**Business Analytics Report**

**August 31, 2021**

**SHERRY AGAGRWAL**

**CID - 01948430**

**Analyzing The Influence Of Twitter Sentiments On Cryptocurrencies – Bitcoin And Ethereum**

Abstract

Table of Contents

Abstract1

Table of Contents1

1. Introduction1
2. Related Literature4
3. Method1

3.1 Data Preparation2

3.1.1 Timeframe3

3.1.2 Twitter Data3

3.1.3 Financial Data3

3.2 Sentiment Analysis3

3.3 Final Dataset3

3.4 Predictive Modeling3

3.4.2 Correlation Analysis3

3.4.2 Classification Models3

3.5 Model Evaluation3

4. Results & Analysis4

4.1 Exploratory Analysis3

4.2 Correlation Results3

4.3 Classification Model Results3

4.4 Causality Analysis Results3

**5. Discussion5**

**6. Limitations6**

**7. References6**

**Appendix3**

1. **Introduction**

Cryptocurrencies are storming the headlines almost every week. With tech leads like Elon Musk changing twitter biography to #bitcoin, cryptocurrency has gained much traction. If one is not a trader or crypto enthusiast, these things will not make much sense. So, why trader community is so interested in the cryptocurrency market, and how can it change the future of finance?

Cryptocurrency is a decentralized method of sending money over the internet that gained popularity after Satoshi Nakamoto (a pseudonym) published Bitcoin, the first cryptocurrency, in 2008. It is a digital asset that can be sold, bought, and traded on various cryptocurrency exchanges. Its value is highly speculative as it is determined by demand and supply. It is based on Blockchain technology, in which every node in the distributed network (usual computers) validates the transactions, and there is no central authority such as banks to regulate the transactions. The way cryptocurrency is transacted and created is using 'mining.' This process is critical in generating new cryptocurrencies and maintaining the blockchain ledger and involves solving various computational heavy mathematical equations (1).

Until 2011, Bitcoin was the only cryptocurrency, and while it still has the greatest market capitalization of all cryptocurrencies, other alternative coins known as altcoins were developed to increase cryptocurrency security, privacy, and speed. There are currently over 4000 cryptocurrencies, including Ethereum, Litecoin, and Dogecoin, to mention a few (2). As anyone can invest in cryptocurrencies, they have emerged as a new asset class with a market capitalization of $2 Trillion as of August 2021 (3). Bitcoin dominates the crypto market with a 44% share of the capital, followed by Ethereum with an 18% market share (3).

Though crypto is a fast-growing asset and Bitcoin has seen a meteoric rise in its value, price volatility and speculation are cited as the two undesirable features for investing. According to the article by the Bank of England (4), oil prices did not change by more than 10% in one day from 2014 to the beginning of 2018, unlike the value of Bitcoin, which changed significantly – rising by 65 percent in one day and falling by 25 percent the next, indicating that crypto investment is not for the faint of heart and requires high-risk tolerance.

Since its inception, crypto volatility has mainly been influenced by social media channels, with Twitter being a central platform for disseminating crypto news and public conversations to the point that Forbes coined the term 'Crypto twitter' to describe it. It is a world that provides free thinkers with a centralized platform to connect and learn from one another, as well as insights and perspectives from some of the world's finest financial brains. As a result, many crypto early adopters are millennials who engage in Twitter debates and make trading decisions based on social opinion. Crypto traders monitor twitter to anticipate the ebb and flow of prices depending on broader public sentiment.

This report tries to capture the influence of Twitter sentiments on crypto returns, specifically for Bitcoin and Ethereum. A further attempt is made to understand the correlation between stock markets and cryptocurrencies\*\* The rationale for this study is further supported by existing related research work showing promising results in a similar space.

Why

The tool is also helpful for Bitcoin daytraders or short-term traders that can use the real-time Bitcoin Twitter sentimen tdata to determine entry and exit points for their trades

The tool is very good at showing the market perception of Bitcoin long term as well. We know that Bitcoin is highly volatile, but increasing positive tweet sentiment and volume over a long period of time could indicate upward growth, stability and acceptance.

To what degree public opinion influences bitcoin

Bitcoin

Why so important? Like money, % share of markets

1. **Literature Review**

This section reviews the exiting work on this topic and discusses how it is used as a foundation for the analysis done for this project.

One of the most prominent concepts in market trading is efficient market hypothesis(EHM) which states that stocks always trade at fair value and it’s very difficult to time the buy/sell strategy based on expert knowledge. The only way to increase the returns is to invest in riskier investments. When there isn't an open market, liquidity is low, or behavioural biases are imposed, however, this hypothesis fails. Cryptocurrency markets are decentralised markets with a set supply and no central authority. As a result, crypto presents a significant challenge to the EHM theory.

Furthermore, the primary source of information for crypto investors is social media and press coverage of the digital currency, with Bitcoin being published via a crypto mailing list and Ethereum being unveiled at a conference, as opposed to market stocks traded on national exchanges. With more than 4000 crypto currencies existing as of January 2021 (9), twitter hosts official channels of many of these crypto currencies for open conversations and news dissemination inducing human emotion when making decisions.

Many attempts have been made to understand the impact of twitter sentiments on market behaviour. Rao & Srivastava (2012)(10) investigated the complex relationship between twitter sentiments and financial market instruments and discovered a 0.88 correlation for returns. They used granger causality analysis and observed that twitter debates had a significant impact on stock price movement in the short run. In the same vein, Nisar & Yeung (2017) found promising evidence of correlation between general public mood and price changes of FTSE 100 index during a UK political event.

Similar studies have been extended to cryptocurrencies by exploring their relationship with public sentiments on Twitter. Colianni et.al 2015 used supervised machine learning algorithms to predict hour-to-hour and day-to-day predictions of Bitcoin price movements. They achieved 90% model accuracy when used tweet text as the input feature and 86% accuracy when used third party text sentiment API as the feature vector. Similarly, Valencia et at. 2019 did a similar study and used common machine learning tools including support vector machines (SVMs), neural networks (NNs) and random forest (RF) to predict price movements for four crypto currencies using different input features and found NNs outperforming other models with accuracy of 72% for Bitcoin, 44% for Ethereum and 64% for Ripple. Otabek el at. 2020 implemented Random Forest Classifier to classify the prices direction based on twitter sentiments and achieved 62.48% accuracy. Kraaijeveld, O & De Smedt, J 2020 expanded the study of understanding the predictive power of twitter sentiments on multiple crypto currencies and developed an approach to identify tweets by bots. They implemented bivariate granger causality test to understand the factors driving the crypto prices.

Researchers have extended the social media domain to other platforms other than Twitter by using Google trends and Reddit sentiments to predict the price movements of cryptocurrencies. Abraham et. al (2018) used Google trends together with Twitter to predict price movements of Bitcoin and Ethereum. They found no relation between Twitter sentiments and price movements. They analysed tweet volume and google trends instead using a linear model approach and found the relationship to be robust in time of high variance and non-linearity. Twitter being an unreliable indicator from previous study was corroborated by Alejandro 2021 who used both Reddit and Twitter data and found that higher sentiment is not correlated with positive returns.

Given cryptocurrency price prediction is still a relatively new field, numerous available literatures contradict one another. This project builds on previous work by exploring the predictive power of twitter sentiments using machine learning models, as well as attempting to understand causal relationships between variables using Granger Causality analysis and implementing a VAR model to capture the time series impact. Furthermore, since stock market trends and crypto prices have not been studied together till now, this study incorporates stock market data, namely Dow Jones Index prices, as one of the model's input features and investigates its association with crypto prices.

1. **Method**

* 1. **Data Collection**
     1. **Timeframe**

Before starting the data collection, it was necessary to choose a timeframe during which cryptocurrency prices fluctuated, so 1 July 2021 till 15 August 2021 was chosen because crypto prices were low in early July and then increased later (see Figure1), providing enough data points to effectively evaluate the models. Furthermore, Bitcoin and Ethereum were chosen as currencies since they have the highest market capitalization, data gathering will not be a problem as they are actively discussed coins on Twitter, and finally, the results will be easy to compare to previous work in this field.

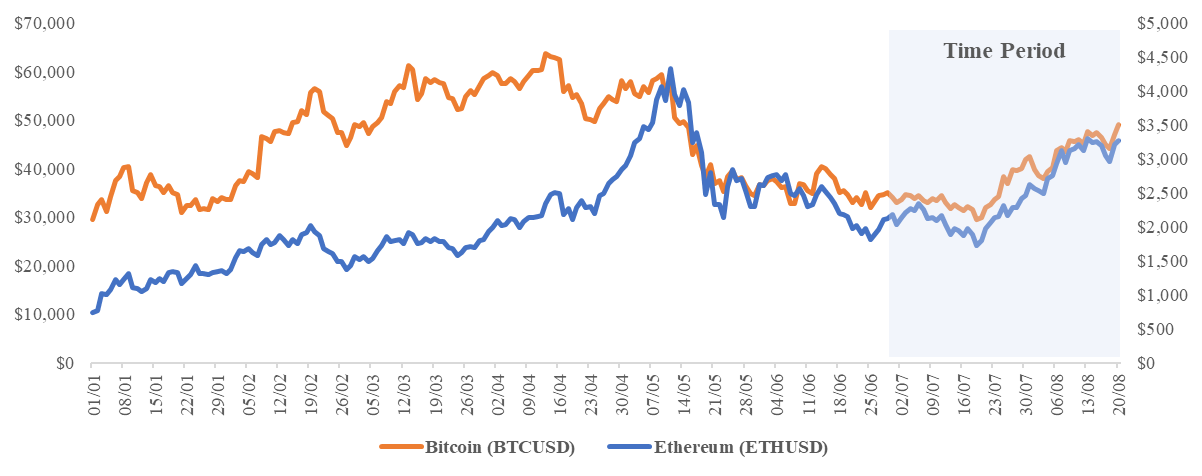
****

Figure1: Daily Bitcoin and Ethereum prices for 2021 (source: <https://www.cryptodatadownload.com/> Gemini Exchange)

* + 1. **Twitter Data**

The first step in this analysis was to collect tweets data which was downloaded using an official Twitter API. A developer account with academic research product track was created that gave access to v2 endpoint that allowed downloading Twitter user attributes, media information, likes, and retweet information based on search terms and hashtags. The maximum tweets that could be scraped per request was 500; however,10 million tweets could be retrieved in a month. The collection of crypto-specific tweets was done using # or $ the crypto coin name (for example, #BTC or #Bitcoin) as Twitter users uses these hashtags while posting tweets relevant to a cryptocurrency. In total, 16,848,867 English tweets were downloaded from 1 July 2021 to 15 August 2021 for both Bitcoin (11,589,792 tweets) and Ethereum (5,259,075 tweets) in a batch of seven days.

* + 1. **Financial Data**

Since this report explores both the hourly and daily level relationships for cryptocurrency price movements, data was downloaded from https://www.cryptodatadownload.com/. It has historical prices both at daily and hourly levels across different exchanges and data from two exchanges - Gemini and Bitstamp was considered as these are big websites and had prices in USD. Then mean of the ‘Close’ prices across these exchanges was calculated to determine the final crypto price for that time of the day. Other financial data included was - ‘Open’, ‘High’ and ‘Low’ prices and daily trading volume.

Dow Jones Index (DJI) data is acquired for stock market prices from barchart.com, which provides data for both interday and intraday trading. Because stock markets are only open for a few hours a day, unlike crypto currency, which is available for trade 24 hours a day, missing prices were replaced with the index's last close price.

* 1. **Sentiment Analysis**

The next stage was to do sentiment analysis on tweets to determine the trader's thoughts and intentions. To analyze the text and generate the sentiment score, the VADER (Valence Aware Dictionary for Sentiment Reasoning) model (Hutto, C.J. & Gilbert, E.E. (2014)) is utilized. It is a sentiment analysis tool with a lexicon-rule based designed particularly for social media content. It can handle negations ("not good"), capitalization and punctuation for emphasis, emoticons, and sentiment-laden acronyms (e.g., 'lol'), among other things. A sentiment result is a dictionary of negative, positive, and neutral scores, as well as a compound score derived by adding the valence ratings of each word in the lexicon, adjusting according to the criteria, and then normalizing to a range of -1 (most extreme negative) to +1 (most extreme positive) (most extreme positive) in case unidimensional score to be used.

Twitter data is well recognized for its unstructured and noisy nature. So, before computing the sentiment score, the tweet data was cleaned using the procedures below:

1. Remove the RT from retweeted tweets.

2. Getting rid of the @Twitter handles

3. Getting rid of the URLs that do not say anything about the emotion

4. Special characters and digits have also been deleted.

The rest of the text was left in its current state as VADER is specifically used to handle other expressions for this analysis.

Please note that retweets are not excluded from this analysis since increased activity on Twitter due to retweets indicates content dissemination, which is critical for this analysis.

* 1. **Final Dataset**

After data collation, cleaning, and score calculation at tweet level, entire data was merged to prepare model input files at the hourly and daily levels. The input file contained input feature set for time t with both twitter sentiment scores and financial data (see Table1).



Table1: Explanation of the variables in the input feature set

The dependent feature was a binary variable with values 1 and 0, with 1 representing a positive change in crypto returns between time t and t+1 and 0 representing a negative change in the returns. The selection of this variable was based on the learning from existing literature that the direction of crypto return is a good indication to sell or buy the currency and indicate profits.

The split of direction is little imbalanced for both hourly and daily level data for both currencies but both directions are well represented in the datasets.



Table2 : Distribution of price movements for Bitcoin and Ethereum

* 1. **Predictive Modeling**

To understand Twitter data's influence on crypto returns, this analysis uses machine learning algorithms to predict the directional change in cryptocurrencies. Further, it explores the causality behavior between the variables using the granger causality test and VAR model to capture the time series lag.

**3.4.1 Correlation Analysis**

Correlation is a bivariate analysis that measures the relationship between variables and the direction of that relationship. If the correlation coefficient value is between 0 and 1, the variables are positively correlated; if it is between -1 and 0, the variables are negatively correlated, with a weaker relationship around 0 and a strong relationship represented by the +- 1 coefficient. There are several techniques for calculating the correlation, including the Pearson, spearman, and Kendall methods. Pearson's coefficient is a parametric test that requires that the data is normally distributed. Spearman and Kendall's techniques are employed for non-normal series since they work on data rankings rather than actual values. (5).

The Shapiro-Wilk test (6), a standard normality test that quantifies if data is derived from a gaussian distribution, is used to determine if our data is normally distributed. According to the normality test, the majority of the variables in our data did not follow a normal distribution; therefore, the Kendall approach was chosen since it is the most robust and efficient method, and it works best with small samples.

Because activity on Twitter and previous crypto prices might impact future crypto returns, cross-correlation analysis is used to determine the correlation between the variables up to two lags to capture the lagged series relationship. This is done to understand that increasing delays will reduce the sample dataset; thus, maxlag is kept at two. Only variables with strong correlation are kept for prediction as part of feature selection based on cross-correlation coefficients, as uncorrelated variables add noise to prediction models and reduce performance.

**3.4.2 Classification Models**

Multiple ML Classifiers such as Logistic Regression, XGBoost, SVMs, Random Forest, and Neural Networks were built using the input features obtained through cross-correlation analysis to conduct the classification process. The input features were divided into two sets: one with only Twitter data and the other using a combination of both Twitter and crypto data. Then, based on out-of-sample accuracy for all models, these two input feature approach was compared to see if Twitter can forecast crypto price fluctuations.

The input data is split 70:30 for training and test data for training and evaluating the models. Furthermore, models are customized by tuning the hyperparameters called hyperparameter optimization. Every model has a range of hyperparameters that can result in better model accuracy based on specific classification problems and input data. These hyperparameters are discussed in detail for each model. However, two approaches are specifically performed for optimizing - GridSearchCV and RandomSearchCV.

GridSearchCV defines a search space and evaluates all possible values to find an optimum value, whereas RandomSearchCV creates a bounded domain of hyperparameters and randomly samples points in that space. Both techniques use cross-validation, a resampling method used on sample data to evaluate model performance and reduce model bias caused by sample selection. This analysis uses both approaches with 5-fold cross-validation for hyperparameter tuning and is implemented in Python using functions from the scikit-Learn library.

1. **Logistic Regression**

Logistic Regression is a popular classification algorithm that predicts the probability of occurrence of an event using the logistic/sigmoid function. The matrix notation of the probability function is –

where represents the matrix for coefficients for each feature and x represents the observations.

Despite its name, it is a discriminative classifier that outputs the chance of an event being categorized as 0 or 1, with threshold values that can be modified based on the cost of misclassification rate. The ROC curve was used to determine the best threshold, and two parameters were tuned, namely the regularisation value C and the penalty parameter. The model's complexity is controlled by C, while the sparsity of the solutions is controlled by different penalty options (7).

1. **XgBoost**

The following model is XGBoost (Extreme Gradient Boosting), a gradient boosting machine learning approach that implements gradient boosted decision trees. Gradient boosting uses the gradient descent approach to minimize the loss when adding new models. Boosting is an ensemble method in which models are added sequentially to enhance the accuracy of the preceding ones. XGBoost is popular as it provides results with high accuracy in both classification and regression problems. Before implementing XGBoost, four parameters were tuned [8]:

1. n\_estimators: number of trees to be built.
2. max\_depth: maximum depth of a tree. a higher value can make the tree overfit

Gamma controls the partition of a leaf node of the tree based on reduction loss. the higher value of gamma fewer the splits and the more conservative the algorithm will be

scale\_pos\_weight: parameter that controls the balance for positive and negative weights. Since the data is slightly imbalanced, few values are tuned to see if it improves the results.

To tune these parameters, GridSearchCV() which performs an exhaustive search over specified parameter values for a model.

1. **SVMs**

Support Vector Machines (SVMs) are a supervised learning technique that outperforms other classifiers in terms of accuracy. A non-probabilistic classifier works well in high-dimensional spaces and uses a regularisation term to prevent overfitting. It generates an optimal hyperplane, called maximum marginal hyperplane, to divide the dataset and classify the points. It can also convert non-linear issues to linearly separable instances using various kernel methods, such as polynomial or radial basis functions.

This analysis uses all three kernel methods – linear, radial, and polynomial to build SVM Models and optimize hyperparameters accordingly. Python’s scikit.learn library is used to run the models using SVC(), SDGClassifier() and LinearSVC() functions. Three parameters are tuned-

1. loss: the loss function to be used for classification
2. alpha: a constant that multiplies by the regularisation term
3. penalty: the penalty or regularisation term
4. **Random Forest**

Support Vector Machines (SVMs) are a supervised learning technique that outperforms other classifiers in terms of accuracy. A non-probabilistic classifier works well in high-dimensional spaces and uses a regularisation term to prevent overfitting. It generates an optimal hyperplane, called maximum marginal hyperplane, to divide the dataset and classify the points. It can also convert non-linear issues to linearly separable instances using various kernel methods, such as polynomial or radial basis functions.

This analysis uses all three kernel methods – linear, radial, and polynomial to build SVM Models and optimize hyperparameters accordingly. Python’s scikit.learn library is used to run the models using SVC(), SDGClassifier() and LinearSVC() functions. Three parameters are tuned-

1. loss: the loss function to be used for classification
2. alpha: a constant that multiplies by the regularisation term
3. penalty: the penalty or regularisation term
4. **Neural Networks**

The neural Networks method is superior at capturing complex relationships between variables by searching through different models in its black box to find patterns. In its black box, there can be one or more linear or non-linear hidden layers.

A multi-layer perceptron model, or MLP, is created in sklearn to depict a neural network made up of perceptrons. The most critical parameters are activation (the type of activation function), alpha (regularisation constant), hidden layer sizes (which represent the hidden layers), learning rate (which controls how quickly the network tunes weights depending on training feedback), and max iter (the number of iterations of training).

* 1. **VAR Modeling**

In the next step, granger causality is tested to determine various factors driving price changes. It is important to note that granger causality determines if X has predictive power for Y and relates to the critical fact in statistics that correlation does not mean causation. To run the causality test, the series must be stationary, so both ADF and KPSS tests were run to check trends and seasonality in the series for each variable. Further granger causality test was applied to the stationary series. Finally, the VAR model was built to capture the price change due to time lagged Twitter and financial data.

**3.6 Evaluation**

To evaluate the robustness of each model, confusion matrix (ref Table3) was drawn and various measures were used which are defined as –

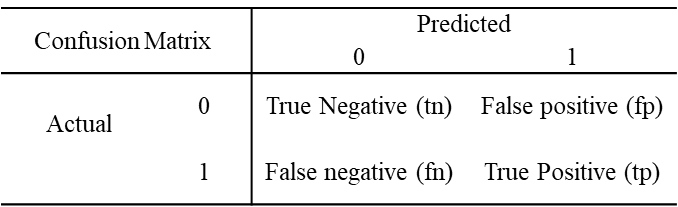


Table3: Confusion Matrix

Accuracy is an important metric as it measures the test records predicted correctly, precision is the ratio of relevant samples correctly predicted for that class, recall is the ratio of relevant class predicted correctly out of actual relevant class samples, and f1 score is when both the classes are equally important. Here, for our analysis, since we are interested in both classes as it indicated when to sell or buy the cryptocurrency, we will look at the model's overall accuracy and f1 score.

1. **Results and Analysis**

**4.1 Exploratory Analysis**

The results of the sentiment scores can be viewed in Figures 2(a), 2(b), 2(c), and 2(d). The average compound score (avg\_cmp\_scr) appears to be positively skewed for both cryptocurrencies, with an average score of 0.22 for both hourly and daily data. This is consistent with all of the existing literature reviewed earlier, which shows that users on social media prefer to use reassuring language. When looking at hourly average scores (ref Figure3(a), 3(b), 3(c) & 3(d)), they appear high in the early hours of the day, but they gradually decline as the day progresses. However, daily tweet volume seems to have increased from mid-July onwards as crypto prices went up. Both Bitcoin and Ethereum are following a similar pattern (refer to EDA Notebooks for interactive charts)

Please note that during EDA, Bitcoin tweet volume appeared to be relatively low between 1 July and 15 July, which could be a data scraping issue for that period, but data downloading is limited by Twitter, so this could not be confirmed, and the entire analysis was carried out assuming Twitter volume was genuinely low.

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated | Chart, histogram  Description automatically generated |
| Figure2(a): Distribution of daily average sentiment score for Bitcoin tweets | Figure(b): Distribution of daily average sentiment score for Ethereum tweets |
| Chart, histogram  Description automatically generated | Chart, histogram  Description automatically generated |
| Figure2(c): Distribution of hourly average sentiment score for Bitcoin tweets | Figure2(d): Distribution of hourly average sentiment score for Ethereum tweets |
| Chart, line chart  Description automatically generated | Shape  Description automatically generated with medium confidence |
| Figure3(a): Daily mean sentiment metrics and Bitcoin prices for Bitcoin-related tweets between 1 July – 15 August 2021 | Figure3(b): Hourly mean sentiment metrics and Bitcoin prices for Bitcoin-related tweets between 1 July – 15 August 2021 |
| Chart, line chart  Description automatically generated | Chart, line chart, histogram  Description automatically generated |
| Figure3(c): Daily mean sentiment metrics and Ethereum prices for Ethereum-related tweets between 1 July – 15 August 2021 | Figure3(d): Hourly mean sentiment metrics and Bitcoin prices for Etheruem-related tweets between 1 July – 15 August 2021 |

**4.2 Cross Correlation results**

The Table4 displays the findings of a cross-correlation analysis conducted better to understand the relationship between lagged variables and crypto values. For daily level data, the correlation score varies from -0.3 to 0.3, which is relatively low and indicates a poor correlation between the variables and crypto prices. The results were the worst at the hourly level as the correlation coefficient ranged from -0.06 to 0.06, indicating no meaningful association. This pattern was observed in both bitcoin and Ethereum.

The selection of essential variables for daily level prediction for bitcoin was made with cut-offs of [0.15, -0.10] and [0.10, -0.10] for Ethereum due to low correlation coefficients. However, no lagged variables were employed for hourly level prediction, and models were created based on variable knowledge.



Table4: List of variables selected after cross-correlation results for Bitcoin and Ethereum daily classification models

**4.3 Classification Models results**

The tables below show the results of each of our models built using two different input features for both the cryptocurrencies at hourly and daily levels. The best models for each subset are highlighted in yellow.

As seen in the table5(a), model accuracy is low when Twitter data is used for the majority of the models, but accuracy improves significantly when crypto data is utilized, such as open prices, the past volume traded and closure prices, and DJI

data (ref Table5(b)). SVM methods outperform simple logistic regression models for daily bitcoin predictions, with an accuracy of 81 percent and precision of 83 percent. Random Forest and XgBoost performed the poorest; however, when twitter data was utilized alone, Random Forest performed the best.



Table5(a): Classification model results for daily bitcoin price movements using twitter data



Table5(b): Classification model results for daily bitcoin price movements using twitter and price data

Model accuracies with Twitter data were relatively poor at the hourly level (ref Table5(c)), but they greatly improved when financial data was added (ref Table5(d)), with the Random Forest Classifier being the top-performing model with an accuracy of 0.85.



Table5(c): Classification model results for hourly bitcoin price movements using twitter data



Table5(d): Classification model results for hourly bitcoin price movements using twitter and financial data

Using Twitter data at the daily level, the highest performing models for Ethereum (ref Table6(a)) were SVM and Random Forest; however, SVMSDGClassifier with radial kernel outperformed other models when financial data was used (ref Table6(b)). Few other models performed better, but their recall rates for negative price movement (indicated by \*\* in the table) were zero; thus, they were ignored.



Table6(a): Classification model results for daily Ethereum price movements using twitter data



Table6(b): Classification model results for daily Ethereum price movements using twitter and financial data

At the hourly level (ref Table 6(c)), XgBoost had better performance with an accuracy of 0.58 when twitter data was used. However, model accuracies improved when financial data was added to the input feature list. SVM models achieved the model accuracy of 84% with linear kernel (ref Table6(d)).



Table6(c): Classification model results for hourly Ethereum price movements using twitter data



Table6(d): Classification model results for hourly Ethereum price movements using twitter and financial data

4.4 Causality Analysis Results

The Appendix contains the findings of the Granger causality test. Twitter volume (p<0.05) granger-caused Bitcoin prices data at daily intervals, indicating that twitter volume has a strong predictive power on Bitcoin. Sentiment scores, on the other hand, had no significant effect on the crypto data, which is consistent with our previous findings. The DJI prices have a big effect on Bitcoin prices and vice versa, which is yet to be investigated because both markets trade separately. However, this result can be further investigated to see if there is any correlation between stock and digital currency markets. Bitcoin prices have a considerable impact on twitter activity on an hourly basis, yet the Bitcoin close price is unaffected by twitter sentiments or activity. Interestingly, there is a two-way causation between Bitcoin prices and the DJI. Also, Twitter activity such as likes, retweets, and user follower count impacts the Bitcoin trading volume.

When VAR model was implemented to capture these causality relationships, it performed with an accuracy of 77% on daily dataset but since twitter sentiments were having no predictive power at hourly level, model performed poorly with an accuracy of 55%.

For Ethereum, both tweet sentiment and volume showed predictive power for ‘Close’ price at daily levels and inverse relation existed between Ethereum prices and tweet volume. At hourly level, polarity and DJI prices granger caused the Ethereum prices and Ethereum prices showed significant effect on tweet volume. Despite the fact that Ethereum has many meaningful correlations, VAR model performed poorly with just 55% accuracy at both daily and hourly level data.

Discussion

The bitcoin and ethereum coins show that the crypto price movements can be predicted using Twitter and historical price data. Though twitter data alone could not predict the directional change in prices, adding historical market data improved the model performance significantly at both hourly and daily levels. Bitcoin's best model was an SVM using a linear classifier with an accuracy of .81 and precision of .83 at the daily level and a random forest classifier with 0.85 accuracies at an hourly level. Ethereum's top-performing model was SVM with radial Classifier with 0.77 accuracies at the daily level and SVM with linear Classifier with 0.84 accuracies at the daily level.

This prediction was limited to the direction change in prices, but it proves that Twitter data drives price movements.

**References**

1. <https://www.investopedia.com/tech/how-does-bitcoin-mining-work/>
2. <https://www.dummies.com/personal-finance/what-is-cryptocurrency/>
3. <https://uk.tradingview.com/markets/cryptocurrencies/global-charts/>
4. <https://www.bankofengland.co.uk/knowledgebank/what-are-cryptocurrencies>
5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107969/>
6. <https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/>
7. <https://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic_l1_l2_sparsity.html>
8. <https://www.datacamp.com/community/tutorials/xgboost-in-python>
9. <https://www.investopedia.com/tech/most-important-cryptocurrencies-other-than-bitcoin/#:~:text=One%20reason%20for%20this%20is,communities%20of%20backers%20and%20investors>.
10. [1212.1107.pdf (arxiv.org)](https://arxiv.org/pdf/1212.1107.pdf)

**Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.**

**Appendix**



Table: Granger Causality results for daily Bitcoin data



Table: Granger Causality results for hourly Bitcoin data



Table: Granger Causality results for daily Ethereum data



Table: Granger Causality results for hourly Ethereum data