# A Predictive Model for the Global Cryptocurrency Market

## A Holistic Approach to Predicting Cryptocurrency Prices

Minul Wimalagunaratne
Informatics Institute of Technology
Colombo, Sri Lanka
e-mail: minulwimala8@gmail.com

Guhanathan Poravi
Informatics Institute of Technology
Colombo, Sri Lanaka
e-mail:guhanathan.p@iit.ac.lk

***Abstract*—The realm of cryptocurrency has grown exponentially over the past decade, with the most rapid advances seen in the past few years as more and more parties around the world recognize the value of holding digital assets online. Statistics from Twitter support this statement where, approximately 1,500 Tweets about Bitcoin alone is recorded per hour. Consequently, many people are beginning to become more aware and accepting of the nature of digital currencies, and traders in particular seek to know how they can make profitable crypto-coin trades and investments. Although a number of research projects have been undertaken to develop systems that can effectively predict price movements in the cryptocurrency market, they display significant efficiency gaps, which this paper further explores. The authors then attempt to learn from past studies and construct a more holistic approach to a predictive price model for the cryptocurrency market. This focuses on assessing key factors that affect the volatility of the market – public perception, trading data, historic price data, and the interdependencies between Bitcoin and Altcoins - and how they can be best utilized from a technological aspect by applying sentiment analysis and machine learning techniques, to increase the efficiency of the process.**

***Keywords-machine learning, predictive models, sentiment analysis.***

## I. INTRODUCTION

In today's highly commercialized global landscape, the demand for a more accessible and transparent medium of currency has grown rapidly [2]. As the world moves forward with new advancements in technology, so too has the realm of monetary exchange evolved with the concept of digital currency (or cryptocurrencies) following the release of the first virtual currency in 2009, Bitcoin. The concept of cryptocurrency focuses on having faster and more secure monetary transactions online. The technology used to enable this is called the Blockchain, where there is no central party to verify transactions and instead the whole system is decentralized, hence making it much more secure. This proves significantly more advantageous than the current system used to verify and conduct fiat currency transactions, where the ability to easily abuse their use in transactions has led to some of the biggest financial scandals of the century. One such example is the Libor Scandal of 2016 where banks manipulated interest rates for bigger profit margins.

When looking at the growth and success of digital currencies, data from the world's first cryptocurrency survey draws four main conclusions. First, cryptocurrency market capitalization has increased radically as a result of the growing strength and acceptance of crypto-coins as a digital asset; the market cap lies at USD 592 billion with the Bitcoin market price at over USD 17,000 [6]. Second, since 2009 Bitcoin has remained the market leader for cypto-coins, with a dominance of over 55% in the cryptocurrency market (which consists of a total of 1360 digital coins) (as of December 2017); this attests to its strong influence over the behavior of Altcoins. Third, out of all interested parties (traders, miners and investors), traders are the largest stakeholder group who engage the most with the cryptocurrency market to profit from buying and selling digital assets in online cryptocurrency exchanges. And fourth, the extreme volatility of the market makes it risky for traders to hold or trade their assets profitably. All these observations illustrate the dynamism of the market and highlight the need for a cryptocurrency price prediction system.

## II. LIMITATIONS OF CURRENT STUDIES

Attempts thus far to construct such a model has faced significant limitations. The biggest limitation is the fact that current research has been heavily restricted to a few more popular crypto-coins in the market - Bitcoin, the market leader, closely followed by Ethereum, Dash, Monero, Ripple and Litecoin. But as of today, 1360 other Altcoins exist in the market (as of December 2017). At the same time, the cryptocurrency market is highly unstable and experiences periods of extreme volatility which often makes it difficult to predict behavioral patterns. Past studies only take into consideration one or two market variables when attempting to predict the price of crypto-coins, failing to account for all factors that may affect the market. These two observations result in predictive models with limited accuracy, and the problem of limited access to information in the global cryptocurrency market continues to persist.

The predictive model suggested by the authors aims to be more holistic in nature. It takes into consideration multiple factors affecting the market, and applies a range of technological methodologies, tools and techniques, in order to provide an accurate prediction so that users will be able to better benefit through investing, trading or mining cryptocurrencies more effectively.

978-1-5386-6539-8/18/$31.00 ©2018 IEEE
DOI 10.1109/ISMS.2018.00024

## III. Factors That Can Be Used Predict The Price of Cryptocurrencies

### A. Public Perception

As discussed in the previous sections, people are beginning to understand and use cryptocurrencies more frequently. The amount of chatter online has also increased significantly over the past few years, giving rise to the increased popularity of the cryptocurrency market [11]. Subsequently, an interesting relationship between Bitcoin prices and the amount of news online has been identified in research conducted by D'Alfonso et al. (2016). The study's findings made it clear that the increased amount of talk online had a considerable effect on the price; in this case, positive talk about Bitcoin lead to an increase in its price [2]. To further confirm this point, Google trends show that for Etherum, Bitcoin's biggest competitor, there is an extremely high co-relationship between its increase in price and the amount of google searches online [8].

### B. Trading Data

With the growth of cryptocurrencies and the general public starting to accept its benefits, many online exchanges have been set up so that people all over the world can invest in cryptocurrencies or trade their digital assets in order to gain a profit. With time the number of users on exchanges has increased tremendously. With this increase, stakeholders show a strong interest in trading information as this information can be used to make informed decisions about trades or investments. Due to the demand for information, popular exchanges such as Bittrex and Poloniex now expose data to their users through API's. This is a clear indicator that direct stakeholders are interested in futuristic information about the behavior of digital currencies, and hence adds to the need for a price prediction system.

### C. Historic Price Data

In the world of stock markets, people use graphs of historic price data to identify behavioral trends and patterns of currencies in order to try and predict their future price. Research carried out by Patel et al. (2015) study the techniques used to identify price patterns to make price predictions of a few selected companies on the Indian Stock Market [13].

In a similar manner, patterns have also been identified on cryptocurrency price graphs which suggests that this could also be a valuable attribute to consider when predicting the price of cryptocurrencies. The historic prices of cryptocurrencies can be obtained through online exchanges as explained in the previous section

### D. Interdependencies between Bitcoin and Altcoins

Research carried out by Ciaian et al. (2016) attempt to identify factors that can create an interdependency between Bitcoin and Altcoins. As a result of Bitcoin's dominance (55%) in the market, it is being traded off to buy Altcoins. To further elaborate on this point, 68% of all Altcoins are purchased through Bitcoin while only 14% of Altcoins are bought directly through the US dollar. Another factor that is considered is the similar price developments of Bitcoin on Altcoins. This is because the vast majority of Altcoins are largely clones of Bitcoin with minor changes in parameter values (different block times, currency supplies or issuance schemes). A visual inspection of Bitcoin and Altcoin price indices show a similar price drop in both crypto-coins in the same period of time; this trend is again observed when Altcoins follow an increase in Bitcoin prices. The research states that the similar behavioral patterns could be as a result of the fact that Altcoins are following the price of the market leader, Bitcoin [5].

## IV. Areas of Sentiment Analysis Applicability

Sentiment analysis, also known as "opinion mining", is the process of digitally understanding the polarity of text. Polarity in this context refers to the ability to understand if the specific text expresses a positive or negative opinion. Determining this fact allows a system to process what a particular individual's sentiment is. Therefore, this technique can be used to understand what kind of opinion people have with regard to a certain topic, given that the data provided to be analyzed is of the same topic [12].

The world we live in today is highly globalized, and people of all ages have access to social media which they use to keep in touch with others, as well as to express their opinions on specific topics of interest. The increase in use of social media has brought in many social media networks such as Facebook, the largest social media network where there are over 1,200 million active users per day, and Twitter, which produces over 500 million tweets per day [7]. These statistics prove that there is a great deal of information which can be used to in order understand what exactly people feel about a particular topic. Taking a closer look at the statistics from Twitter, around 1500 tweets about Bitcoin has been recorded per hour [1]. The massive number of posts/tweets, and the number of users they come from, makes social media a great data source from which to gather primary information about Bitcoin and Altcoins, which the proposed system can then use to better understand through the application of sentiment analysis.

## V. Areas Of Machine Learning Applicability

Machine learning is a method of getting systems to act in a certain manner without explicitly programming it to do so. This technique is used in areas like self-driving cars, speech recognition and also prediction systems like stock prediction systems. Having said that, machine learning will have to be applied in multiple areas and using a variety of techniques in order for this approach to be a success. This section discusses and justifies how machine learning techniques will be applied to make the predictive model a success.

### A. Machine Learning Applied to Trading Information

Studies show that trading information is a reliable factor that can be used for predictions, especially when applying machine learning as there are existing relationships already found

within trading information, such as between the opening price, closing price and trading volumes. Machine learning techniques pick up on these relationships and base their predictions out of these factors. Greaves and Au (2015) used trading information from online exchanges to predict the price of Bitcoin. Initially the research was based on making predictions using information from the blockchain itself. Having no success in this approach the study continues by focusing on trading information. The study elaborates on an interesting relationship where the price increases when Bitcoin is sold for more than it is bought at. The justification for this affect is that the price of Bitcoin increases due to an increase in demand for the coin [9].

### *B. Machine Learning Applied to Historic Price Data*

By using historic price data, machine learning algorithms can attempt to identify price patterns. This would help a predictive system anticipate prices more accurately. A similar technique has been used by Patel et al. (2015) to predict price indexes for four companies listed on the Indian Stock Market. In order to gain outcomes with high accuracy, the system they designed used ten technical parameters in the data set which was used to train and predict prices. To further improve accuracy multiple machine learning algorithms were employed. This study uses a trend deterministic data preparation layer which manipulates continuous data to discrete values (+1 or -1), according to the results show an increase in accuracy. Results of the experiment show that the algorithm Naïve Bayes performed the best with an accuracy of 90.19%. A study carried out by Rowland (2014) also used this approach to predict the price of Bitcoin only. This system used an ensemble voting system together with two machine learning algorithms which ran synchronously in order to maximize on the accuracy of the predictions. The ensemble learning approach generates multiple classifiers which are later combined. This approach tries to optimize accuracy by eliminating errors generated by a single classifier [14].

### *C. Machine Learning Applied to the Data Results of Sentiment Analysis*

Determining the sentiment of user posts alone is not enough to come up with an actual price prediction. For this to be done effectively, machine learning techniques will have to be used. First, a count of negative and positive tweets from the relevant time period will be fetched from the database. The number of negative and positive tweets acts as the data set which will allow the machine learning program to understand the opinion of the users. Using this information, the system can now be trained to predict the price change which will then allow us to make a prediction on the future price of cryptocurrencies.

## VI. MACHINE LEARNING TECHNIQUES

Machine learning technologies are being constantly developed over time to bring many new algorithms to life. While some algorithms can be used to produce results with high accurate levels when addressing one type of problem, they can also produce a low accuracy when attempting to address a problem of a different nature. Hence it is important to first assess the properties, strengths and weaknesses of different available algorithms in order to understand and determine which algorithm forms the best fit for this system.

### *A. Artificial Neural Networks*

Artificial Neural Network (ANN) is one of the most commonly used machine learning techniques. Its structure is based on two layers called the input and output layers. Apart from these layers, an unknown number of nodes can exist in a layer known as the hidden layer. The hidden layer enables the algorithm to have a strong learning capability. One of the biggest advantages this algorithm holds is its capability to have a finite number of layers and yet still be able to accurately predict values for any continuous function. This property is referred to as the universal approximator [4]. Due to this property, the algorithm also faces an issue know as over-fitting. This is when the algorithm detects a considerable amount of non-existent relationships which could lead to a major drop in accuracy [3].

### *B. Support Vector Machine*

Support Vector Machine (SVM) is a supervised learning model. The algorithm defines classes by separating data using clear boundaries while considering errors caused during the process [9]. In certain real-world situations where boundaries need to be defined in a more complex manner, an implementation called the kernel implementation, also known as the non-linear kernel, is used [14]. This algorithm has the ability to automatically find optimum parameters for high levels of accuracy [10]. However, it must be noted that this algorithm might use a very high amount of computational power, as well as time, to produce results.

### *C. Random Forest*

This algorithm is based on the logic of decision trees. The algorithm generates ‘n’ number of trees, thereafter each node selects three random features. When the data is analyzed multiple outputs are generated, and the final output is chosen by selecting the output that went through the most number of trees [13]. This algorithm is fairly easy to implement as it does not require too much fine-tuning in order to reach high limits of accuracy [15]. However, this technique uses up a lot of memory depending on the classification at hand.

### *D. Naïve Bayes*

Naïve Bayes is a simple and efficient algorithm. It is based on the assumption that any one of the attribute identified is completely unrelated to each other attribute; this gives the algorithm the ability to ignore irrelevant attributes. Although this assumption does not hold true in all real-world situations, it has been proven to reach high accuracy levels in certain real-world scenarios [16]. As each feature is defined into the prediction model there are no issues in scaling it, which is extremely advantageous. However, if a new feature needs to

be considered then a new prediction model needs to be built from scratch [14].

## VII. The Proposed System

In the previous sections of this paper the methods and techniques of the proposed system have been discussed. The proposed system aims to capture the main factors that affect the price of cryptocurrencies, and employ various machine learning techniques on the data, in order to make an accurate prediction of the price of cryptocurrencies. A high-level diagram of the proposed system is depicted below in Figure 1.

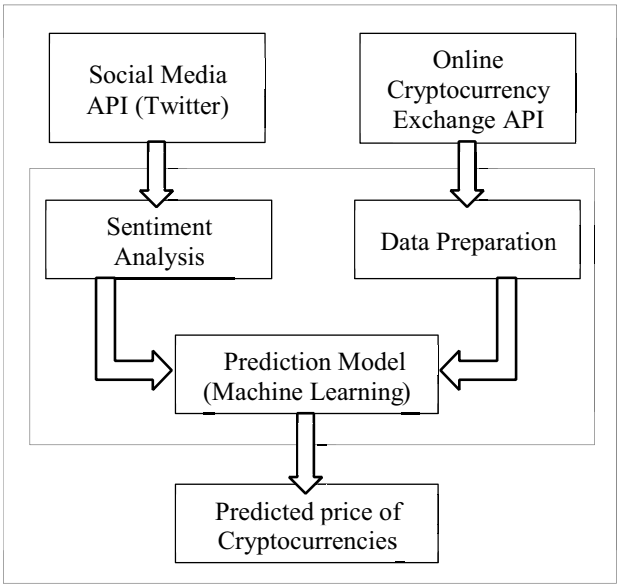


Figure 1: High Level Diagram of the Proposed System.

### A. Data Capture and Data Preparation

One of the factors that affect the price of cryptocurrencies is public perception of cryptocurrency stakeholders on the health of the market. This factor is captured through social media API's, Twitter to be precise. As Twitter's API exposes all tweets based on a certain topic, only the needed tweets can be filtered out and gathered. Thereafter, sentiment analysis is performed in order understand the public perception of Twitter users who identify with the concept of cryptocurrency. Unfortunately, this technique cannot be applied to all cryptocurrencies as most online chatter is limited to a few more popular coins like Bitcoin, Etherum and Litecoin. Due to this reason, there is a limitation where sufficient data on other cryptocurrencies cannot be gathered to make an accurate enough prediction. In order to overcome this problem, sentiment analysis will be applied only to tweets related to Bitcoin. This is further elaborated below in Section B 'Prediction of Bitcoin Prices'. Once the sentiment of a tweet is determined, the tweet will be saved in the system's database along with its sentiment so that this data can be easily reused for further research purposes in the future.

The data needed to capture the factors of trading information, historic prices and information related to the price interdependency between Bitcoin and Altcoins can be obtained through online exchanges. When the data is collected, it will be adjusted so that the system can use this information to train machine learning algorithms which will then predict prices into the future. The adjusted data set will also be saved in the system's database as this information can be used to train the system again in future to adapt to new trends or patterns. The maintenance of data will be a scheduled process so that the system can be as up to date as possible.

### B. Prediction of Bitcoin Prices

In order to predict the price of Bitcoin three factors are considered; the aspect of public perception, as discussed previously, will be captured through Twitter API's and the database will contain the tweets and their respective sentiments. The key information that is needed will be a count of negative and positive posts mapped to its time sequence. The other two factors that will be considered for this prediction is the trading information and the historic price of crypto-coins. The dataset will contain values such as the percentage change and trading volume, which can be derived from the trading information when initially saved in the database. The dataset should also have data from a big enough time period so that the machine learning algorithms will be able to identify the price patterns. Once the data set consists of all the necessary data, a selected machine learning technique can be applied in order to obtain an accurate prediction.

### C. Prediction of Altcoin Prices

One of the biggest problems faced in past studies is the limitation of data online about newer or less popular cryptocurrencies. Although this factor cannot be directly captured here, information based on the prediction of Bitcoin, such as the percentage change of the new price prediction, can be used to weigh in on the price prediction of Altcoins. Considering information about the change in price of Bitcoin when predicting the prices of Altcoins is beneficial as there is an interdependency between the price fluctuations of Bitcoin and that of Altcoins. Therefore, including this data will most likely create a new relationship which will in turn help increase the accuracy of the prediction. Consequently, the interdependency between Bitcoin and Altcoins can then be used to predict the prices of Altcoins and therefore acts as a substitute for the factor of public perception which cannot be used for altcoins.

### D. Initial Evaluation and Testing

To make sure that the captured tweets are useful and that they have an effect on the market, the polarity and the actual percentage change of the price change in Bitcoin was compared repetitively while applying various filters. Figure 2 represents a comparison of the polarity and the percentage

change of price in Bitcoin for period of time with high volatility on a daily basis.

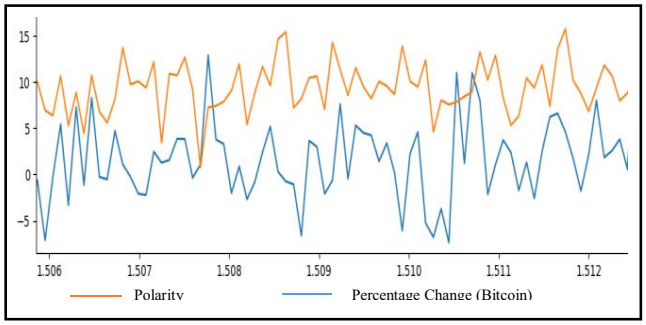


Figure 2: Comparison between polarity and percentage change in Bitcoin

When evaluating the above figure, it is clear that in some instances there is a clear relationship whereas in other instances this relationship isn't as clearly visible. Observations also show there is a delayed reaction to a spike or drop in polarity, which suggests that in certain periods the market takes time to respond to the online chatter. As this system is employing a neural network, its learning capability helps it to reduce the impact on accuracy due to these factors. A fact that should be remembered is that the final prediction of the system is based on the frequency of tweets and trading information in addition to the feature discussed above.

Three main coins have been used in order to test the accuracy of the system. The data from predictions made on the past 3 months were considered for each coin. The table below summarizes the reasons as to why these specific coins were used to testing purposes, as well as their relevant accuracy levels.

TABLE 1: REASONS AND ACCURACY OF CRYPTOCURRENCY USED FOR TESTING

| Cryptocurrency | Reason | Accuracy |
|---|---|---|
| Bitcoin | The most popular cryptocurrency. This coin is needed in order for other coin predictions to happen. | 85% |
| Ethereum | An altcoin which is comparatively popular and Bitcoins biggest competitor. Used to test how the system will handle altcoins. | 93.33% |
| Bitcoin Cash | A fairly new cryptocurrency. Used to monitor how the system will handle coins with small data sets | 70% |

Results show that that predictions of Ethereum has the highest accuracy which can be a result of including the predicted percentage price change of Bitcoin as one of the factors in the predictive model. Bitcoin also has a comparatively high accuracy. Overall, the predictive model did perform better in time periods of less volatile market fluctuations. Bitcoin cash, on the other hand, displayed the lowest accuracy level of 70%, which is reflective of the limited size of the dataset.

## VIII. CONCLUSION

This study has identified and discussed how different market factors can be used in an attempt to make accurate predictions on cryptocurrency prices. The techniques and technologies that will allow these factors to be captured and manipulated in order to make such predictions has also been explored. This solution aims to address the behavior of, and thus capture the value of, all cryptocurrencies in the market instead of focusing solely on a few of the more popular digital coins, while extracting from and taking into account information from the wide amount of data available.

## ACKNOWLEDGMENT

My heartfelt thanks goes out to my supervisor and mentor Mr. Guhanathan Poravi for all the constant guidance provided. I would also like to thank my family for the endless motivation and invaluable support provided.

## REFERENCES

[1] 30 bitcoin hashtags popular on Twitter | RiteTag: Find the best hashtags: https://ritetag.com/best-hashtags-for/bitcoin. Accessed: 2017-12-20.

[2] Alexander D'Alfonso, Peter Langer, Z.V. 2016. The Future of Cryptocurrency. International Security. 7, 5 (2016), 7–45.

[3] Anantwar, S.G. and Shelke, R.R. 2012. Simplified Approach of ANN : Strengths and Weakness. International Journal of Engineering and Innovative Technology (IJEIT). 1, 4 (2012), 73– 77.

[4] Byvatov, E. et al. 2003. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. Journal of Chemical Information and Modeling. 43, 6 (2003), 1882–1889.

DOI:https://doi.org/10.1021/ci0341161.

[5] Ciaian, P. et al. 2016. Virtual Relationships: Short- and Long-run Evidence from BitCoin and Altcoin Markets. (2016).

[6] CryptoCurrency Market Capitalizations: https://coinmarketcap.com/. Accessed: 2017-07-22.

[7] Definitive portal for social media statistics globally | Socialbakers: https://www.socialbakers.com/statistics/. Accessed: 2017-08-28.

[8] Ethereum - Explore - Google Trends: https://trends.google.com/trends/explore?q=Ethereum. Accessed: 2017-12-17.

[9] Greaves, A. and Au, B. 2015. Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin. (2015).

[10] Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 1398, (1998), 137–142. DOI:https://doi.org/10.1007/s13928716.

[11] Kim, Y. Bin et al. 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PLoS ONE. 11, 8 (2016), 1–17.

DOI:https://doi.org/10.1371/journal.pone.0161197.

[12] Pang, B. and Lee, L. 2008. Opinion Mining and Sentiment Analysis.

[13] Patel, J. et al. 2015. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications. 42, 1 (2015), 259–268.

DOI:https://doi.org/10.1016/j.eswa.2014.07.040.

[14] Rowlands, J. 2014. Bitcoin Algorithmic Trading Bot. April (2014).

[15] Statnikov, A. et al. 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer

classification. BMC Bioinformatics. 9, 1 (2008), 319. DOI:https://doi.org/10.1186/1471-2105-9-319.

[16] Yang, Y. and Webb, G.I. 2001. Proportional k-Interval Discretization for Naive-Bayes Classifiers. (2001), 564–575.