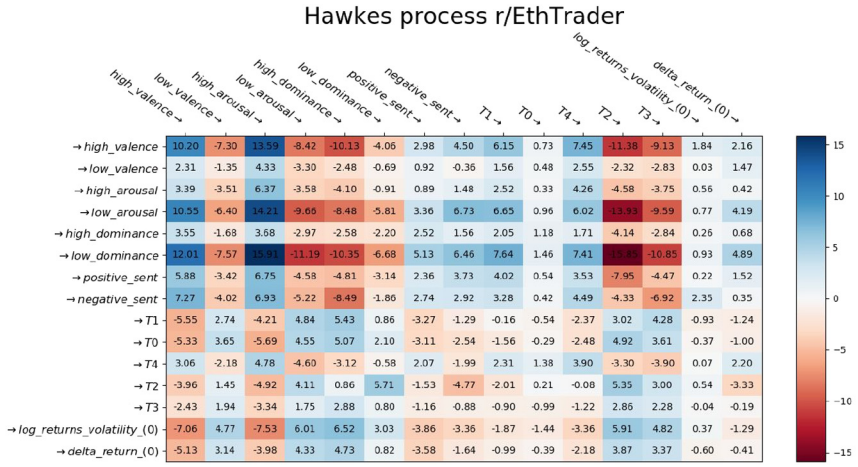


Fig. 22. Hawkes matrix for *r/BitcoinMarkets* *max_lag* 6 from 28th March to 1st April 2019.

The words present in this topic (“f ** k”, “remove”, “downvote”, “question”) suggest little discussion’s serene. The influence of topic T_3 on prices is, on the contrary, negative. Discussions related to the government, money, investment portfolios cause prices to rise significantly stable and gradual.

6.3.4. Case 4: The Price Fall of Bitcoin on September 24th

For the last case study, we consider the period preceding the great collapse of the Bitcoin price, more precisely, 24th September 2019 around 8 pm. The time interval considered is between 20th and 24th September, about 1 h before the sudden collapse of the value of over $

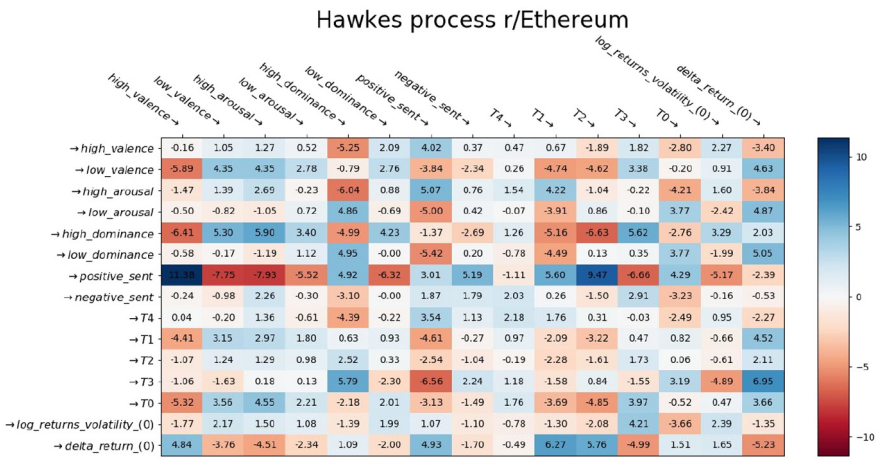


Fig. 23. Hawkes matrix for *r/Ethereum* *max_lag* 12 from 31st January to 6th February 2019.

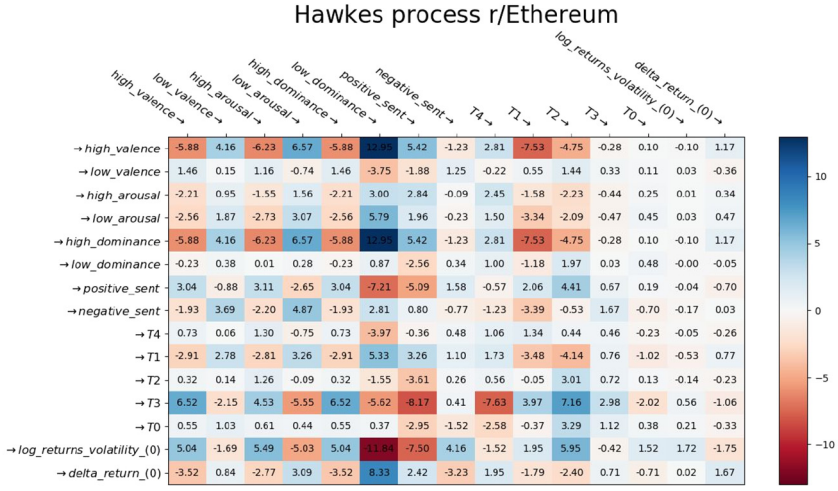


Fig. 24. Hawkes matrix for *r/Ethereum* *max_lag* 12 from 11th July to 13 July 2019.

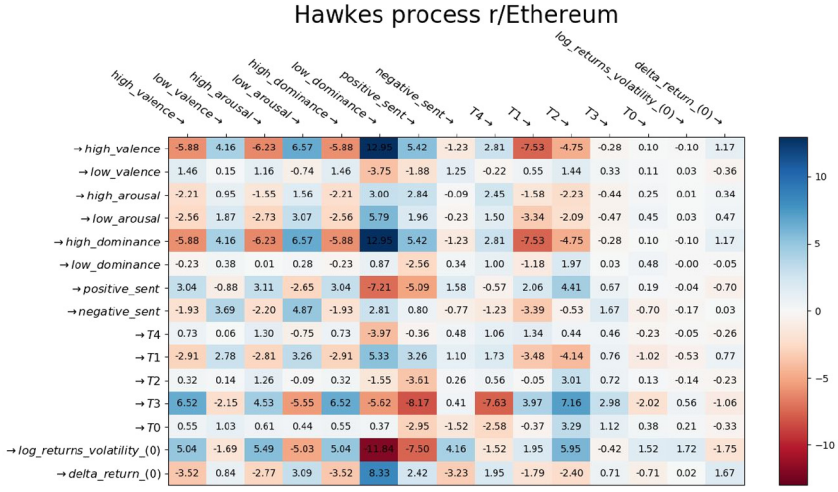


Fig. 25. Hawkes matrix for *r/EthereumTraders* *max_lag* 12 from 11th July to 13 July 2019.

2000 (over 10% of the total value). Figs. 26 and 27 show the two coefficient matrices corresponding to the r/Bitcoin and r/BitcoinMarkets subreddits. With regard to the first figure, it is mainly the emotions linked to the VAD metrics that drives the greatest influences on prices. In particular, events related to *low_dominance* (which, among other things, has a negative self-excitement value) cause positive effects in

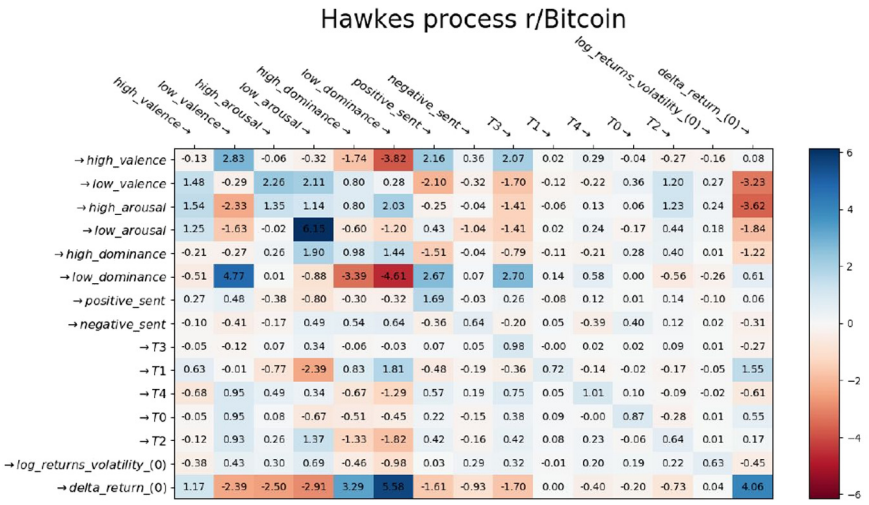


Fig. 26. Hawkes matrix for *r/Bitcoin max_lag* 12 from 20th to 23rd September 2019.

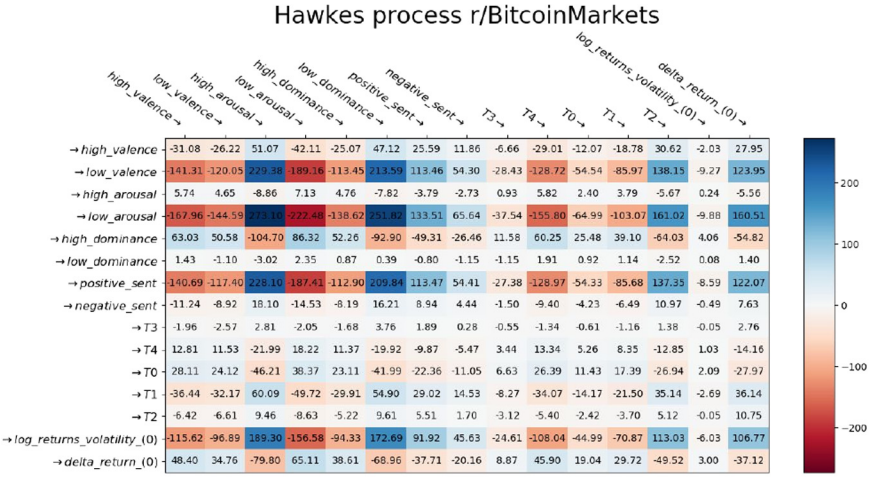


Fig. 27. Hawkes matrix for *r/BitcoinMarkets max_lag* 12 from 20th to 23th September 2019.

delta_return. The *low_dominance* events are gradually decreasing over time, and this generates a slight rise in prices, before the final fall of the 8pm on September 24th.

Other remarkable results are shown by analysing the matrix relating to discussions of the trading theme (Fig. 27). In particular, low *Dominance* and high *Arousal* events generate positive effects on price volatility (*log_volatility_return*). Conversely, high *Dominance* and low *Arousal* events have negative effects on *log_volatility_return*.

The case in Fig. 27 shows how low *Dominance* events cause an increase in events linked to equally low levels of *Arousal*, and which in turn have repercussions to rises in price volatility and a continuous reduction of prices.

7. Validation

In this section, we propose a validation methodology based on empirical evidence. We will use the striking case of "WallstreeBets and Game Stop"[3] which took place in January 2021 and is the first recorded case where millions of social media users coordinate their stock market's behaviour with the purpose of short squeezing a stock market title.

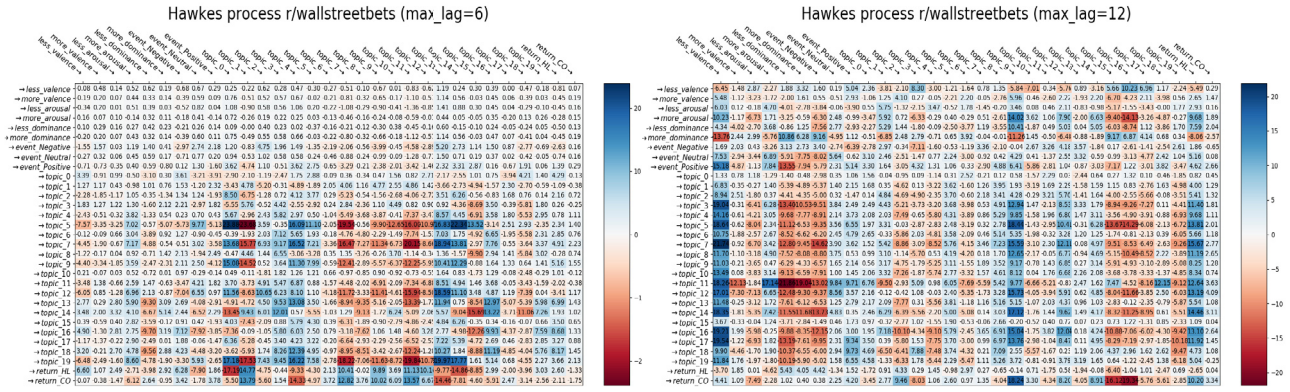
In brief, users on the forum WallStreetBets, a subreddit of Reddit.com,[4] had noticed what was happening with GameStop,[5] a publicly-traded company, which was having economic difficulties. The forum users realized that the hedge funders were betting against GameStop and shorting more shares than the existing ones (multiple landing of a share). Investors on WallStreetBets coordinate a buy action of GameStop shares, pushing the price up and disrupting the hedge fund investors' plan. The move worked, creating higher demand for the shares in relation to supply, and forcing the hedge funds to repurchase shares at a higher price, generating a loss for millions. As a result of these actions, the GME title registered +400% gain (on the 12th of January 2021, the value of a GME share was $19.95, while on the 29th of January it reached $325). The value of the title skyrocketed after the action of a coordinated (through the forum WallStreetBets) group of online traders who targeted several hedge funds which decided to short-sell shares of the video game company.

We used this use case as an empirical validation benchmark for our model based on Hawkes model. From one side, we have the evidence of social media users who coordinate themselves using the Reddit platform, and on the other side, we can measure the price movements of a stock market title. We ask whether it is possible to highlight social media signals of what was going to happen using our model.

[3] https://www.economist.com/finance-and-economics/2021/02/06/how-wallstreetbets-works.

[4] https://www.reddit.com/r/wallstreetbets/

[5] https://www.gamestop.com/.



(a) Hawkes coefficients' matrix before price arise from 15th to 25th January 2021

(b) Hawkes coefficients' matrix before price arise from 18th to 25th January 2021

Fig. 28. Hawkes coefficients' matrix before price arise from 15th to 25th January 2021.

Although the crypto-markets and stock markets are very different in many ways, they also exhibit many similarities Al-Yahyaee et al. (2018), Liang et al. (2019), Soloviev et al. (2020), and this use case provide a perfect benchmark since we already know the causal effect.

Figs. 28(a) and 28(b) shows the Hawkes model coefficients' matrix (max_lag equal to 6) in two different time windows: i) from 15th to 25th of January 2021; ii) from 18th to 25th of January 2021. We aim to record events in social media activities that are precursor of rises and fall in the price of the GME share.

The time difference between the first (15–25 January) and the second period (15–18 January) is one day. In fact, during the weekend, the stock market is closed. We note that as the period of analysis approaches 25th January, the coefficients' scale decreases, passing from a range of [15; -15] to [3; -3]. The colour of the coefficients also tends to brighten more as the days approach 25th January, the day before the prices increase beginning. The price time series cells light up more. This is mainly due to greater interest from people, which probably generated a substantial increase in the number of comments in the subreddit. There were few comments at the beginning of the month (compared to the end of the month, when the stock value skyrocketed). During the period from 23rd to 27th January 2021, there was an exponential increase in prices. For this reason, this is the most exciting period in which users turn on their discussions the most. Few relationships are seen in general, and there are mostly mutual-excitement relationships. During this period, the rwallstreetbets subreddit closed for two days (January 25th and 26th), so they are not considered in the analysis. Specifically, in the matrix in Fig. 29(a), the matrix coefficients of the days 23 and 27 January are reported, two distant days in which a lot has happened. Events of positive and negative sentiment generate mutual excitement. In particular, positive sentiment events cause a positive increase in price variability. Sentiment still has a significant effect on topics.

The period January 25 and February 5 is the most chaotic of the entire analysis. The price of the GME share is going through a phase case of enormous expansion (over 300%) and then to a phase of sharp decline (with a fluctuating trend between the 25th and 2nd of February). The latter is a black day for GME investors, with the price dropping to 90 dollars. In this period, the coefficient matrix shown in Fig. 29(b) highlights different causal relationships with the price variables. Among all, the high valence events (more_valence) negatively affect the rising and falling price events (CO_price_events_up and CO_price_events_down). On the other hand, neutral sentiment events have a negative effect on the latter, probably due to a period of uncertainty (users ask questions about what is happening to the stock price, generating additional uncertainty and price falls).

8. Conclusions

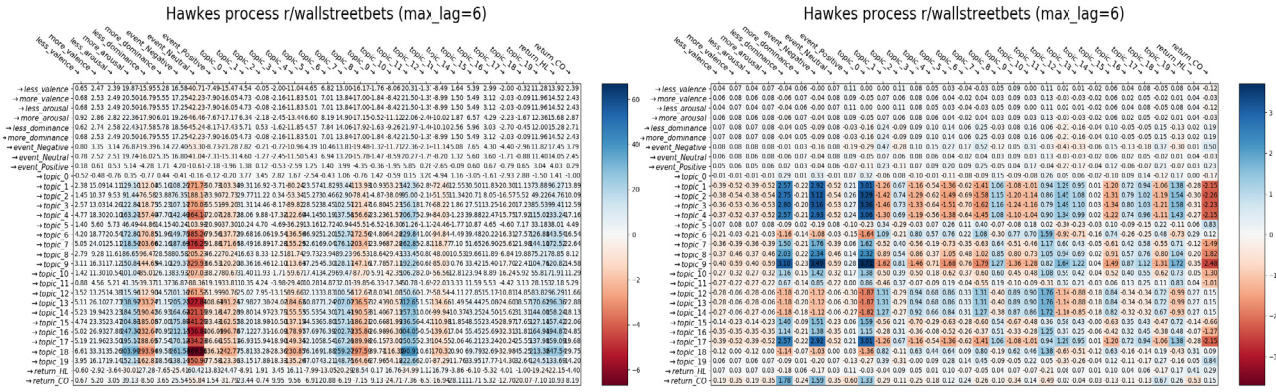
This work aimed to identify the relationships between topic discussion occurrences on social media and cryptocurrencies market price changes. Dynamic topic modelling was first applied to social media content, and then a Hawkes model was used to decipher relationships between topics and cryptocurrency price movements.

We started our study considering the evolution of sentiment and discussed topics on two developers communities: Bitcoin and Libra. We analysed discussions of developers communities in two popular social platforms: Reddit and Telegram. Technical discussions range from software development to fintech, and our goal was to understand if and how the discussions and sentiments in one Blockchain community influence one another. We considered a time span of one month, June 2019. We chose this particular month due to the announcement of Libra, which has stirred the cryptocurrency market.

Our analysis showed that the price of Bitcoin is highly positively correlated with the sentiment arising from all the considered discussion's groups except for *Ripple Group (XRP)*, which has a robust negative correlation equal to −0.71. This result is not fortuitous and explains that this is the only group in our analysis that does not appreciate Bitcoin's price growth. Although limited to one month, this study shows how Blockchain development communities influence each other from the sentiment expressed in technical discussions. Furthermore, further analysis that involves more sophisticated methodologies such as Hawkes processes (Phillips & Gorse, 2018) have been applied to deeper understand how discussions in different Blockchains communities influence each other and influence the cryptocurrencies' price (Bartolucci et al., 2020).

The additional analyses continued in this direction and revealed a link between the cryptocurrency market and the content in social media dedicated to these topics. In particular, it is demonstrated that users' opinions can explain capital movements and consequent price fluctuations.

The work was divided into four phases: data extraction, data preprocessing, the construction of variables and, finally, the modelling of metrics for applying Hawkes' processes. In the last section of the work, we presented several case studies in which we focused on particular periods close to financial stresses for both cryptocurrencies Bitcoin and Ethereum. Although the intrinsic causes of price fluctuations are still being researched today, experiments have shown interesting and significant causal relationships between metrics extracted from social media and price fluctuations. The extraction and processing of data from different sources (subreddit of type technical and trading type) made it possible to deduce various recurring themes. From the construction of the matrices of the coefficients obtained by applying the Hawkes



(a) Hawkes coefficients' matrix from 23rd to 27th January 2021

(b) Hawkes coefficients' matrix before price arise from 25th January to 5th February 2021 during price fall.

Fig. 29. Hawkes coefficients' matrix before price arise from 23rd January to 5th February 2021 during price fall.

processes, it has been shown how users in the multiple communities react differently to price changes.

In general, considering the whole year 2019, the r/Bitcoin subreddit presented self-excitement reports, while r/Ethereum mutual-excitement relations were mainly caused by price fluctuations. However, in the trading-focused subreddits, it is shown how in r/BitcoinMarkets price changes have adverse effects on metrics extracted from the Redditors' comments. Furthermore, analysing four specific cases between Bitcoin and Ethereum, we noted that the topics have less influence than the VAD and Sentiment metrics. The case of Ethereum was an exception when users discussed the upcoming protocol change in February. It is mainly the VAD metrics that cause prices changes and High and Low Arousal events. Thanks to the metrics and analysis tools shown in this paper, it is possible to have a new overall point of view capable of capturing new patterns. We validated the model using a real study case: the "WallstreetBets VS GameStop" event that took place in January 2021. We found that it was possible to detect warning signals of imminent financial distress with our model.

Through these applications, it is possible to programme a real-time alarm system capable of highlighting **precursors** warning events of financial stress as a sort of "**social media seismograph**" which anticipate financial earthquakes, similarly to how seismographs identify earth tremors that anticipate an earthquake. In addition to being effective in the financial field, the Hawkes processes shown here can be applied in various fields, including forecasts in electoral polls, measurement of a company's brand reputation, and identification of particular social events in real-time. Consequently, the Hawkes coefficients in such a system can be tracked, and as a coefficient begin to change colour, in our example from deep blue (high positive influence) to deep red (high negative influence), cryptocurrencies stakeholder may take informed actions on their stocks.

From the research point of view, future works will focus on the methodological aspects and implications of Hawkes' model, especially in constructing an extended Hawkes' model that can more accurately fit the social media data explored in the present study. These future studies will help better understand the role of social media in the online cryptocurrencies environments.

CRediT authorship contribution statement

Marco Ortu: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft. **Stefano Vacca:** Investigation, Data curation, Software. **Giuseppe Destefanis:** Writing – review & editing. **Claudio Conversano:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

No funding was received for this work.

Intellectual property

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

References

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.

Bacry, E., Bompaire, M., Gaiffas, S., & Poulsen, S. (2017). Tick: a python library for statistical learning, with a particular emphasis on time-dependent modeling. arXiv e-prints.

Bartolucci, S., Destefanis, G., Ortu, M., Uras, N., Marchesi, M., & Tonelli, R. (2020). The butterfly "affect": Impact of development practices on cryptocurrency prices. *EPJ Data Science*, *9*(1), 21.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120).

D. Blei, A. N., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 993–1022.

Destefanis, G., Ortu, M., Counsell, S., Swift, S., Tonelli, R., & Marchesi, M. (2017). On the randomness and seasonality of affective metrics for software development. In *Proceedings of the symposium on applied computing* (pp. 1266–1271). ACM.

Garcia, D., & Schweitzer, F. (2015). Social signals and algorithmic trading of bitcoin. *Royal Society Open Science*.

Hawkes, A. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*. Vol. 8. No. 1.

Keskin, Z., & Aste, T. (2019). Information-theoretic measures for non-linear causality detection: application to social media sentiment and cryptocurrency prices. arXiv: 1906.05740.

Kristoufek, L. (2013). Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific Reports*, 197–215.

Kyriazis, N. A. (2019). A survey on efficiency and profitable trading opportunities in cryptocurrency markets. *Journal of Risk and Financial Management*, *12*(2), 67.

L. Cocco, M. M. (2019a). An agent-based artificial market model for studying the bitcoin trading. In *IEEE access IEEE Access*, 42908–42920.

L. Cocco, M. M. (2019b). An agent based model to analyze the bitcoin mining activity and a comparison with the gold mining industry. *Future Internet*, 8.

Liang, J., Li, L., Chen, W., & Zeng, D. (2019). Towards an understanding of cryptocurrency: a comparative analysis of cryptocurrency, foreign exchange, and stock. In *2019 IEEE international conference on intelligence and security informatics* (pp. 137–139). IEEE.

Murgia, A., Ortu, M., Tourani, P., Adams, B., & Demeyer, S. (2017). An exploratory qualitative and quantitative analysis of emotions in issue report comments of open source systems. *Empirical Software Engineering*, 1–44.

Neuts, M. F. (1979). A versatile Markovian point process. *Journal of Applied Probability*, 764–779.

Ortu, M., Destefanis, G., Kassab, M., Counsell, S., Marchesi, M., & Tonelli, R. (2015). Would you mind fixing this issue? An empirical analysis of politeness and attractiveness in software developed using agile boards. In *Lecture notes in business information processing212* (p. 129),

Phillips, R. C., & Gorse, D. (2018). Mutual-excitation of cryptocurrency market returns and social media topics. In *Proceedings of the 4th international conference on frontiers of educational technologies* (pp. 80–86). ACM.

R. C. Phillips, D. G. (2017). Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *IEEE symposium series on computational intelligence*.

R. Rehurek, S. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*.

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

S. McNally, S. C. (2018). Predicting the price of bitcoin using machine learning. In *26th Euromicro international conference on parallel,and network-based processing* (pp. 339–343).

Soloviev, V., Yevtushenko, S., & Batareyev, V. (2020). Comparative analysis of the cryptocurrency and the stock markets using the Random Matrix Theory. In *CEUR Workshop Proceedings*.

Stuart G. Colianni, M. S. (2015). Algorithmic trading of cryptocurrency based on Twitter sentiment analysis. *Computer Science*.

Uras, N., Vacca, S., & Destefanis, G. (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* (pp. 779–782).

V. Y. Naimy, M. R. H. (2018). Modelling and predicting the bitcoin volatility using GARCH models. *International Journal of Mathe-Matical Modelling and Numerical Optimisation*, 197–215.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Wołk, K. (2020). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems*, *37*(2), Article e12493.

Zheng, Z., Xie, S., Dai, H.-N., Chen, X., & Wang, H. (2018). Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services*, *14*(4), 352–375.