Università
Ca'Foscari
Venezia

Master's Degree

Economics and Finance

Ca' Foscari University of Venice

Double Degree

Financial Analytics

Stevens Institute of Technology

Final Thesis

Social Media Impact on Cryptocurrencies

Supervisor
Ch. Prof. Monica Billio
Dr. Ionut Florescu

Graduand
Giovanni Scalzotto
863480

Academic Year
2020 / 2021

Abstract

This paper studies the impact of Twitter on two cryptocurrencies: Dogecoin and Ethereum. Using hourly data from 01/01/2021 to 04/30/2021, various ARIMAX models are implemented to analyze the influence of exogenous variable as Twitter Volumes, Retweet, Reply and Like on the log-returns for the selected cryptocurrencies.

The model are performed follwing the requirements for stationarity condition and the absence of serial correlation on residuals.

Results show a strong impact of Twitter on Dogecoin while the impact for Ethereum has not been identified.

Introduction

Since the pandemic has started, social media have played an important role on the market because it has been the only way to stay connected around the world. In the past years, social networks were used only for interacting with people, expressing their own opinions and make new friends.

Recently, many studies have been made considering the impact of social media on the stock market and in the way they can drive the and influence it.

The purpose of this project is to understand and analyze how Twitter conditionates the blockchain, especially taking into consideration the second largest cryptocurrency, Ethereum, and a new cryptocurrency in term of fame as Dogecoin.

Section 1 of the paper describes how the data for Twitter and for the two cryptocurrencies have been downloaded from the Twitter API and Coin Base API. It illustrates how the data have been cleaned and adjusted for the purpose of the project.

Section 2 explains the methodology that have been used.

Firstly, the Cross Correlation is performed to analyze the highest periods of correlation between Twitter, Ethereum and Dogecoin. Secondly, the time series are checked for stationarity condition to remove any possible biases effects on them. Lastly, the ARIMAX models are designed to analyze and quantify the

impact of the different exogenous variables of Twitter on the returns of the two cryptocurrencies, Ethereum and Dogecoin.

Section 3 describes the results of the five different regressions, explaining which variables have the most important impact on returns and which cryptocurrencies is influence by Twitter, looking to RMSE, Standard Error and P-Values.

Section 4, finalize the project explaining the reasons behind those impacts and why one of the two cryptocurrencies is more affected by Twitter.

Literature Review

Recently, many professors and researchers have tried to analyze the impact of external factors on different sectors, as the stock market, technology, and cryptocurrencies.

On of the most important paper regarding the influence of the new technology as cryptocurrencies and social media on media platforms and cryptocurrencies have been by J. Abraham, D. Higdon et al. (2018).

They investigated and predicted the prices change in Bitcoin and Ethereum considering Twitter data and Google Trend.

The authors discovered through a linear model that the volumes of Twitter, rather than the sentiment of Twitter, can be used for prices direction.

In the same way, S. Coliannni, S. Rosales et al. (2019) have studied the possible uses of Twitter on Bitcoin to make trading strategies.

They performed various supervised learning algorithms such as Logistic Regression, Support Vector Machines and Naïve Bayes.

Results demonstrate that those models improve their strategy of 25%.

Furthermore, T. Ray Li, X. Fong, et al. (2019) studied Twitter signals to predict the prices fluctuation of a small-cap cryptocurrency, ZClassic.

The project used a Gradient Boosting Tree model and results have shown that the model had an accuracy of 80%, and it ha been used for price predictions.

Those project have been really helpful and they made a great contribution on the development of the ARIMAX models used to analyze the influence and impact of Twitter on Dogecoin and Ethereum.

In additon, they have been used as references to study the use of Twitter and design the paper.

1 Data Cleaning

1.1 Twitter

The first step to collect the data from Twitter was to use the Twitter Developers account that enables academic researcher to get access to the full archive of historical tweets and download 10,000,000 of tweets per month with a frequency of 500 tweets per second and a maximum requests of 500 every 15 minutes.

The Academic Research account provides a bear token for the Twitter API that allows the user, using different programming languages as Python or RStudio, to retrieve tweets and collect them in compressed files.

The account allows the researcher to collect data from the Full Archive search of Twitter and gather tweets related to different type of queries.

For example, it is possible to retrieve tweets for a specific author or ID, select a time period in which searching words or hashtags have been mentioned or select the language and location.

For the purpose of this project, it has been taken into consideration specific hashtags for a time period of 10 months.

The period starts form 07/01/2020 to 04/30/2021.

To retrieve tweets for cryptocurrencies as Dogecoin and Ethereum, the most accurate method to the select them have been to look for tweets where hashtags have been disclosed.

The hashtags that have been considered are:

- #Dogecoin
- #Doge
- #Ethereum
- #Eth

The API returned every tweet where the selected hashtags were mentioned plus the Timestamp, User ID, Retweet count, Reply count and Like count.
Retweet, Reply and Like count are additional features created from the API to enable the researchers to analyze how many times a specific tweet has been liked, re-posted or commented.
However, the Full Archive Search of Twitter does not provide the volume of tweets posted on a a given period.
To analyze how many tweets have been posted on a specific period of time, a column of ones has been added to each data collected to get the Volume them.
The final Twitter dataset is composed by 3,237,672 tweets on a second frequency form 07/01/2020 to 04/30/2021.

1.2 Cryptocurrencies

The second step to analyze the influence of Twitter on cryptocurrencies, the Coinbase API has been used to collect historical price for Dogecoin and

Ethereum.

Coinbase, one of the most important cryptocurrencies exchange in United States, allows the free download of historical price for cryptocurrencies throughout the usage of their API.

The Application Programming Interface allows to collect historical price for Dogecoin and Ethereum from 07/01/2021 to 04/30/2021.

The data was downloaded on an hourly time base.

Due to the fact that the cryptocurrencies are traded on the OTC, they do not include an open and close price.

In order to get a specific price in which analyze the data, the adjusted closing price has been calculated using the mid-price formula:

$$mid - price = \frac{(open\ price + close\ price)}{2}$$

In addition, the continuously compounded return has been calculated on the mid-price to analyze the performance of the two cryptocurrencies.

$$r = log(\frac{P_t}{P_{t-1}})$$

1.3 Final Datasets

The last step to get the final dataset in which to work with, the four different set of data have been combined in two dataset based on the cryptocurrencies' name.

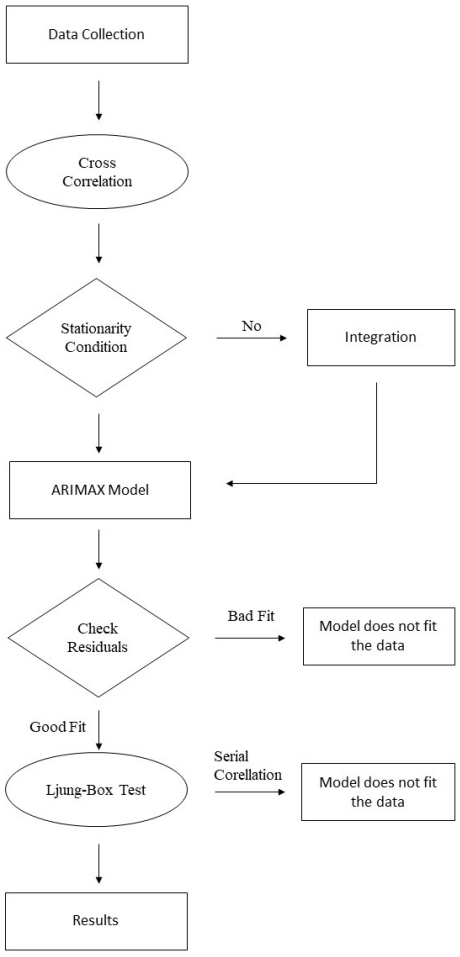
Saying that, the Twitter datasets have been rounded in hourly timestamp and the value of Retweet, Reply, Like and Volume have been aggregated.

Through RStudio, the datasets have been merged based on time and identified in which period there were the most important trends between volumes of Twitter, Price and Return of the cryptocurrencies.

The selected period starts from 01/01/2021 to 04/30/2021 with 6,830 hourly observations.

2 Methodology

With the data collected, cleaned, and adjusted, the project has been structured in the following steps:



2.1 Cross Correlation

The first step of the project is to analyze at which point there is the highest correlation between the Twitter variables, Dogecoin and Ethereum, the Cross Correlation function has been performed.

$$r_k = \frac{\sum_{i=1}^{n-k}(X_i - \overline{X})(Y_{i+k} - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Results show that the highest correlation values for Dogecoin appear at a time lag of 1, i.d. one hour, while for Ethereum appear at time lag of 2, i.d. two hours. In consequence, the variables of Twitter related to Dogecoin have been lag 1 hour before and for the variables related to Ethereum have been lagged 2 hours before.

Considering the Twitter Volume variables, if there is an increase in the tweets posted on the platform, it will influence the Dogecoin Price and Return one hour later while Ethereum will be influenced two hours later.

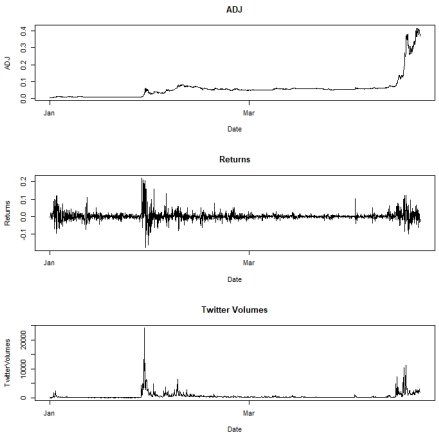


Figure 1: Dogecoin

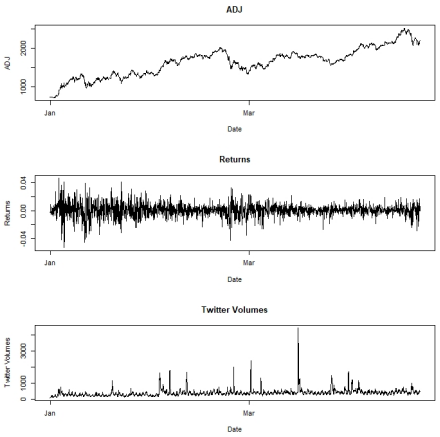


Figure 2: Ethreum

Looking to Figure 1 and 2, they display the impact of Twitter Volumes on cryptocurrencies Prices and Returns.

For Dogecoin, there is a strong impact of Twitter them at the end of February and at the end of April.

This impact is caused by the extreme influence of famous people, like Elon Musk, that tweets and creates posts related to cryptocurrencies, especially for Dogecoin.

Those tweets affect users because they trust successful people and, even though they are not informed on the blockchain, they post and invest money on it causing a large impact on the prices and returns of the crypto.

This is due to the fact that Dogecoin has started to be know in 2021 and it has been created as a 'joke' to make fun of cryptocurrencies, and having a low

market it is affordable to everyone.

Considering Ethereum, it seems that tweets regarding the currency do not affect returns and prices. In fact, it is the second largest cryptocurrencies for market capitalization and market value, second only to Bitcoin.

This mean that although there is high volume of tweets for Ethereum, this does not influence the currency because it is a well known cruptocurrency with a stable capitalization and market value.

Table 1 and 2 show the correlation coefficients of all the different variables for Ttwitter as Volumes, Retweet, Reply and Like on Dogecoin and Ethereum.

	Return	ADJ	Volume	Retweet	Reply	Like	Twitter Volume
Return	1	0.04264	0.2802	0.2136	0.1762	0.2156	0.2810
ADJ	0.0426	1	0.1393	0.2459	0.2949	0.2009	0.3303
Volume	0.2802	0.1393	1	0.3793	0.3534	0.3640	0.5508
Retweet	0.2136	0.2459	0.3792	1	0.8082	0.8882	0.7300
Reply	0.1762	0.2949	0.3534	0.8082	1	0.9347	0.6915
Like	0.2156	0.2009	0.3640	0.8883	0.9347	1	0.7033
Twitter Volume	0.2810	0.3303	0.5509	0.7301	0.6915	0.7033	1

Table 1: Correlation Matrix, Dogecoin and Twitter

	Return	ADJ	Volume	Retweet	Reply	Like	Twitter Volumes
Return	1	-0.1997	-0.0362	0.0318	0.0317	0.0209	0.0643
ADJ	-0.0199	1	-0.1817	0.2168	0.1622	0.1674	0.2703
Volume	-0.0362	-0.1816	1	-0.0237	-0.0038	0.0034	0.0170
Retweet	0.0317	0.2168	-0.0237	1	0.6969	0.5950	0.1594
Reply	0.0209	0.1622	-0.0038	0.6969	1	0.4559	0.1323
Like	0.0643	0.1674	0.0034	0.5950	0.4559	1	0.2787
Twitter Volumes	0.0284	0.2703	0.0170	0.1594	0.1323	0.2787	1

Table 2: Correlation Matrix, Ethereum and Twitter

For Dogecoin, there are a high correlation values of 33% and 28%, between the adjusted closing price, returns and Twitter Volumes.

In addition, there is a strong correlation between Retweet, Reply and Like variables on Dogecoin returns and prices with a coefficient of 21%, 18% and 21%.

For the same variables, Ethereum does not disply any significant correlation coefficients.

This means that there is an influence of Twitter on Dogecoin but not on Ethreum, implying that an increase in volumes of tweets will positively impact returns and the prices only for Dogecoin.

2.2 Stationarity Condition

Stationarity condition is one of the most important assumption in time series modelling because it allows to remove trend, seasonality and make the data independent.

To be stationary, a time series must follow three main criteria:

- $E(Y_t) = \mu$ constant for all T
- $Var(Y_t) = \sigma^2$ constant for all T
- $Cov(Y_{t_1}, Y_{t_2}) = E[(Y_{t_1} - \mu_{t_1}(Y_{t_1}] = \gamma(t_1 - t_2) = f(t_1 - t_2)$

  The autocovariance function between X_{t_1} and X_{t_2} only depends on the interval t_1 and t_2.

To check for stationarity condition, the Augmented Dickey- Fuller Test analyzes if a variable follows a unit root process i.d. nonstationary.

The null hypothesis (H_0) represents the variable that contains a unit root, while the alternative hypothesis (H_1) represents the variable that has been generated by a stationarity process.

The test rejects H_0 when the p-value of the Augmented Dickey-Fuller test is smaller than 0.05, meaning that the process is stationary.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta \Delta y_{t-1} + ... + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

$$H_0 : \gamma = 0$$

$$H_1 : \gamma < 0$$

$$ADF_\tau = \frac{\widehat{\gamma}}{SE(\widehat{\gamma})}$$

P-Value

	Twitter Volume	Retweet	Reply	Like
Dogecoin	0.0002	0.0076	0.0033	0.0098
Ethereum	0.0021	0.0002	0.0078	0.0053

Table 3: P-value Matrix for Dogecoin and Ethreum

Table 3 represents the results for the Augmented Dickey-Fuller test.

They shows that for Twitter variables connected to Dogecoin and Ethereum, all are stationary.

This means that the processes do not include any trend and patterns of seasonality, implying that there is not the need for integreation.

2.3 ARIMAX Model

The ARIMAX model is composed by the combination of the autoregressive process AR, integration for stationarity I, moving average process MA and the exogenous variable X.

$$y_t = \beta x_t + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 z_{t-1} - ... - \theta_q z_{t-q} + z_t$$

Rewrite with Backshift Operator:

$$y_t = \frac{\beta}{\phi(B)} x_t + \frac{\theta(B)}{\phi(B)} z_t$$

where

$$\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - ... - \theta_q B^q$$

The model includes y_t and y_{t-p} that represents the lag period for the AR process, z_t and z_{t-q} that represent the error terms for MA and X that explains the exogenous variable that identify and quantify the impact of an external factor on a time series.

Given a time series of data y_t, the ARIMAX model can be implemented to analyze past data for the time series and enables to forecast future value.

With the purpose of defining the impact of Twitter on Dogecoin and Ethereum, the exogenous variables took into consideration are:

- Twitter Volume
- Retweet
- Reply
- Like
- Combination of all factors together

The exogenous variable called X can be explained as an external factor that influences the endogenous variable y_t and allows to understand how the model performs with a new component.

In order to define which parameters have to be chosen to identify the best ARIMAX models, the Bayes Information Criteria has been performed for model selection.

BIC is used to find the best model among the set of candidate parameters selected.

$$BIC = -2LogLikehood + log(N) * k$$

where

N : number of examples

k : number of parameters

The best parameters (p, d, q) have been chosen running different models and identifing the one with the lowest BIC values.

Dogecoin		
	ARIMAX	BIC
Twitter Volume	(4, 0, 5)	-13,481.16
Retweet	(4, 0, 5)	-13,378.20
Reply	(3, 0, 5)	-13,344.86
Like	(3, 0, 8)	-13,344.63
All	(3, 0, 9)	-13,353.09

Ethereum		
	ARIMAX	BIC
Twitter Volume	(3, 0, 5)	-18,699.14
Retweet	(4, 0, 5)	-18,693.96
Reply	(3, 0, 6)	-18,697.45
Like	(3, 0, 3)	-18,701.42
All	(3, 0, 3)	-18,701.35

Having found the best parameters (p, d, q) for the ARIMAX models, the models have been run for the 5 different exogenous variables and determined the good fit of the models looking to the ACF, PACF of residuals and checked for the serial correlation of residuals using Ljung-Box Test.

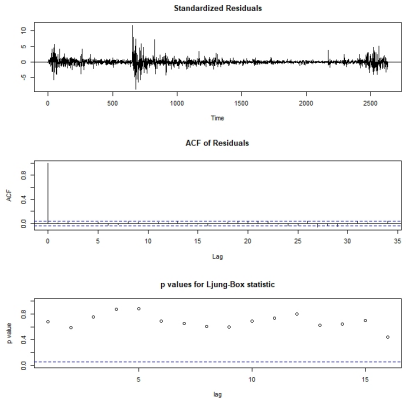


Figure 3: Dodecoin Volumes

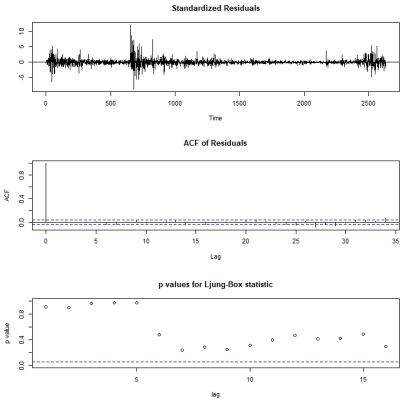


Figure 4: Dodecoin Retweet

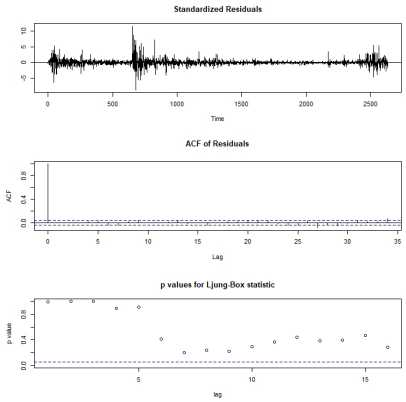


Figure 5: Dodecoin Reply

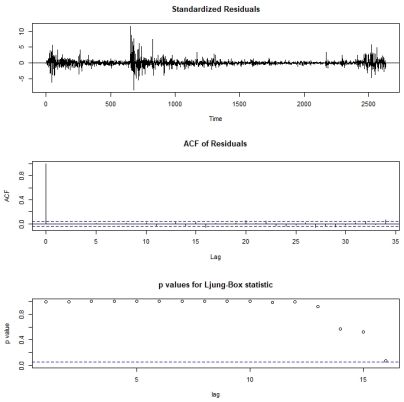


Figure 6: Dodecoin Like

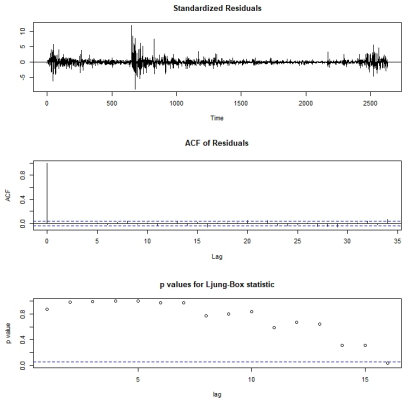


Figure 7: Dodecoin Combination

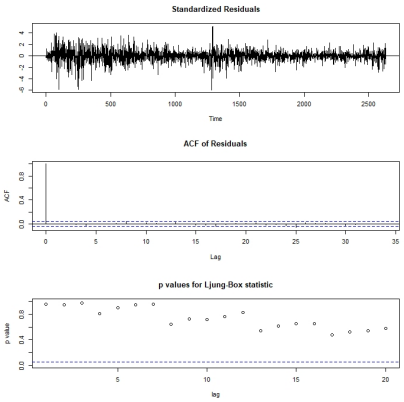


Figure 8: Ethereum Volumes

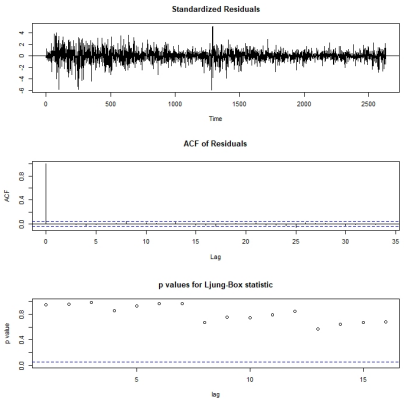


Figure 9: Ethereum Retweet

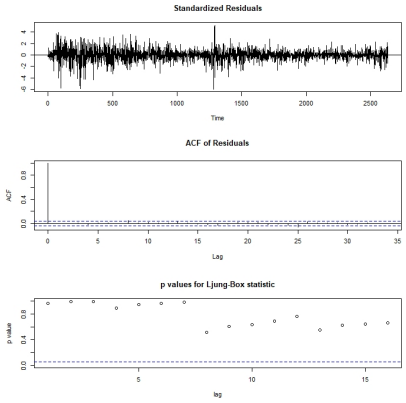


Figure 10: Ethereum Reply

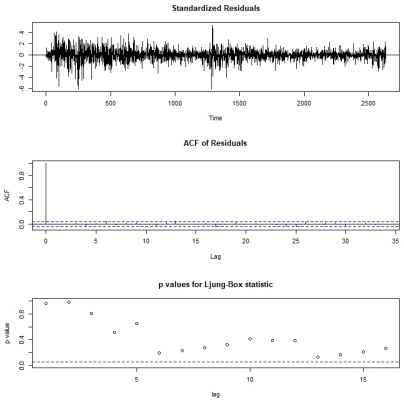


Figure 11: Ethereum Like

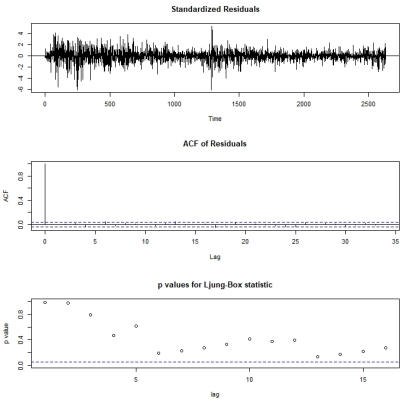


Figure 12: Ethereum Combination

For the purpose of determining the presence of serial correlation residuals, the ACF functions display from Figure 3 to Figure 10 show that for all the ARIMAX models with the 5 different exogenous variables have not autocorrelation on residuals.

This assumption is supported by the Ljung- Box Test.

The test calculates the absence of serial autocorrelation of residuals for different lag periods.

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\widehat{p}_k^2}{n-k}$$

H_0 : Model does not show lack of fit

H_1 : Model does show lack of fit

where

$$Q > X^2_{1-\alpha,h}$$

Ljung-Box Test

	Twitter Volume	Retweet	Reply	Like	All
Dogecoin	0.6995	0.4904	0.4681	0.4349	0.313
Ethereum	0.6489	0.6745	0.6463	0.2116	0.2198

Table 4: Augmented Dickey-Fuller Test for Dogecoin and Ethereum

Table 5 shows the values of the Ljung-Box test for the different ARIMAX models. The test rejects the null hypthones if the P-Values are greater than 0.05, meaning that for Dogecoin and Ethereum there is absence autocorrelation on residuals.

3 Results

After checking the stationarity conditions and satisfied all the requirements for the good fit of the models, the ARIMAX has been performed on the 5 different variables for both Dogecoin and Ethereum.

Dogecoin

	ARIMAX	RMSE	Stand. Error	P-Value
Twitter Volume	(4, 0, 5)	0.0183	0.0005	2.2e-16 ***
Retweet	(4, 0, 5)	0.0196	0.0003	0.0077 **
Reply	(3, 0, 5)	0.0157	0.0004	0.0372 *
Like	(3, 0, 8)	0.0287	0.0003	0.0026 **
All	(3, 0, 9)	0.0186	0.0001	5.5e-06 ***

Table 5: ARIMAX models for Dogecoin

Etherum

	ARIMAX	RMSE	Stand. Error	P-Value
Twitter Volume	(3, 0, 5)	0.0068	1.6048e-04	0.5366
Retweet	(4, 0, 5)	0.0068	7.1908e-05	0.6780
Reply	(3, 0, 6)	0.0054	7.2781e-05	0.4429
Like	(3, 0, 3)	0.0028	7.9856e-05	0.6412
All	(3, 0, 3)	0.0068	3.4923e-05	0.9557

Table 6: ARIMAX models for Ethereum

The Root Mean Square Error represents the standard deviation of residuals and helps to identify how the residuals are spread around the best fit model. The lower the value, the better the residuals are concentrated on the line of the ARIMAX model.

Results show that for Dogecoin and Ethereum the RMSE values are low, meaning that residuals fit good.

The Standard Errors, that represents the spread of data on the actual data, show low values meaning that the data have a good accuracy on the models.

Lastly, P-Values analyze the statistical significance of the models.

If the values are lower then 0.05, it implies that they are statistically significant.

For the purpose of the project, it measures if the exogenous variable X impacts the endogenous variables Y.

Results show that for Dogecoin all the P-Values are statistically significant, demonstrating that all the Twitter variables impact and affect the returns of the cryptocurrency.

On the other hand, there are not any significant P-Values for Ethereum, meaning that the cryptocurrency has not been influenced by Twitter variables.

Conclusion

The aim of the project was to identify any impact of Twitter on the two different cryptocurrencies: Dogecoin and Ethereum.

The digital currencies have shown completely different results since they have opposites applications on the market.

Ethereum is a decentralized application cryptocurrency in which users can interact with. It is used to create NFTs that are non-interchangeable token that can be linked to a digital work or object and traded as a unique digital asset. Furthermore, Ethereum is used as way of exchanging money.

Differently, Dogecoin was born as a joke to make fun of the largest and most known cryptocurrencies, Bitcoin.

The main uses of the altcoin are for digital and peer-to-peer payments.

Since 2021, Dogecoin started to be well known and gain a lot of visibility around social media like Twitter and Reddit because many famous people, as Elon Musk, started to talk on TV-shows and tweets about the crypto as the future for the world of cryptocurrency.

For this reason, Dogecoin became to be noticed, and well known around the world of investors and traders.

Having clarified that, results reflect perfectly the different behavior of the cryptocurrencies.

The data has been downloaded from 07/01/2020 to the end of 04/30/2021, collecting a total of 3,237,672 tweets and 6,830 values for Ethereum and Dogecoin. Due to the fact of time's inconsistency, tweets have been rounded to hourly as the price for cryptocurrencies and the Twitter Volumes has been created based on the tweets that have been made.

Starting from the correlation, tweets related to Dogecoin have demonstrated influence on returns, 28%, while tweets for Ethereum have not shown any influence of tweets on them, 2%.

This is caused by the fact that Ethereum is a well-known cryptocurrency with a stable capitalization and a delineate use on the market.

On the other hand, Dogecoin is a high volatile cryptocurrency with a spike in prices from January to April of 253%.

Thanks to its high volatile prices that can give high returns, Dogecoin started to get noticed around people especially on social media.

This has had a second effect on the market, because the more the people talk about the cryptocurrency the more the masses start to invest on it and make price growing fast.

In order to proof the influence of Twitter on Dogecoin and Ethereum, the ARIMAX models on the data collected have been implemented.

The first step has been to check the stationarity condition for all the different variables of Twitter using the Augmented Dickey Fuller test.

The test has shown that all the dummies are stationarity, with a constant mean, constant variance and no autocorrelation between lags.

This mean that there were not the need for integration to make the data stationary.

Having checked the stationarity condition, the ARIMAX model has been performed to analyze the impact of the different Twitter variables on returns of the cryptocurrencies.

The best models have been chosen looking to the lowest Bayesian Information and performed for all the 5 different exogenous varibles for Dogecoin and Ethereum.

Subsequently, the Ljung-Box Test has been used to check for serial correlation of residuals, showing that all the ARIMAX models had not lack of fit and performed good.

Finally, the P-Values have demonstrated that for Dogecoin there is a significant influence of all the different exogenous variable on returns, particularly for the Twitter Volumes.

On the other hand, the ARIMAX models considering the Ethereum's variables have not revealed any significant value.

This mean that for the second most known crypto there is not any impact or influence of Twitter.

This absence of impact can be explained by the fact that Ethereum has a stable

history and a remarkable value with secure application for specific purposes. Oppositely, Dogecoin is a new cryptocurrency on the market with recent renowned name on social media and internet.

In conclusion, the project can be used for future work, comparing the results with other models that can capture other types of impacts on cryptocurrencies. In addition, the ARIMAX model may be implemented to predict the direction of returns based on new adjustments as the high volatility nature of the cryptocurrencies and the trend that might appear in the following periods.

Bibliography

[1] Abraham Jethin, Higdon Daniel, NelSon John, Ibarra Juan. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis, SMU Data Science Review, 2018.

[2] Colianni Stuart, Rosales Stephanie,Signorotti Michael Signorotti. Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis, Standford University Journal, 2017.

[3] Hasitha Ranasinghe, Malka N. Halgamuge. Twitter Sentiment Data Analysis of User Behavior on Cryptocurrencies: Bitcoin and Ethereum, IGI Global Publisher, 2020.

[4] Jing Fan, Rui Shan, Xiaoqin Cao, Peiliang Li. The Analysis to Tertiary-industry with ARIMAX Model, Journal of Mathematics Research, 2009.

[5] Kraaijeveld Oliver, De Smedt Johannes. The predictive power of public Twitter sentiment for forecasting cryptocurrency prices, Journal of International Financial Markets, Institutions and Money, 2020.

[6] Mustafa Ali. ARIMA vs. ARIMAX – which approach is better to analyze and forecast macroeconomic time series, Academia, 2016.

[7] Tianyu Ray Li, Anup S. Chamrajnagar, Xander R. Fong, Nicholas R. Rizik, Feng Fu. Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model, Frontiers in Physics, 2019.

[8] Tudor-Mircea Dulac, Mircea Dalen. Cryptocurrency - Sentiment Analysis in Social Media, Sciendo, 2019.

[9] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, Chang Hun Kim. Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies, Plos One Journal, 2016.