Twitter and Reddit posts analysis on the subject of Cryptocurrencies

1st Ramón Hinojosa Alejandro
MCCi
Tecnológico de Monterrey, Campus Mty
Monterrey, Mexico
a01382300@itesm.mx

***Abstract*—Cryptocurrencies are digital assets for exchange between individuals. They are decentralized assets as they do not require a central authority to manage them, thus, like other exchanges, questions arise over it about which types of forces may influence its movement.**

The beginning of 2021 has gifted us with examples where social media influences the stock market and the cryptocurrency one, such as the GameStop case, or Elon Musk's tweets about cryptocurrency investment, respectively.

The present papers approaches a sentiment analysis on post published on the Twitter and Reddit platforms, to see its interactions with the cryptocurrency market.

There being hundreds of crypto assets, this paper only focuses on three: Bitcoin, Ethereum, and Litecoin.

***Index Terms*—Social Media, Cryptocurrency, Sentiment Analysis, Machine Learning**

I. Introduction

Cryptocurrencies are decentralized digital assets, that work as a medium of exchange between individuals. The history of cryptocurrencies dates to 1983, where the term "ecash" started to appear in scene by the cryptographer David Chaum [Chaum, 1983]. Nevertheless, the first cryptocurrency, bitcoin, was created until 2009, by a pseudonymous Satoshi Nakamoto [Brito and Castillo, 2013], [Berentsen and Schar, 2018].

There are many reasons why cryptocurrencies have become so popular nowadays. One of the reasons is that it is a decentralized asset, in other words, no government or main authority manages it. The transactions that are done with the cryptocurrency are stored in what is called Blockchain, a database of all transactions that have been done with the respective cryptocurrency. Anyone can access the Blockchain, that is why it is said that it is decentralized, because this database is of public record ownership.

As in many other markets, people have tried to predict its movements using different tools as Machine Learning ([Livieris et al., 2021], [Livieris et al., 2020]), or sentiment analysis ([Yasir et al., 2020]), where the incorporation of these new features brought an increase in accuracy levels for the proposed models.

Over the topic of sentiment analysis, significant cases have attracted the attention of analysts and investors due to social media influence on the market.

Elon Musk on January 2021 changed his tweeter biography to "bitcoin" creating a rise of the crypto asset value from $32,000 USD Dollars to $38,000 USD Dollars in matter of hours. Days later he talked about the Dogecoin, making not reference about investment on it, nevertheless, the Dogecoin rise too in value [Ante, 2021], reasons to it can also be attributed to Snoop Dogg's tweet which at the time responded to Musk's tweet, with a picture referring Doge coin [Popper, 2021].

In more recent times, Elon Musk commented negatively about the decentralization of the crypto assets and other subjects, comments that brought a downfall in prices for many cryptocurrencies.

Reddit is a social media platform, like Twitter, that gathers information from various sources and puts it in one place; one significant difference is that you do not follow people, but a topic of interest.

In the USA's stock market, a historic event took place at the ends of January 2021, where a hedge fund[1] short sold[2] GameStop stocks, making the stock price to lower, in response, a group of Reddit (Wallstreetbets), who have noticed what the hedge fund was doing, started to long stocks, significantly rising the price, driving the hedge fund to bankruptcy [Di Muzio, 2021]. Government intervention was necessarily to set regulations.

As seen by these examples, social media has become a powerful tool that affects financial markets. The present paper will not try to deepen on prediction models over the cryptocurrency market, but to see how the sentiment of these social media platforms relates with day-by-day returns.

Abdessamad et al. (2020) showed the importance of retweeting (reblogging, mechanism that allows users to share someone else's posts) as a method for propagations of information. Applied to the present project, retweeting can be seen as a mechanism for sentiment propagation. Huang et al. (2018) demonstrate how likes and comments show the sentiment opinion of users over hot topics, and how the most liked posts influence more users. More social media variables such as likes, comments, and shares are integrated into the analysis.

[1]Hedge fund is an investment fund that trades in liquid assets and uses complex investment tools like short-selling, leverage, etc. [Lemke et al., 2015]

[2]Investment methodology where profits are made from stock prices falls. You borrow stock from the market, at a price, expecting it to fall, to re-buy it at a lower price and get a profit from the difference. [Elder, 1993]

By December 2020, according to CoinMarketCap, there exits around 7,800 cryptocurrencies, and rising [Septhon, 2020]. This paper will only be focused on Bitcoin (BTC), Ethereum (ETH), and Litecoin (LTC).

II. Methodology

A. Data

Cryptocurrencies prices are extracted from the Binance cryptocurrency exchange. Binance is the top cryptocurrency exchange according to CoinMarketCap at the time of March 2021. CoinMarketCap ranks and scores exchanges based on the following: Web Traffic Factor; Average Liquidity, Volume, as well as the Confidence that the volume reported by an exchange is legitimate.

For sentiment analysis, we are required to extract posts from Twitter and Reddit. Apart from it, relevant information about them such as the number of likes, comments, and shares are also recollected.

The posts are scrapped considering the name of the cryptocurrency and the symbol of it; the corresponding keywords are bitcoin, BTC, ethereum, ETH, litecoin, and LTC.

Twitter is a social media platform that according to [Sayce, 2020], 500 million tweets are posted every day, thus an immense number of tweets can be expected from the tags considered. The Twitter API (Application Programming Interfaces) only allows a search for 450 tweets in a window of 15 minutes, this rate limit implies a constraint for the present project, therefore a Twitter crawler is applied. A Twitter crawler will allow us to recover present and older tweets without compromising the Twitter servers.

For Reddit, 303.4 million posts are produced per month [Hutchinson, 2020], less than Twitter. The Reddit API does have search limitations, nevertheless, due to the lower number of posts produced per day, the API is more than enough to fulfill what is required.

It will not only be considered the sentiment of the posts published on these social medias, the number of comments, likes, and shares too. The data hold information from March 1 to June 8, 2021.

B. Data Preparation

The data preparation process starts by cleaning the corresponding posts from emoticons, special characters, and avoiding exclamation and interrogation symbols.

Many research papers have approached the problem that social media bots impose on sentiment analysis methodologies on the cryptocurrency subject, [Wright and Anise, 2018] [Kraaijeveld and De Smedt, 2020], or others [Anwar and Yaqub, 2020]. In [Kraaijeveld and De Smedt, 2020] four heuristics are proposed for bot's identification on Twitter, if a post meets two of the next considerations, then it is said to be a bot. The next rules will be applied for both Twitter and Reddit posts.

- The post contains "give away" or "giving away".
- The post contains "pump", "register", or "join".
- The post contains more than 14 hashtags.
- The post contains more than 14 ticker symbols.

With the remaining posts, a sentiment analysis is run based on the Valence Aware Dictionary and sEntiment Reasoner (VADER) python-tool [Hutto and Gilbert, 2014]. For the only case of the Reddit posts, two sentiments are calculated, Reddit posts are composed by a title and the text, the sentiment of both is computed.

The present paper tries to see the relation of sentiment analysis and the cryptocurrency market movement, so, our target variable is defined as the lead net returns.

Market data is characterized by being nonstationary, one way of dealing with this is differentiating data, this project will inquire over the relation of the High and Low price, percentage difference between them is computed. According to Chong et al. (2017), past returns have proven to have a predictive power which can be exploited by machine learning techniques.

The data of the cryptocurrencies is concatenated to form a panel data dataset. Panel data is multidimensional data that involves measurement over time of multiple phenomena of the same subject [Diggle et al., 2002].

C. Statistical Tests

To see the relation the sentiment of social media has with cryptocurrencies returns different statistical tests are applied.

1) Correlation : Correlation is a statistic that measures the degree at which two variables move in a similar direction. This will allow us to see how correlated the social media variables with the market data are.

2) Multicollinearity: Multicollinearity is another statistical measure, that will complement us with the results observed with correlation, telling how correlated the independent variables are.

There can be multicollinearity but no correlation. When making a regression model, it will try to search for parameters that represent how much the dependent variable changes with a unit of change in the independent variables, while maintaining the remaining constants. If a model detects multicollinearity between two variables, then, while it tries to set the parameter for one of these variables, the other will change too, influencing the inferences that can be made with the model.

The VIF (Variance Inflation Factor) is used to detect it.

3) Homoskedasticity: Homoskedasticity is a statistical characteristic that tells if a variable has a finite variance. In a regression model, it is applied to the residuals to see how the error terms vary according to the dependent variable, in this case, if homoskedasticity is present, then our regression model can explain good the changes in the dependent variable as the residual does not differ for each observation, but if homoskedasticity is violated, leading to heteroskedasticity, more variables are needed to explain the dependent variable.

The White test and Breusch-Pagan test are used parting from a Pool OLS (Ordinary Linear Squares) model. The White test allows us to identify non-linear heteroskedasticity.

For the Pool OLS, and the regression models that follow, data is normalized with the Min-Max method.

4) Endogeneity: Endogeneity is present when the predictor variables are correlated with the error term. The reasons for it to be present are that a loop of causality is present between the independent and dependent variables, or when variables are omitted.

Exogeneity is the opposite of endogeneity and refers to the fact that the independent variables are not dependent on the dependent variable.

The analysis conducted on the present paper tries to see the relation of social media variables towards crypto returns, ignoring that the same returns can influence the sentiment of investors on social media. This dual causality is the one to be tested with the present statistical measure.

It is tested using the Hausman test on a FE (Fixed Effects) and RE (Random Effects) model.

D. Lead net returns prediction

Based on the observations given with the statistical tests, a prediction of lead net returns is carried out.

Panel data models such as Pool OLS, FE and RE, and Machine Learning regression techniques such as Random Forest (RF), AdaBoost (AB), k-Nearest Neighbors (kNN), and Multilayer Perceptron are applied.

Train and test dataset are formed with an 80-20 ratio. RMSE (Root Means Squared Error) and MAE (Mean Absolute Error) will be used to compare the models.

III. Results

Figures 1 and 2 shows the distribution of sentiment over the lead log returns (leadnet) of the three cryptocurrencies. In both cases we can see how the sentiment distribution of Bitcoin is on top, and Litecoin at the bottom, this is expected due to their popularity. Another observation to be made, there is no clear correlation that higher returns bring more positive returns.

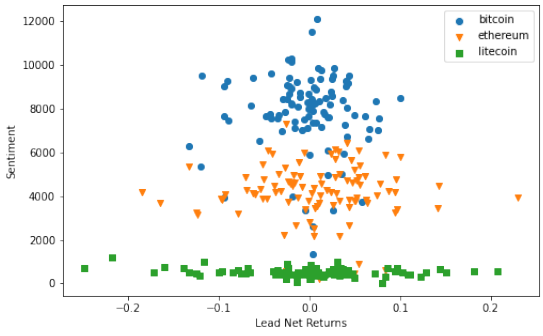

Fig. 1. Log returns and Twitter sentiment

A. Correlation

Figure 3 shows how the features of each social media are highly correlated between them, and even the strong correlation between both social media.

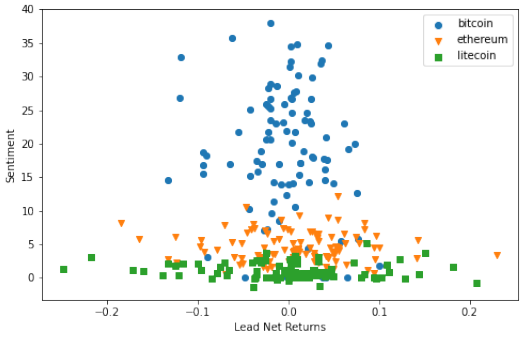

Fig. 2. Log returns and Reddit sentiment

No social media feature showed a high correlation with the market data. Lastly, Volume is positively correlated to the percentage difference of the High and Low prices (H-L), and negatively with the social media variables.

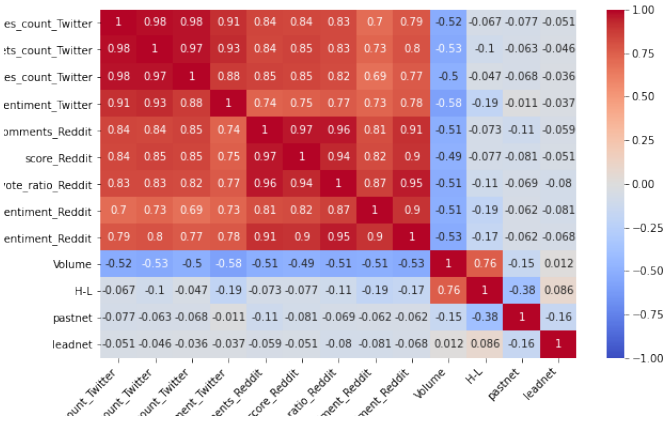

Fig. 3. Person's correlation Heatmap

B. Multicollinearity

VIF values closer to zero represent that there is no multi-collinearity between that variable and another one, between 1 and 5, a moderate correlation, above 5, concern of correlation, and above 10, a problematic correlation [Choueiry, 2021]. Table I shows the VIF values obtained of the independent variables.

For Test 1, it shows high values on VIF, which represents multicollinearity. Iteratively this test is repeated while removing the highest VIF value attribute until no multicollinearity is seen, represented by Test 2.

Our dataset is left only with the sentiment of Twitter, sentiment of the titles of post in Reddit, the score of those posts, the Volume of transactions, and past net returns.

C. Homoskedasticity

For homoskedasticity, the White and Breusch-Pagan tests are carried out. Table II shows the results of these two tests

	Normal Data	
Feature	Test 1	Test 2
replies_count_Twitter	75.42	-
retweets_count_Twitter	81.00	-
likes_count_Twitter	55.93	-
sentiment_Twitter	24.84	4.86
comments_Reddit	50.67	-
score_Reddit	27.73	5.9
upvote_ratio_Reddit	44.79	-
title_sentiment_Reddit	10.41	5.89
text_sentiment_Reddit	30.43	-
Volume	9.17	1.17
H-L	12.11	-
pastnet	1.30	1.03

TABLE I
VIF FACTORS

considering the LaGrange Multiplier (LM) and the F-statistic.

For both, p-values below a 5% significant level are seen, rejecting their null hypothesis of the presence of homoscedasticity, indicating heteroskedasticity. Our model is incapable of explaining our target variable, meaning that extra variables are required.

Test	LM T-stat	LM P-value	F T-stat	F P-value
White test	54.649	0.0	3.120	0.0
Breusch-Pagan test	35.777	0.0	7.987	0.0

TABLE II
HOMOSKEDASTICITY TESTS

D. Endogeneity

Lastly, endogeneity is tested. Based on the p-value (91%) seen in Table III, we accept the null hypothesis that our variables are endogenous, they are exogenous, they are no correlated with the error term.

Chi-squared	Degrees of Freedom	P-Value
0.767	6	0.993

TABLE III
HAUSMAN TEST ON FE AND RE

E. Lead net returns prediction

The corresponding parameters for each regression model are the next ones:

Table IV shows a comparative of the different prediction methodologies used. For both metrics, we see similar ranges for all models considered. These metrics are in units of our target variable, meaning if a return of 1% is estimated, we can still expect return between the range of 7% to –5%.

IV. DISCUSSION

From the distribution plots we could gather that there is no clear evidence that higher sentiment correlates with more positive returns, observation that was corroborated by the correlation matrix. In it, we too could see that the social media variables were positively correlated between them, meaning that the sentiment of investors is shared over these platforms. Lastly, from a higher sentiment, lower quantity transacted can be expected, due to the negative correlation between them.

name	RMSE	MAE
RE	0.074	0.053
FE	0.075	0.053
PoolOLS	0.074	0.053
RF	0.076	0.057
AB	0.072	0.051
kNN	0.073	0.051
MLP	0.070	0.048

TABLE IV
PREDICTION ERRORS

Multicollinearity demonstrated how certain predictor variables were dependent between them, leaving us to only consider the sentiment of the corresponding posts. Something to mention is that the score of the Reddit posts, unlike the amounts of comments and upvotes, do not show dependency with any other variables considered, even when a strong positive correlation was shown.

The proposed model lacked homoskedasticity, meaning that the considered variables were not sufficient to explain the expected returns of the cryptocurrencies, which was corroborated by the prediction tests.

We could also see that the proposed variables were exogenous, they were not correlated with the residuals of our model, and no dependent of the target variables, still, this last statement needs to be further examined.

V. CONCLUSIONS

The relation of social media variables related to the topic of cryptocurrency (Bitcoin, Ethereum, and Litecoin) and their market was analyzed. Different statistical tests were performed to see the dynamics of these variables, how the relate between them, and towards the next day returns (lead net returns).

From the tests computed, we could see that as much as the social media variables prove a certain level of information to predict future cryptocurrency price movements, they were not enough, information is still lacking, by which we can determine that these price movements do not only depend on the investor's sentiment, but on other factors as macroeconomic variables, political landscapes, etc.

VI. FUTURE WORK

For future work, more social media platforms and news sources are to be considered, to see if a better grasp of the sentiment of investors can be enclosed.

Analyzed the long-run influence instead of the short-run, as the one focused on this paper.

And the application of methodologies to augment the quality of the social media variables, as data propagation weights.

ACKNOWLEDGMENT

Special thanks to PhD. Francisco Javier Cantú Ortiz and PhD. Héctor Gibrán Ceballos Cancino, Data Science professors on Tecnológico de Monterrey, for their guidance in the development of this project, and help in results interpretation.

REFERENCES

[Ante, 2021] Ante, L. (2021). How elon musk's twitter activity moves cryptocurrency markets.

[Anwar and Yaqub, 2020] Anwar, A. and Yaqub, U. (2020). Bot detection in twitter landscape using unsupervised learning. In *The 21st Annual International Conference on Digital Government Research*, pages 329–330.

[Berentsen and Schar, 2018] Berentsen, A. and Schar, F. (2018). A short introduction to the world of cryptocurrencies.

[Brito and Castillo, 2013] Brito, J. and Castillo, A. (2013). *Bitcoin: A primer for policymakers*. Mercatus Center at George Mason University.

[Chaum, 1983] Chaum, D. (1983). Blind signatures for untraceable payments. In *Advances in cryptology*, pages 199–203. Springer.

[Chong et al., 2017] Chong, E., Han, C., and Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205.

[Choueiry, 2021] Choueiry, G. (2021). What is an acceptable value for vif? https://quantifyinghealth.com/vif-threshold/ . Accessed: 2021-05-20.

[Di Muzio, 2021] Di Muzio, T. (2021). Gamestop capitalism. wall street vs. the reddit rally (part i).

[Diggle et al., 2002] Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Heagerty, P. J., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

[Elder, 1993] Elder, A. (1993). *Trading for a living: psychology, trading tactics, money management*, volume 31. John Wiley & Sons.

[Essaidi et al., 2020] Essaidi, A., Zaidouni, D., and Bellafkih, M. (2020). New method to measure the influence of twitter users. In *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pages 1–5. IEEE.

[Huang et al., 2018] Huang, Y.-P., Hlongwane, N., and Kao, L.-J. (2018). Using sentiment analysis to determine users' likes on twitter. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 1068–1073. IEEE.

[Hutchinson, 2020] Hutchinson, A. (2020). Reddit shares insights into top posts, trends, users numbers and more. https://www.socialmediatoday.com/news/reddit-shares-insights-into-top-posts-trends-users-numbers-and-more/591822/: :text=Here's%20a%20look%20at%20the,million%20posts%20%E2%80%93%20up%2052.4%25%20YoY . Accessed: 2021-03-23.

[Hutto and Gilbert, 2014] Hutto, C. and 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*, volume 8.

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

[Lemke et al., 2015] Lemke, T. P., Lins, G. T., Hoenig, K. L., and Rube, P. S. (2015). *Hedge funds and other private funds: Regulation and compliance*. Thomson Reuters Westlaw.

[Livieris et al., 2021] Livieris, I. E., Kiriakidou, N., Stavroyiannis, S., and Pintelas, P. (2021). An advanced cnn-lstm model for cryptocurrency forecasting. *Electronics*, 10(3):287.

[Livieris et al., 2020] Livieris, I. E., Pintelas, E., Stavroyiannis, S., and Pintelas, P. (2020). Ensemble deep learning models for forecasting cryptocurrency time-series. *Algorithms*, 13(5):121.

[Popper, 2021] Popper, N. (2021). Elon musk and snoop dogg push cryptocurrencies to record highs. https://www.nytimes.com/2021/02/08/technology/dogecoin-bitcoin-elon-musk-snoop-dogg.html . Accessed: 2021-02-18.

[Sayce, 2020] Sayce, D. (2020). The number of tweets per day in 2020. https://www.dsayce.com/social-media/tweets-day/ . Accessed: 2021-03-23.

[Septhon, 2020] Septhon, C. (2020). How many cryptocurrencies are there?. https://currency.com/how-many-cryptocurrencies-are-there . Accessed: 2021-02-17.

[Wright and Anise, 2018] Wright, J. and Anise, O. (2018). Don't@ me: Hunting twitter bots at scale. In *Blackhat USA 2018*.

[Yasir et al., 2020] Yasir, M., Attique, M., Latif, K., Chaudhary, G. M., Afzal, S., Ahmed, K., and Shahzad, F. (2020). Deep-learning-assisted business intelligence model for cryptocurrency forecasting using social media sentiment. *Journal of Enterprise Information Management*.