Community impact on a cryptocurrency: Twitter comparison example between Dogecoin and Litecoin

Edouard Lansiaux (✉ edouard.lansiaux@globalvariations.com)
Global Variations  https://orcid.org/0000-0002-4342-8707

Noé Tchagaspanian
Global Variations

Joachim Forget
National Assembly

Research Article

Keywords: Shannon information entropy, conditional probabilities, Pearson correlation coefficient, Dogecoin, Litecoin, community impact, Twitter

Posted Date: January 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1265494/v1

License: ☇ ① This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Community impact on a cryptocurrency: Twitter comparison example between Dogecoin and Litecoin

Competing interests: All authors declare: no support from any organisation for the submitted work other than that described above; no financial relationships with any organisation that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Word count: 2,400

Number of figures: 13

Number of tables: 5

Supplemental material: 4

Supplementary Figures: 0

Supplementary Tables: 2

Supplementary Files: 2

Keywords: Shannon information entropy; conditional probabilities; Pearson correlation coefficient; Dogecoin; Litecoin; community impact; Twitter
Abstract

Context: The 3rd generation of cryptocurrencies groups together cryptocurrencies as diverse as they are sulphurous, like Dogecoin or Litecoin. While one qualifies as memecoin, the other is of interest to a different category of investors. In our knowledge, no study has independently assessed crypto community economic impact concerning this comparable cryptocurrencies.

Method: Our study has retrospectively studied (from 01/01/2015 to 03/11/2021), using open access data, the association strength (using normalized mutual information) as well as the linear correlation (using Pearson’s correlation) between Twitter social networks markers and cryptocurrency economic markers. In addition, we have computed different models in order to forecast past prices values and then we have assessed their precision.

Findings and conclusions: While average Dogecoin transaction value is impacted by tweets, Litecoin transactions number and average Litecoin transaction value impacted tweets. Concerning whales, tweets are impacted by Dogecoin whales but any significant relationship was found between Litecoin whales and tweets. According to our ARIMA(0,0,0) models, we have a past values forecasting error of 0.08% on Litecoin and 0.22% on Dogecoin. Therefore, there are thus the beginnings of a scientific rationale in order to build a trading robot based on these big datas. This paper is only for academic discussion, conclusions need to be confirmed by further research.
Introduction

Since Satoshi Nakamoto’s whitepaper in 2008, cryptocurrencies have grown to a huge market capitalisation — currently over $2T. This huge rise in cryptocurrency market capitalisation seems at first glance, deeply linked to the cryptocurrency’s community. Indeed, most coins have a strong community promoting it through social networks. One of the most relevant examples when talking about online promotion of a coin might be Elon Musk’s tweets. He seemed to have a huge impact on the cryptocurrency market as value seems to increase or decrease as he tweets, which could constitute an insider delay. However, according to a big data study [1], it can clearly be stated that Elon Musk can not control the utter volatile world of cryptocurrencies and especially Bitcoin and Dogecoin.

But the price of cryptocurrency could in the end be more related to the Lindy effect than anything else. According to this effect, the future life expectancy of certain non-perishable goods — such as a technology or an idea — is proportional to their current age. Indeed, the longer something has been around, the more chances there are that it will survive longer. Among memes, competition for survival is fierce. In this jungle the average lifespan is roughly 4 months. When compared to other memes, Doge is kind of a venerated elder. By surviving for eight years — The Doge meme first became popular in 2013 — it has already proven to be one of the most resilient memes of the whole internet history. The Lindy Effect suggests that, for this reason alone, Dogecoin is more likely to persist into the future than any other meme. Just as the US Dollar is backed by America’s hegemonic power, Dogecoin is backed by some of the most powerful memes in existence — and the communities behind them. In the meme economy, Doge is as close to a stable thing as you can get. Indeed Dogecoin has a real fan base promoting its use through social networks. Part of what has made Dogecoin a successful cryptocurrency is the non-tribalism of its community.

On the other hand, while being technically very similar (i.e. almost same PoW, and use cases) : Litecoin has a less loud community, despite being older and a more stable cryptocurrency Litecoin doesn’t have the same online popularity. Litecoin’s users aren’t that loud over social networks and do
not mean — for most of them — to organise coordinated buying in order to influence the currency’s value.

Litecoin and Dogecoin have been selected in particular for this study due to their parentage. Indeed, on one hand, Litecoin was an early Bitcoin spinoff or altcoin, starting in October 2011 [2]; and on the other one, Dogecoin's protocol based on existing cryptocurrencies Luckycoin and Litecoin[3], allowed a launch of this cryptocurrency on December 6th, 2013 [4]. On the first view, these cryptocurrencies are different and might not be comparable. As explained previously, the Lindy effect allowed Dogecoin, this "memecoin" to free itself from its initial status as a joke. Thus, after a relative period of abandonment [5], it was pushed thanks to its community both at the level of its adoption and at the level of its technical development [6] (reduction of transaction costs, bridge with ETH, first NFT forged on its network, ...). Consequently, we observe a real decentralized approach of this cryptocurrency since the whole community makes it evolve. Conversely, Litecoin has a more centralized development via its foundation, which does not prevent it from having one of the largest capitalizations. Therefore, Litecoin was one of the first altcoins and Dogecoin was one of the first memecoins; their evolutions as single bitcoin daughters were both dependent on their community. Other memecoins cannot be used for comparison due to their capitalist minds so far removed from those libertarians of cypherpunks [7].

As it was broadly studied and well-documented, crypto-economy and wider the financial economy has a big behavioral part with about 20 cognitive biases [8]. Here we will be addressing that question: how much of an impact does online activity — through Twitter in this paper — have on the cryptocurrency market. We plan to investigate two main cryptocurrencies (Litecoin and Dogecoin) for this study:

- Are cryptocurrency prices and tradeoffs affected by the fluctuation in cryptocurrencies communities tweets?
- Can the price of Litecoin and Dogecoin be forecasted, according to this assumption?
- Can Dogecoin or Memecoin actually become a currency of the future?
Methods

Our analysis was divided in two parts: the first one the correlation/causality analysis and then the possible price predictions of those cryptocurrencies based upon causality/correlation analysis.

Causality and correlation analysis

We used two different methods in order to evince association, or lack thereof, between X and Y. These two methods were, on one hand, the classical Pearson correlation and, on the other hand, normalized Shannon mutual information.

Settings

We used historical data, spanning from 01/01/2015 to 03/11/2021, by extracting various economic trackers detailed below.

Variables

For each method, we have studied the following variables:: ‘date (quantitative variable)’, ‘top_100_percent’ 100 first addresses with a large wallet on the studied crypto blockchain (i.e; “whales) (quantitative variable), ‘median_transaction_value’ (quantitative variable), ‘market_cap’ (quantitative variable), ‘average_transaction_value’ (quantitative variable), ‘active_addresses’ on Twitter (i.e. most important influencers) (quantitative variable), ‘tweets’ (quantitative variable).

Data sources

The data frame mainly comes from 3 websites [9-11], but the reason why there are always two data frame versions is that data was lacking for specific days. The first file is the original one which contains some “null” values. But in order to work with our algorithm they have been filled (in the second file) with the average value of the last existing value and next one. This allows us to work with our files without introducing new bias in our correlations.
Statistical method

Obviously, correlation is not causation; but absence of correlation implies absence of causality. Correlation (which might be negative as well as positive) is therefore a key component of the scientific process, for it evinces collections of variables that may interact with each other, hereby warranting further study. Conversely, this methodology also accounts for the early dismissal of unwarranted hypotheses regarding such interplay between variables.

The first method we used is based on the standard Pearson correlation matrix [12], whose computation was performed with the Python Numpy library, that we controlled with two Pearson formulas, for discrete series and continuous series. Specifically, we used the following function:

\[
\text{numpy.corrcoef (df [cols].values.T)}
\]

where:

- df is the dataframe of the data

- cols is the list of columns used for the matrix

About Pearson correlation, it is a commonly formulated criticism that one may not establish a linear correlation between a series of quantitative variables and another one of qualitative variables. However, it will help us to identify those correlations as we are studying them [13,14,15].

Pearson's correlation evaluates the linear relationship between two continuous variables. A relationship is said to be linear when a modification of one of the variables is associated with a proportional modification of the other variable. Spearman's correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data. Therefore, in view of the continuous variables that we will study, a Pearson correlation seems more suitable to study the immediate impact of tweets on the economic parameters of Dogecoin and Litecoin.
The second method we used was based on mutual information entropy \cite{16, 17}, allowing us in particular to free ourselves from the limiting assumption of monotony required by linear correlation. It is a measure of the quantity of information (in the sense defined by Claude Shannon in 1948 \cite{18}) that two distributions share. In other words, it is a measure of the association ("clustering") between two variables: it is important to stress the fact that his approach is NOT linear correlation, but classical information entropy. In this approach, we compute a dimensionless quantity generally expressed in units of bits \cite{19}, which may be thought of as the reduction in uncertainty about one random variable given knowledge of another. For instance, high mutual information indicates a large reduction in uncertainty about one variable, given the other; whereas low mutual information indicates a small reduction about this uncertainty; and of course zero mutual information between two random variables entails no association between the two distributions\cite{20}. Furthermore, Shannon’s source coding theorem establishes strict bounds on what can be known about one data series might be compressed -- which in turn tells how to what extent one variable might be a proxy of another one without data loss. More broadly, Shannon information entropy has been demonstrated to be especially efficacious to evaluate algorithmic complexity when evaluated with the Block Decomposition Method \cite{21,22}. Moreover and according to N. N. Taleb, entropy metrics solve practically all correlation paradoxes in the field of social sciences (or rather, pseudo-paradoxes) \cite{23}. Another important example of the relevance of this technique is that of mother wavelet selection, where it demonstrated superior sensitivity to quantify the changes of signal structure than classical mean-squared error and correlation coefficient \cite{24}.

Therefore, in order to compute reproducible results, we use the “muinther” R package available on GitHub which uses these two statistical methods \cite{25}.

**Biases**

The first important bias is the community size. Indeed, that could impact Pearson method, more prone to these issues. However, a larger community will be able to reduce the extreme variations of the variables studied (number of tweets). Therefore, for the two methods, we will not be able to compare
the raw data of the samples but only the coefficients (from Pearson or from the normalized information theory) between these two cryptocurrencies.

The second bias is the Pearson method. Indeed, by its definition, Pearson's correlation evaluates the linear relationship between two continuous variables. A relationship is said to be linear when a modification of one of the variables is associated with a proportional modification of the other variable. However, if one moves in a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate[27]. In that case, Spearman’s correlation would be better.

**Prices prediction model**

**Settings**

We used historical data, spanning from 01/01/2015 to 03/11/2021, by extracting various economic trackers detailed below.

**Variables**

Three variables were primarily used: ‘date’, ‘tweets” (a continuous variable describing number of tweets by day with a mention of the interest cryptocurrency), ‘price’ (a continuous variable describing adjusted closure price of the interest cryptocurrency). If one of these values is missing, data was censored; in this way, we obtained 2482 data for each variable.

**Data sources**

Tweets were collected and extracted thanks to Twitter API and price were extracted on Yahoo Finance[4].
Statistical method

Prices correlation/causation with tweets number
Same method as above was used to explore correlation and causation relationship between Dogecoin/Litecoin price and tweets number. In other words, thanks to “muinther” R package, we used firstly Pearson correlation method and then Shannon mutual information entropy to assess that.

Application
So thanks to our correlation/causation analysis, we will be able to determine two functions to establish the relationship between currencies price and tweets number at some point $t$.

Models
Augmented Dickey-Fuller test
Since our Litecoin and Dogecoin data are time series data, it is important to check whether the data has any kind of stationarity. To check this we considered ADF which is an acronym for Augmented Dickey Fuller test. These unit roots are the reason for causing unpredictable results in time series data analysis. This ADF is a significance test, so a null and alternative hypothesis comes into play by that, test statistics are calculated and p-value is reported. Based on p-values, the stationarity of data is determined. Basically, ADF determines the trend of data and determines how strongly or weakly the time series is defined by a trend. However, we use three linear regression models to evaluate this:

1) The first type (type1) is a linear model with no drift and linear trend with respect to time:
   $$dx[t] = \rho*x[t-1] + \beta[1]*dx[t-1] + ... + \beta[nlag - 1]*dx[t - nlag + 1] + e[t].$$
   where $dx[t]$ is an operator of first order difference, i.e., $dx[t] = x[t] - x[t-1]$, and $e[t]$ is an error term.

2) The second type (type2) is a linear model with drift but no linear trend:
   $$dx[t] = \mu + \rho*x[t-1] + \beta[1]*dx[t-1] + ... + \beta[nlag - 1]*dx[t - nlag + 1] + e[t].$$

3) The third type (type3) is a linear model with both drift and linear trend:
   $$dx[t] = \mu + \beta*t + \rho*x[t-1] + \beta[1]*dx[t-1] + ... + \beta[nlag - 1]*dx[t - nlag + 1] + e[t].$$
Stationarity or stationary data means that data's statistical properties do not depend on time. And if suppose the given data is non-stationary then we need to change it to stationary by applying a natural log.

**ARIMA**

Model we choose to use is ARIMA, short for AutoRegressive Integrated Moving Average. ARIMA is a part of linear regression models, mainly used for predicting future values based on the past behaviour of the target. It is said that history doesn’t repeat itself but it surely has and had its own rhythm. The beauty of ARIMA models is that these do not use any exogenous values imposed on them but rather, are completely dependent on the past target values for prediction. ARIMA can be broken as AR, I and MA. As mentioned above, AR stands for Auto-Regressive, it works on the idea of regressing target on its past variable which is nothing but lagging on itself. Eq. (1) indicates a value Y is a linear function of its past n values. These n values can be chosen and are beta values used during fitting of the model. This equation can be made to forecast future values by simply making the following changes as in Eq. (2).

\[
Y = B_0 + B_1 \ast Y_{lag_1} + B_2 \ast Y_{lag_2} + \ldots + B_n \ast Y_{lag_n}
\]

(1)

\[
\hat{Y}_{forward} = B_0 + B_1 \ast Y + B_2 \ast Y_{lag_1} + \ldots + B_n \ast Y_{lag(n-1)}
\]

(2)

The Integrated part in ARIMA deals with stationarizing of data. Here, the differencing is applied on data as shown in Eq. (3), it indicates the future values of Y is some linear function of its past changes. Reason for differencing is that the time series data is not stable and the values of Y should have mean variance stationary.

\[
\hat{Y}_{forward} = B_0 + B_1 \ast (Y - Y_{lag_1}) + B_2 \ast (Y_{lag_1} - Y_{lag_2}) + \ldots
\]

(3)

Moving Average is all summarized in Eq. (4), somewhat similar to AN equation with lags. E indicates the error in the data is nothing but the residual derivations between the model and target value.

\[
Y = B_0 + B_1 \ast E_{lag_1} + B_2 \ast E_{lag_2} + \ldots + B_n \ast E_{lag_n}
\]

(4)
is the standard notation to represent ARIMA models. These parameters can be replaced with integer values to specify the type of the model used. Parameter ‘p’ is referred as the lag order of AR that’s the number of lags in \( Y \) to be included in the model, ‘d’ is the order of differencing required for making data stationary and ‘q’ is referred as order of MA that’s number of lagged forecast errors.

**Criterions**

There are several different models (i.e. based on the orders like AR and ARMA model if certain order or different order) for building time series. Lower the value obtained using these criterions, better suitable will be the model for our time series data. Parameters used in these criterions include - Log likelihood (L), signifies how strong the model is in fitting the data. Generally, it is considered that the more complicated the model is, the better it fits the data. It is true though in fitting but also leverages the concept of overfitting (i.e. model fits the training data better but loses its ability of generalizing on test data). To prevent that, incorporate the number of predictors \( k \) (i.e. number of lags (fixed amount of passing time)) plus a constant. Another parameter to consider here is \( T \), number of samples or observations used for estimation.

Below listed are the criterions used in this experiment for selecting best model-

- **Akaike’s Information Criterion (AIC)** - AIC is used to determine the order of an ARIMA model and also it can be useful for selecting the predictors of the regression model. AIC can be calculated using the formula given below,

\[ AIC = -2\log(L) + 2(p + q + k + l) \]

In cases with small values of \( T \), AIC may tend to predict too many predictors so in order to prevent that bias corrected version of AIC i.e. AICc should be considered.

\[ AICc = \frac{AIC + \frac{2(p+q+k+l)}{(p+q+k+l)(T-p-q-k-2)}}{T-p-q-k-2} \]

where, \( L \) is likelihood value, \( p \) is order of AR model, \( q \) is order of MA model, \( k \) is number of predictors and \( T \) is number of observations which is used for estimation as mentioned above.

In order to get the best model, we need to consider the model with low AIC value. That
means, the value of k should be less and the value of L should be maximum, illustrating that the model will be simple as k is small and fitting the data well with max l.

- **Bayesian Information Criterion (BIC)** - BIC also known as Schwarz information criterion is used for model selection on score obtained.

\[ BIC = AIC + [\log(T) - 2](p + q + k + 1) \]

Here also minimum value should be taken into consideration. BIC with small value illustrates that the model is simple with relatively less number of k, best fits the model and it is trained on a few number of observations. In addition to that, other studies have revealed that the ARIMA model has a strong potential for short-term prediction and can compete favourably with existing techniques for stock price prediction [29].

### Data availability

All data generated or analysed during this study are included in this published article (and its repository [26]).
Results

Correlation and causation analysis

Pearson correlation analysis

LITECOIN

All studied Pearson correlations for Litecoin (Supplementary File 1 and Figure 1.A.) were significant with a p-value under 0.001 excepting the correlation between the Litecoin market cap and the average litecoin transaction value (p-value= 5.938*10^-3).

Tweets have a small negative impact on Litecoin whales behavior (Pearson coefficient=-0.057). They are positively correlated with median Litecoin transaction value (0.449), average Litecoin transaction value (0.2944), Litecoin market cap (0.469), Litecoin transactions (0.296), and Litecoin active addresses (0.376).

Some results are surprising: Litecoin whales are negatively correlated with Litecoin active addresses (-0.398), with transactions (-0.439), with market cap (-0.466), with average Litecoin transaction value (-0.010) but not with median Litecoin transaction value (0.308).

DOGECOIN

All studied Pearson correlations for Dogecoin (Supplementary File 2. and Figure 1.B.) were significant with a p-value under 0.001.

Tweets are positively correlated to all economic variables: with median Dogecoin transaction value (0.534), with average Dogecoin transaction value (0.543), with Dogecoin market cap (0.549), with
Dogecoin whales (0.343), with Dogecoin transactions (0.376), and with Dogecoin active addresses (0.430).

Contrary to Litecoin, Dogecoin whales are positively correlated with Dogecoin active addresses (0.405), with Dogecoin transactions (0.436), with Dogecoin market cap (0.520), with average Dogecoin transaction value (0.452) and with median Dogecoin transaction value (0.476).

Mutual information theory analysis

LITECOIN

"Association" analysis

Community tweets are strongly (with a normalized mutual information coefficient of 0.9 at least) associated (Figure 2.A. and Supplementary Table 1) to all Litecoin variables but with fluctuant p-values.

Indeed, only few association p-values are significant: Litecoin average transaction value with tweets (p-value=0.0005), Litecoin average transaction value with whales (0.003), Litecoin active addresses with tweets (0.03), Litecoin transactions with tweets (0), Litecoin transactions with Litecoin active addresses (0.0005), Litecoin transactions with Litecoin average transaction value (0.016).

"Causation" analysis

We are going to explore and to emphasize only significant association causality (previously described).

Concerning tweets association with the economical trackers, the Litecoin transactions number impacted tweets [conditional information entropy of tweets given Litecoin transaction (0.637) is higher than the conditional information entropy of Litecoin transaction given tweets (0.070) as the Supplementary Table 1 shows], and the Litecoin average transaction value impacted tweets too [conditional
information entropy of tweets given Litecoin average transaction value (0.637) is higher than the conditional information entropy of Litecoin average transaction value given tweets (0.071)].

Concerning others associations, Litecoin active addresses is impacted by Litecoin transactions [conditional information entropy of Litecoin active addresses given Litecoin transactions (0.071) is higher than the conditional information entropy of Litecoin transactions given Litecoin active addresses (0.037)], and the Litecoin transactions are impacted by Litecoin average transaction value [conditional information entropy of Litecoin transactions given Litecoin average transaction value (0.071) is higher than the conditional information entropy of Litecoin average transaction value given Litecoin transactions (0.070)].

DOGECOIN

"Association" analysis

Community tweets are strongly (with a normalized mutual information coefficient of 0.9 at least) associated (Figure2.B. and Supplementary Table 2) to all Dogecoin variables but with fluctuant p-values.

Significant p-values (under 0.05) only concerned some associations: Dogecoin transactions with median Dogecoin transaction value (0.03), Dogecoin transactions with Dogecoin whales (0.003), average Dogecoin transaction value with Dogecoin market cap (0.011), average Dogecoin transaction value with tweets (0), Dogecoin whales with Dogecoin active addresses (3.41*10^-11), Dogecoin whales with tweets (3.22 * 10^-4).

"Causation" analysis

We are going to explore and to emphasize only significant association causality (previously described).

Concerning associations between tweets and Dogecoin economical trackers, average Dogecoin transaction value is impacted by tweets [conditional information entropy of average Dogecoin transaction value given tweets (0.861) is higher than the conditional information entropy of tweets given
average Dogecoin transaction value (0.048)] and tweets are impacted by Dogecoin whales [conditional information entropy of tweets given Dogecoin whales (0.861) is higher than the conditional information entropy of Dogecoin whales given tweets (0.124)].

Concerning others associations with whales, Dogecoin active addresses are impacted by Dogecoin whales [conditional information entropy of Dogecoin active addresses given Dogecoin whales (0.120) is higher than the conditional information entropy of Dogecoin whales given Dogecoin active addresses (0.029)]; Dogecoin whales are impacted by Dogecoin transaction [conditional information entropy of Dogecoin whales given Dogecoin transactions (0.124) is higher than the conditional information entropy of Dogecoin transactions given Dogecoin whales (0.049)].

Concerning other associations, Dogecoin transactions impacted the median Dogecoin transaction value [conditional information entropy of median Dogecoin transaction value given Dogecoin transactions (0.078) is higher than the conditional information entropy of Dogecoin transactions given median Dogecoin transaction value (0.049)]; and average Dogecoin transaction value is impacted by Dogecoin market cap [conditional information entropy of average Dogecoin transaction value given Dogecoin market cap (0.047) is higher than the conditional information entropy of Dogecoin market cap given average Dogecoin transaction value (0.001)].

**Prices prediction model**

Considering historical stock data of Litecoin/Dogecoin, we first analyzed the price and volume evolution over years. Since stock data is a time series data, it’s necessary to check this stationarity for model building.
Stationary testing

We used the ADF test for determining the property of our time series variable. To do so, we considered null and alternative hypotheses as- The representation of test data can be done using unit root or is non-stationarity.

ADF statistics and p-value of time series variables were calculated. The results (Table 1) were:
- Dogecoin non-stationary data show no trend in both of the three models (without a significant p-value).
- Litecoin non-stationary data show no trend in both of the three models but with a significant p-value only in the third one.
- Dogecoin stationary data show a significant trend absence in the first model, a non-significant trend absence in the third model and a non-significant trend presence in the second model.
- Litecoin stationary data show a non-significant trend presence in the first model, and a non-significant trend absence in the second and third model.

To make it stationary and remove the trend we simply applied natural logarithm on price values of Litecoin and Dogecoin. The values of time series before and after removing trends can be found in the table. In this way, the third model (a linear model with both drift and linear trend) was the best one to study possible trends in those data. Figure 3. illustrates the non-stationary data (without and with a square scale transformation) and stationary datas’ price v/s time line plot respectively studied with the third model.

Prices correlation/causation with number of tweets

Concerning Pearson correlation between currencies prices and tweets (Figure 4.A. and Table 2), Dogecoin prices are weakly correlated to Dogecoin community tweets ($r=0.29$) while Litecoin ones are strongly correlated to Litecoin community tweets ($r=0.86$). Those results are significant because our R algorithm computed a 0 rounded p-value.
Concerning association explored thanks to Shannon mutual information entropy (Figure 4.B. and Table 3), both Dogecoin and Litecoin links between their community tweets number and their price are strong (respectively 0.950 and 0.998). However, the first one is not significant (p-value=0.18) as the second one (p-value=0.15).

Concerning causality explored thanks to Shannon mutual information entropy (Table 3):
- Dogecoin tweets number impacted Dogecoin price. In fact, conditional information entropy of Dogecoin price given Dogecoin tweets number (0.724) is higher than conditional information entropy of Dogecoin tweets number given Dogecoin price (0.031).
- Litecoin tweets number is impacted by Litecoin price. In fact, conditional information entropy of Litecoin tweets number given Litecoin price (0.037) is higher than conditional information entropy of Litecoin price given Litecoin tweets number (0.002).

Application

Thanks to previous correlation/causation analysis, we could describe the relationship between currency price P and community tweets number C at a time t (Eq 8 and 9):

\[(8) \quad P_{\text{dogecoin}} = 0.0445124 \times C_{\text{dogecoin}} + 2.160866 \times 10^{-0.6} \times t - 0.0009299092 \]

\[(9) \quad P_{\text{litecoin}} = -36.44013 \times C_{\text{litecoin}} + 0.05336475 \times t + 41.35241 \]

Then, in order to analyze this regression quality, for each cryptocurrency (Figs 5. and 7., Tabs 4. and 5.) , we have performed ANOVA tests, compared Residual data vs fitted data, computed residual scale location, performed normality verification by comparing the quantiles of the population with those of the normal law, and compared residual data vs leverage data.

As the Pr(>F) is always under the alpha threshold (0.05), we reject the null hypothesis and therefore, non-null trends exist for both Litecoin and Dogecoin price. Furthermore, concerning Residual data vs fitted data studies, points are distributed randomly around the horizontal axis y = 0 and show no trend.
"Scale-Location" charts show slight trends which, however, are not obvious. According to “QQ-Norm” charts, residues are normally distributed. The last graph "Residuals vs Leverage" highlights the importance of each point in the regression; as we can see, there is only one point in each sample with a Cook distance greater than 1 (making the data suspect with that aberrant point).

We also have identified the suspect points which are the points whose studentized residual is greater than 2 in absolute value and / or the Cook's distance is greater than 1 (Figs 6. and 8.). In the latter case, the point contributes very / too strongly to the determination of coefficients of the model compared to others. However, there is no one-size-fits-all method for dealing with this type of stitches. That’s why a machine learning modelisation, as performed then, is required.

ARIMA

As given in Methods, ARIMA model predicts future values based on past behaviour. We considered certain criterion’s to know the condition of our model after training and the loss of information during training and to select the best model. Minimum loss indicates better training. Choices were automatically made thanks to the “auto.arima” function of R forecast package[28].

We choose about 0.99% of total litecoin/dogecoin stock data for training which is 24,482 samples in which we considered to predict all past values of litecoin/dogecoin pricing (Figure 13). We then calculated error (Eq. (10)) for each predicted past value based on forecast and actual values. Average errors of prediction were presented in Table 6.

\[
\text{Error} = \frac{|\text{forecasted value} - \text{testprice}|}{\text{testprice}} * 100
\]
Discussion

While average Dogecoin transaction value is impacted by tweets, Litecoin’s transactions number and average Litecoin transaction value impacted tweets. Concerning whales, tweets are impacted by Dogecoin whales but any significant relationship was found between Litecoin whales and tweets. Furthermore, this lack of association was clearly observed using one fundamental approach (mutual information theory), resting on wholly different assumptions and principles: one with classical Pearson correlation, the other with novel (in this field) Shannon mutual information. Furthermore, we observed that these two approaches had contradicted a few but mainly rather nicely complemented each other, in that Pearson correlation made it possible to study the sign of the correlation (positive vs. negative) while normalized mutual information made it possible to assess the association strength in a finer way, independent from the assumption of monotony required by linear correlation.

Concerning our prices study and prediction based upon community tweets volume, most accurate ARIMA models are non-seasonal and especially are white noise [30]. In this way, those non-stationary (e.g. with a log transformation) ARIMA(0,0,0) models are the most accurate to predict cryptocurrencies prices based upon their community tweet volume. White Noise is a random signal with equal intensities at every frequency and is often defined in statistics as a signal whose samples are a sequence of unrelated, random variables with no mean and limited variance. In this way, in our study this means that prices are independent between them (confirmed by prices correlation studies).

We surmised that the main criticism of our work would most likely be grounded in community size. Indeed, a larger and therefore more active community such as that of a “memecoin” could have a greater impact than a weaker community. Also a qualitative impact of some specific tweets by very well known people could be studied in order to better understand how twitter may impact cryptocurrencies values.

Few studies concerning behavioral impact on a cryptocurrency were made. The first concerns exclusively Bitcoin [31]. The second one is a comparison between Bitcoin or Dogecoin [1]; despite its main limitation that Dogecoin and Bitcoin are from different cryptocurrency generations, this study
concludes to an unpredictability of prices according to the tweets of a community. Therefore, we carried out the first study of behavioral impact analysis between comparable cryptocurrencies.

Our study has thus proven the interest of behavioral economics applied to the cryptocurrency world like others before it [31,32,33]. However, many things remain to be improved: the application of this modeling to cryptocurrencies operating under other consensus mechanisms (Proof of Stake (PoS), Delegated Proof of Stake (DPoS), Practical Byzantine Fault Tolerance (PBFT)...), the application to cryptocurrencies with a voting consensus (for such consensus, the formulation of the problem would be more meaningful, and from the point of view of applications to commodity backed cryptocurrency their study would be more meaningful), creation of an online bot to predict cryptocurrencies future prices…
References

1. Tandon C, Revankar S, Palivela H, Parihar SS. How can we predict the impact of the social media messages on the value of cryptocurrency? Insights from big data analytics. International Journal of Information Management Data Insights. 2021;1(2):100035.

2. "Ex-Googler Gives the World a Better Bitcoin". WIRED. Archived from the original on 2018-07-09. Retrieved 2017-10-25.

3. Gilbert, David (December 20, 2013). "What is Dogecoin? The Meme that Became the Hot New Virtual Currency". International Business Times. Archived from the original on September 6, 2015. Retrieved December 20, 2013.

4. Noyes, Jenny (January 22, 2014). "An Interview With The Creator Of Dogecoin: The Internet's Favourite New Currency". Junkee. Junkee Media. Retrieved January 12, 2021. Jackson Palmer accidentally invented Dogecoin in early December ... within the first 30 days there were over a million visitors to Dogecoin.com.

5. Locke, Taylor (July 14, 2021). "The co-creator of dogecoin explains why he doesn't plan to return to crypto: It's 'controlled by a powerful cartel of wealthy figures". CNBC. Retrieved July 20, 2021

6. Dogecoin Foundation (August 16, 2021). "Announcement: Re-establishing the Dogecoin Foundation". Dogecoin Foundation. Retrieved September 30, 2021.

7. A Cypherpunk's Manifesto (Eric Hughes, 1993)

8. Douziech T. Les mécaniques du consensus [Internet]. zonebourse. 2021 [cited 2021 Dec 4]. Available from: https://www.zonebourse.com/actualite-bourse/Les-mecaniques-du-consensus--36769932/?countview=0&fbclid=IwAR2Iy1v174eDAU6Y6mTxO6ZP4agG5QISfDUOzh0J4CwD77BHiOaYQiuvdsoU4

9. https://bitinfocharts.com ,

10. https://fr.tradingview.com ,

11. https://finance.yahoo.com ,
12. Caut J-L, Pébaÿ P, Lansiaux E, Forget J. COVID_Dept [Internet]. 2021. Available from: https://gitlab.com/covid-data-2/lockdown-and-curfew/-/blob/master/code/COVID_Dept.py

13. Lev J. The Point Biserial Coefficient of Correlation. Ann Math Stat. 1949;20.

14. Tate R. Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation. The Annals of Mathematical Statistics. 1954;25.

15. Kornbrot D. Point Biserial Correlation. In: Encyclopedia of Statistics in Behavioral Science [Internet]. American Cancer Society; 2005. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa485

16. Pébaÿ P. Bivariate-cardinalities [Internet]. NexGen Analytics; 2021. Available from: https://gitlab.com/covid-data-2/lockdown-and-curfew/-/blob/master/code/src_bivariate-cardinalities.py

17. Pébaÿ P, Caut J-L, Lansiaux E, Forget J. CF_Dept_entropy [Internet]. 2021. Available from: https://gitlab.com/covid-data-2/lockdown-and-curfew/-/blob/master/code/COVID_Dept_entropy.py

18. Shannon C. A Mathematical Theory of Communication. Bell Labs Technical Journal. 1948;27:379–423.

19. Thompson DC, Pebay PP. Parallel contingency statistics with Titan. 2009; Available from: https://www.osti.gov/biblio/993626

20. McDaid A, Greene D, Hurley N. Normalized Mutual Information to evaluate overlapping community finding algorithms. CoRR. 2011;

21. Zenil H, Soler-Toscano F, Kiani N, Hernandez-Orozco S, Rueda-Toicen A. A Decomposition Method for Global Evaluation of Shannon Entropy and Local Estimations of Algorithmic Complexity. Entropy. 2016;20.

22. Zenil H. A Review of Methods for Estimating Algorithmic Complexity: Options, Challenges, and New Directions †. Entropy. 2020;22:612.

23. Taleb NN. Fooled by Correlation: Common Misinterpretations Social ‘Science’. 2019 Mar; Available from:
24. Wijaya D, Sarno R, Zulaika E. Information Quality Ratio as a novel metric for mother wavelet selection. Chemometrics and Intelligent Laboratory Systems. 2016;160.

25. Lansiaux, Edouard; Caut, Jean-Luc; Forget, Joachim; Pêbaÿ, Philippe Pierre (2021): muinther. figshare. Software. https://doi.org/10.6084/m9.figshare.17161871.v1

26. https://github.com/edlansiaux/Behavioral-Cryptos-Study

27. de Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. Psychological Methods, 21(3), 273–290. https://doi.org/10.1037/met0000079

28. https://www.rdocumentation.org/packages/forecast/versions/8.15/topics/auto.arima

29. A. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014, pp. 106-112, doi: 10.1109/UKSim.2014.67.

30. Mahan, M. Y., Chorn, C. R., & Georgopoulos, A. P. (2015). White Noise Test: detecting autocorrelation and nonstationarities in long time series after ARIMA modeling. In Proceedings 14th Python in Science Conference (Scipy 2015), Austin, TX.

31. Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review: Vol. 1 : No. 3 , Article 1. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/1

32. Li, T. R., Chamrajnagar, A. S., Fong, X. R., Rizik, N. R., & Fu, F. (2019). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. Frontiers in Physics, 7, 98.

33. Beck, J., Huang, R., Lindner, D., Guo, T., Ce, Z., Helbing, D., & Antulov-Fantulin, N. (2019, May). Sensing social media signals for cryptocurrency news. In Companion proceedings of the 2019 World Wide Web conference (pp. 1051-1054).
Figure 1.A. Pearson’s correlation matrix concerning Dogecoin
Figure 1.B. Pearson’s correlation matrix concerning Dogecoin
Figure 2.A. Mutual information theory matrix concerning Litecoin
Figure 2.B. Mutual information theory matrix concerning Dogecoin
A. Trend analysis of non-stationary data

![Graph A](image1)

B. Trend analysis of non-stationary data with Y square scale

![Graph B](image2)

C. Trend analysis of stationary data

![Graph C](image3)

Figure 3. PRICE TREND ANALYSIS
Figure 4. Correlation and association heatmaps concerning tweets and cryptocurrencies price
Figure 5. Dogecoin regression residues analysis
Figure 6. Dogecoin regression Student residues and outliers detection
Figure 7. Litecoin regression residues analysis
Figure 8. Litecoin regression Student residues and outliers detection
Figure 9. Dogecoin non-stationary data
Figure 10. Dogecoin stationary data
Figure 11. Litecoin non-stationary data
Figure 12. Litecoin stationary data
Figure 13. Past price predictions
| Type                  | Column 1   | Column 2   | Column 3   | Column 4   | Column 5   | Column 6   |
|----------------------|------------|------------|------------|------------|------------|------------|
| Doge non-stationary  | -1.761199  | 0.110848   | 0.006812   | 0.205035   | 0.125470   | 0.322597   |
| Doge stationary      | -1.608607  | 0.089453   | 0.242513   | 0.193689   | 0.204278   | 0.779089   |
| LTC non-stationary   | -1.205218  | 0.229230   | 0.226548   | 0.223681   | 0.316928   | 0.078368   |
| LTC stationary       | 0.876762   | 0.892460   | 1.136216   | 0.653051   | 0.726769   | 0.692905   |

Table 1. Augmented Dickey-Fuller test results for the three types of linear regression models
| var1         | var2         | value_r   | value_p |
|--------------|--------------|-----------|---------|
| doge_tweets  | doge_tweets  | 1.0000000 | NA      |
| doge_price   | doge_tweets  | NA        | NA      |
| ltc_tweets   | doge_tweets  | NA        | NA      |
| ltc_price    | doge_tweets  | NA        | NA      |
| doge_tweets  | doge_price   | 0.2943733 | 0       |
| doge_price   | doge_price   | 1.0000000 | NA      |
| ltc_tweets   | doge_price   | NA        | NA      |
| ltc_price    | doge_price   | NA        | NA      |
| doge_tweets  | ltc_tweets   | 0.437468  | 0       |
| doge_price   | ltc_tweets   | 0.7331690 | 0       |
| ltc_tweets   | ltc_tweets   | 1.0000000 | NA      |
| ltc_price    | ltc_tweets   | NA        | NA      |
| doge_tweets  | ltc_price    | 0.5466496 | 0       |
| doge_price   | ltc_price    | 0.6070308 | 0       |
| ltc_tweets   | ltc_price    | 0.8451892 | 0       |
| ltc_price    | ltc_price    | 1.0000000 | NA      |

Table 2. Pearson correlation coefficients and p-value for tweets and cryptocurrencies price relationship
| X: Studied variable | Y: Studied variable | Computed marginal DMI at X | Computed marginal DMI at Y | DMI at X | X: Delta entropy, entropy of X | Information entropy of X | Conditional information entropy of Y given X | Conditional information entropy of X given Y | Mutual information of X and Y | Normalized mutual information of X and Y |
|---------------------|---------------------|-----------------------------|-----------------------------|----------|---------------------------------|-------------------------|---------------------------------------------|---------------------------------------------|-------------------------------|-----------------------------------------------|
| 0                   | 0                   | 2421                        | 2421                        | 2421     | 7.77705462                     | 7.77705462              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
| 0                   | 1                   | 2323                        | 2673                        | 2673     | 7.77703264                     | 7.31774994              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
| 0                   | 2                   | 2323                        | 2428                        | 2428     | 7.77702864                     | 7.38489332              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
| 1                   | 0                   | 2323                        | 2473                        | 2473     | 7.77702864                     | 7.38489332              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
| 1                   | 1                   | 2428                        | 2628                        | 2628     | 7.77702864                     | 7.38489332              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
| 1                   | 2                   | 2473                        | 2428                        | 2428     | 7.77702864                     | 7.38489332              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
| 2                   | 0                   | 2428                        | 2428                        | 2428     | 7.77702864                     | 7.38489332              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
| 2                   | 1                   | 2428                        | 2428                        | 2428     | 7.77702864                     | 7.38489332              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
| 2                   | 2                   | 2428                        | 2428                        | 2428     | 7.77702864                     | 7.38489332              | 0.00000000                                 | 0.00000000                                 | 0.00000000                   | 1.00000000                                 |
Table 4. Analysis of Dogecoin variance table

|     | Df  | Sum Sq   | Mean Sq  | F value | Pr(>F)  |
|-----|-----|----------|----------|---------|---------|
| time| 1   | 4.0240585921 | 4.0240585921 | 2726584.100 | 0.0000  |
| tweet | 1396 | 12.765686556 | 9.14475e-03  | 6196.971 | 0.0000  |
| Residuals | 1084 | 0.001595848 | 1.475874e-06 | NA        | NA        |
Table 5. Analysis of Litecoin variance table

|     | Df | Sum Sq | Mean Sq | F value | Pr(>F)  |
|-----|----|--------|---------|---------|---------|
| timel | 1  | 5080915.2 | 5080915.2177 | 6272.258210 | 0.000000e+00 |
| tweet | 1563 | 6050014.5 | 3970.7706 | 4.016537 | 4.74549e-104 |
| Residuals | 917 | 883720.7 | 963.7085 | NA | NA |


| ARIMA model                                    | Average error of prediction (%) |
|-----------------------------------------------|---------------------------------|
| Dogecoin non-stationary with ARIMA(2,1,2)     | 375.760549959756                |
| Dogecoin stationary with ARIMA(0,0,0)         | 0.222533877557071               |
| Litecoin non-stationary with ARIMA(1,1,2)     | 375.760549959756                |
| Litecoin stationary with ARIMA(0,0,0)          | 0.0954546657750234              |

Table 6. Average errors of prediction
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.pdf
- SupplementaryFile2.pdf