Bitcoin Currency Fluctuation

Marius Kinderis, Marija Bezbradica, Martin Crane

School of Computing, Dublin City University, Ireland
marius.kinderis3@mail.dcu.ie, {marija.bezbradica,martin.crane}@dcu.ie

Abstract. Predicting currency prices remains a difficult endeavor. Investors are Continually seeking new ways to extract meaningful information about currency price direction. Recently, cryptocurrencies have attracted a lot of attention due to their unique way of transfer of value as well as its price behaviour differing from the known stock market price fluctuation. A method proposed in this project involves using data mining techniques: mining text documents such as news articles and tweets to learn the relationship between information contained in such items and cryptocurrency price direction. The Long Short-Term Memory Recurrent Neural Network assists in creating a hybrid model which consists of sentiment analysis techniques, as well as a predictive machine learning model. The success of the model was evaluated within the context of predicting the direction of Bitcoin price movements. Our findings have revealed that our system yields more accurate and real-time predictions of Bitcoin price fluctuations compared to other existent models in the market.

1 Introduction

There are more than 900 cryptocurrencies currently available to invest in online; this number is consistently growing [6]. Of these cryptocurrencies, undoubtedly the most popular is Bitcoin and was also the first cryptocurrency in the market [31].

Several techniques have been used to give investors an advantage in predicting Bitcoin price direction. The strategies range from Statistical [19] and econometric [17] approaches to using machine learning to extract nonlinear relationships in the data [34]. In addition to technical analysis, traders gain necessary information about the market from drawing information from peers and news articles, which are often influenced by human emotion: whether market participants are feeling optimistic or pessimistic about the future state of the economy or a particular currency has an impact on that currencys price [24].

This project aims to study the impact of human emotions in the price movements of the cryptocurrency, particularly Bitcoin, by analysing the effect of sentiment contained in Twitter posts (tweets) and news articles. This is done by implementing data mining techniques to collate tweets and scrape news articles relating to Bitcoin. Another objective of this project is to build a diverse trading model which gives traders an extra tool for predicting price direction using Natural Language Processing (NLP) techniques in addition to technical analysis in the form of a Long Short-Term Recurrent Neural Network model (LSTM RNN).

This paper tries to understand the human influences behind Bitcoin's popularity and uncover the variables that influence its fluctuation from the financial perspective. The main research question we address here is to what extent Bitcoin currency fluctuation can be predicted.

To standardize our main approach to the problem, we followed the example of other authors [30] who used Cross Industry Standard Process for Data Mining (CRISP-DM) model; this serves as a baseline for our project implementation. CRISP-DM has been implemented elsewhere [30] in financial applications. The methodologies we used here are divided into four main parts: (i) exploring the factors that influence Bitcoin's price fluctuation from the financial perspective, (ii) data collection, (iii) predictive modeling and (iv) system deployment which includes keeping track of the changes occurring in the other related fields.

2 Literature Review

2.1 Creation of Bitcoin

In October 2008, Satoshi Nakamoto published the first paper [31] on Bitcoin outlining its properties as a decentralized payment system. Shortly after, he published the first Bitcoin open-source software [1] launching the network and the first units of the Bitcoin. Since then, a growing community that actively use Bitcoin has formed. An increase in users' interest has also been shown, as evidenced by over 14 millions wallets registered worldwide [32]. Statistics on the growth of bitcoin and similar cryptocurrencies can be found in [3].

2.2 Assessment/Commentaries

2.2.1 Technological Assessment In contrast to other financial systems that require a third party (i.e. banks) to validate transactions, Bitcoin is based on blockchain technology [18] and is designed to make the transfer of value easier based on its peer-to-peer system. Further, the affordability of Bitcoin transactions (around 1$ per large transaction [2]), makes the system very attractive to consumers. However, just like all other currencies, Bitcoin units can be stolen, lost or confiscated. Hence, the risks associated with using Bitcoin units should not be underestimated by casual users.

The Blockchain technology operates as follows: payments on the users' network [18] are done by the chronological transactions recorded in a public ledger, called the Blockchain. It is important to understand that the whole technology relies mainly on cryptography, digital signatures and hashes to encode the transaction (the technical details will not be outlined here but may be found in sources such as [18]). A recent study [33] has provided us with insights on the current improvements of Blockchain technology through designing a map of raised and solved issues. This technology, however, has yet to tackle some issues[1]. Its security aspects have also been found to cause problems[2]

These issues will not be tackled in this paper, but are worth noting as they must be addressed for the success and trustworthiness of Bitcoin.

2.2.2 Financial Assessment In [23] the authors explored the financial asset capabilities in hedging using generalized autoregressive conditional heteroscedasticity (GARCH) models highlighting the fact that Bitcoin reacts fast to sentiment. They also found that its status in the market is in between a commodity and a currency as it combines some of the properties [23] of both. The GARCH method has been found to be useful for portfolio management and risk analysis. This allows an exploration of the relationships between Bitcoin and other commodities such as Gold, Copper, Cocoa etc.

Several attempts have been made to predict the Bitcoin price using blockchain network-based features [25, 27] The consensus of such studies shows up-down Bitcoin price fluctuations with classification accuracy of roughly 55%; the model which showed the best accuracy used two hidden layers of neural network. One of the conclusions drawn is that only a limited amount of predictive information is embedded in network features proving that a better approach in predicting the price relies on the information related to the financial exchanges. The model that produced the highest accuracy on the price prediction was done by LSTM RNN in [30] although Support Vector Machine (SVM), Random Forest and Binomial Generalized Linear Model (GLM) were previously used to explore how to efficiently trade with Bitcoin [27].

A study has explored the relationship between time-series [24] and sentiment analysis using SVM algorithms to determine the factors that influence the price of Bitcoin. The authors found a positive correlation between Bitcoin price and Bitcoin users' sentiment and activity on Twitter regarding Bitcoin. Bitcoin price was also revealed to be correlated with the exchange rate between the USD and the Euro, the number of Bitcoins in circulation and the level of the Standard and Poor's 500 (S&P500) stock market index. This relationship is outdated, however, as the study was done in 2014-2015, when the overall price of Bitcoin was generally going down. The direction of the price has changed since then and has moved with S&P500 index. Several articles have also shown how social media aside from Twitter, blogs, articles and other sources of information impact on Bitcoin price [29, 28] thus providing an opportunity to test other models on analyzing Bitcoin data.

---

[1] Latency , Throughput, Developer support, Size & Bandwidth issues.

[2] Currency exchanges and large mining pools are major targets of Distributed Denial of Service (DDoS) attack [8], Various types of Bitcoin financial scams, Market-based centralization on mining power, and Duplicate key generation of elliptic curve cryptography (ECC) [4].

2.3 Future ideas

Blockchain is a huge source of information for Big Data [20]; and with its consistent growth, the pace of technology has to keep up with increasing demands for cryptography services. The fact that a deep and organized market for high-quality Bitcoin-denominated bonds could emerge in the near future [20], is adding a new point of view to the study of price discovery process. At this point, it is still uncertain if liquidity, political and technological risks could influence interest rates on Bitcoin deposit, exchange rate stability or on Bitcoin-denominated bonds.

Overall, building a strong credibility appears to be the biggest challenge that Bitcoin faces in developing a viable bond market. Thus, analyzing people's views about Bitcoin and other financial market influencers can help us predict the next stage of Bitcoin evolution.

3 Design of System

3.1 Influencing factors

It is important to first determine the main variables that influence a model before designing a system. Based on [24, 23], we observed two main categories that influence fluctuations in the price of Bitcoin: Finance and Sentiment. In finance, stock markets such as S&P500 and several commodities influence the price of Bitcoin.

In the same way, media is known to play a very important role in Bitcoin price fluctuation [29]. That is, if more and more people put their trust in Bitcoin by simply relying on others' positive social media posts on Bitcoin, its price will likely go up as the demand will increase; the reverse is true when the posts about Bitcoin have a negative tone. To effectively understand these fluctuations, we explored and analyzed the most informative sources of sentiment concerning Bitcoin. We used Twitter [15] feeds and articles [5] from www.coindesk.com to achieve this goal.

We also determined the most appropriate input and output data before attempting to build the system. This was performed by investigating correlations between different variables. Here we used Quandl.com's API of [9], huge database of financial datasets to collate our financial data. Moreover, we carried out correlation checks on 152 different commodity prices and S&P500 index and we found a high positive correlations between the price of Iridium, Palladium, Aluminum, Cobalt and Random Length Lumber Futures and that of Bitcoin.

Finally, to streamline the tedious process of collecting and preparing text-related data, we generated an automated pipeline that scraped articles and tweets directly from the Twitter website.

3.2 System Architecture

Using our vast collection of finance and sentiment data inputs, we reduced our objective to applying a classical binary classification approach that predicted the daily direction of bitcoins currency (whether upward or downward). Each sentiment source, tweet and article on a given day, had 3 polarity variables: positive, negative and neutral as well as subjectivity variable, while every financial source has a single variable.

To qualify a sentiment, Natural Language Processing was performed with TextBlob [14] library by extracting an informative set of data from the text, consequently giving us an overall sentiment indicator of the day. This then served as the predictor to the next days' sentiment on Bitcoin.

Upon determining the values that describe the sentiment of different days and the financial predictions for the next day, we then used different classification algorithms to predict the direction of Bitcoin's price.

In summary, our system was built by combining multiple models. Predicted prices were identified by analyzing financial data from Quandl's API [9] using LSTM RNN on the actual Bitcoin price. LSTM RNNs are very efficient in predicting time-series related data [3]. Sentiments were determined from posts/news collected from Twitter [15] and www.coindesk.com [5] by implementing NLP. Classification algorithms[4] were then used to classify these two sets of data into several factors that influence the (upward or downward) direction of Bitcoin price (Fig. 1).

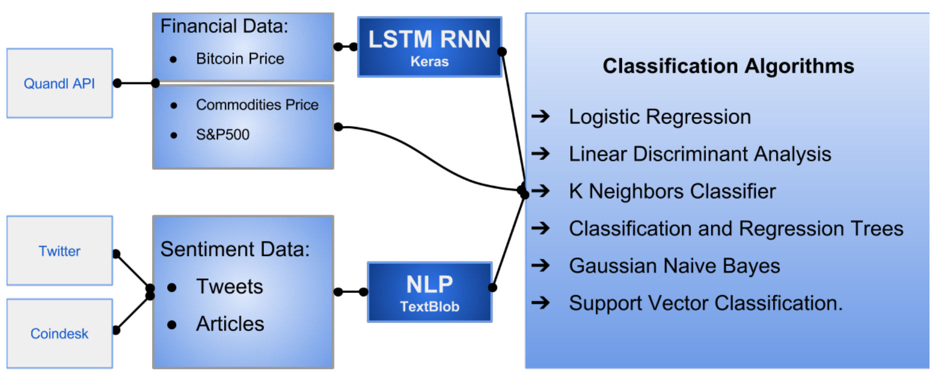

Fig. 1: Design of the system architecture.

4 Implementation of System

We implemented our system using several high-level Python modules listed in the code uploaded in DCU gitlab [12].

[3] e.g. financial (price) data, streaming pixels or images etc

[4] Logistic Regression, Linear Discriminant Analysis, K Neighbors Classifier, Classification and Regression Trees, Gaussian Naive Bayes, Support Vector Classification.

4.1 Dataset choice

To precisely implement our idea, we first decided which data sets and timestep to use. The financial data was only available on a daily basis due to commercial reasons. Thus, we were only able to work with the timestep of one day.

4.2 Data Collection

Upon collection of financial data from Quandl's API, we were faced with the difficulty of collating Bitcoin-related tweets and articles necessary for our sentiment analysis. So we created multiple bots to accomplish this task. We used Selenium [11] as a crawler to emulate a real browser. This is essential to get an Ajax [16] type of data stream from Twitter. The same method was applied for article collection, the difference was that instead of scrolling down, we had to click on the page that brings us to the previous article until the end. Taken together, we were able collect 8620 articles and over 7,000,000 tweets.

4.3 Data preparation

The preparation of our collected financial data was straightforward: the missing values were supplemented with the previous date since commodity financial markets are not open during the weekends.

The tweets, on the other hand, had to be rid of emoticons, hashtags (#), '@', '\n', etc. leaving only words for downstream analysis. Duplicated tweets or retweets were also removed. Running this process brought the tweet count down to approximately 5,000,000. The Bitcoin-related articles were already structured so preprocessing of the text context was not necessary.

4.4 Modeling

The modelling process is divided into two stages: (1) converting the dimension of our thesis problem from multiple to binary, and (2) resolving the binary-classified problem.

4.4.1 First Phase The first stage involved conversion of our text data to numerical values using NLP in tandem with a Python library called TextBlob. This module returns the polarity of a text, which can either be positive (>0), negative (<0) or neutral (=0). This module gives the variable of subjectivity which can be measured by the number in the interval [0, 1], 0 corresponding to an objective statement and 1 to a subjective statement. TextBlob was also used to calculate the number of positive, negative and neutral texts(tweets or articles) of the day with a degree of subjectivity and then we output the percentage of each one of them. This gave us 8 variables for both Twitter and Bitcoin-related article data outputs. These variables were consequently used to to obtain the best predictions for our Bitcoin price data. For this we utilized LSTM RNNs. Our main goal was to predict the direction, not the actual price. Thus, we trained our model onto getting the shape of the actual fluctuation using Keras [7] (including Tensorflow [13]), which has a built-in LSTM network off the shelf. To accomplish this task, we first had to differentiate the series. This step was essential as this gave us stationary time series independent from time, detrended and without seasonality as shown in Figure 2.

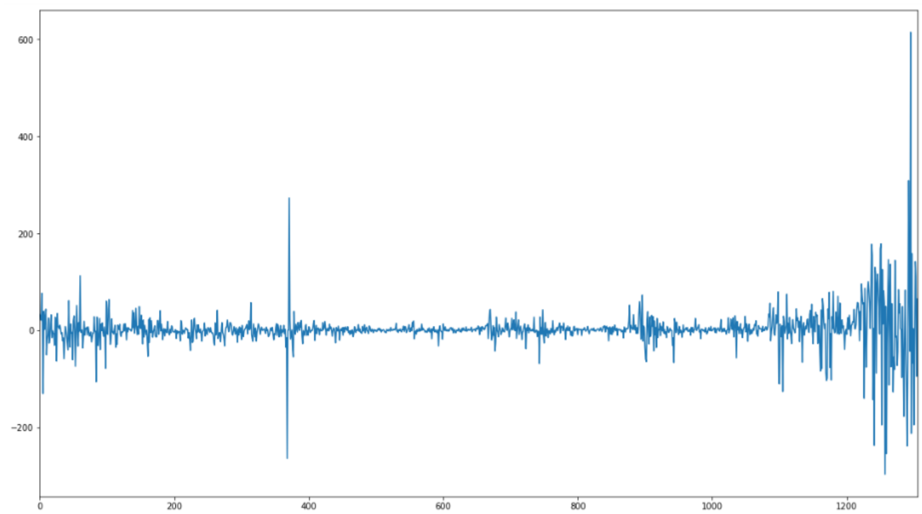

Fig. 2: Stationary time-series used as an input for the LSTM

To train our LSTM model,we tried to minimize the mean absolute error (Figure 3) and improved the accuracy (Figure 4). By doing so, we managed to make our LSTM RNN behave the same way as the actual Bitcoin price.

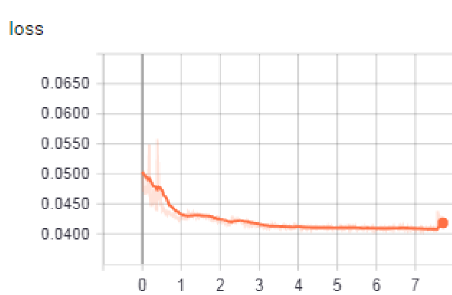

Fig. 3: Mean absolute error of LSTM RNN

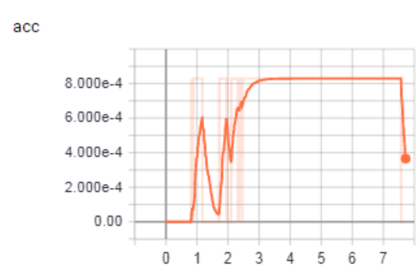

Fig. 4: Accuracy of LSTM RNN

4.4.1.1 Number of Neurons and Optimizers

To optimize our model, we identified the optimal amount of neurons to use before selecting the best optimizer. We further performed several tests on our model with a high processing capacity. We used Google cloud services, by setting up an instance with multiple GPUs (Graphics Processing Units) which considerably accelerated our testing flow by at least 10 times. To keep things simple, we have used built-in optimizers to accomplish this task. The optimizers used were: Stochastic gradient descent [7], RMSProp [7], Adagrad [22], Adadelta [7, 35], Adam [26], Adamax, Nesterov Adam [21].

The procedure to find out the best optimizer and the amount of neurons needed was to loop through the list of optimizers and by testing different values of neurons. Our results showed that the best tool for predicting the upcoming price is Adam [26]. By progressively increasing the number of neurons we were able to minimize the mean absolute error shown in Figure 5. The amount of neurons that were tested in the Figure are: 40, 50, 60, 70 and 80 represented by Dark Blue, Dark Red, Light Blue, Bright Pink and Dark green, respectively. By observing this example we can see that to avoid a local minima, we need to gradually increase our training time and reinitialise the learning rate.

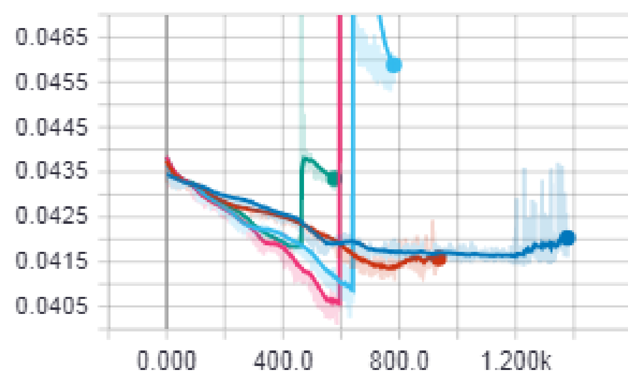

Fig. 5: Mean absolute errors while training the LSTM different configurations (40 to 80 neurons) and ADAM optimizer

We needed to make sure that on every reduction of mean absolute error we saved our state of the model and kept the error as low as possible so that we could continue training from where we left off. We used the built-in callback function to constantly save our model on each reduction of mean absolute error.

4.4.1.2 Training time

We trained our model for 3 days to obtain the right fluctuations shown in Figure 6, where the blue and yellow lines represent the actual and predicted values, respectively. Despite low accuracy, it was clear that our model was able to anticipate or predict the sudden shift of direction, conforming to our main goal. These results allowed us to compare the predicted against the actual price direction. The success rate of the direction prediction is on average 61.3%. These results show that not only LSTM RNN is capable of predicting the direction, but it is also capable of learning the fluctuation.

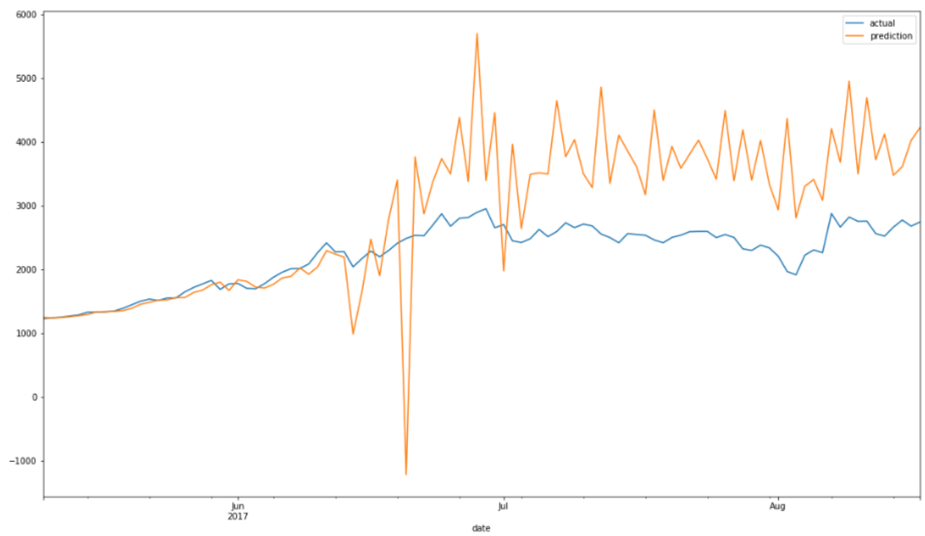

Fig. 6: Actual (blue line) vs Predicted Bitcoin price (yellow line) determined using LSTM

4.4.2 Second phase The second phase of our analysis largely involved classification of the collected data. This was initiated by splitting our dataset into training and testing portions, usually 0.7/0.3, respectively.Several inputs were provided in this phase: predicted Bitcoin price direction(0/1), Iridium price, Palladium price, Aluminum price, Cobalt price, Random Length Lumber Futures price and S&P500 index , as well as 8 sentiment variables taken from Bitcoin-related articles and Twitter resulting to 15 inputs in total. Our main target variable was the actual Bitcoin price direction.

Multiple algorithms in the scikit-learn [10] library were utilized to implement the classification process. These include Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbour, Classification and Regression Trees, Gaussian Naive Bayes and Support Vector Machine. In addition, we used Keras to generate customized neural networks to compare whether we can beat specialized algorithms. We then identified which of these algorithms have highest accuracy rate to finalize our model.

5 Evaluation of system

We evaluated the performance of our system using test samples. The results on average were:

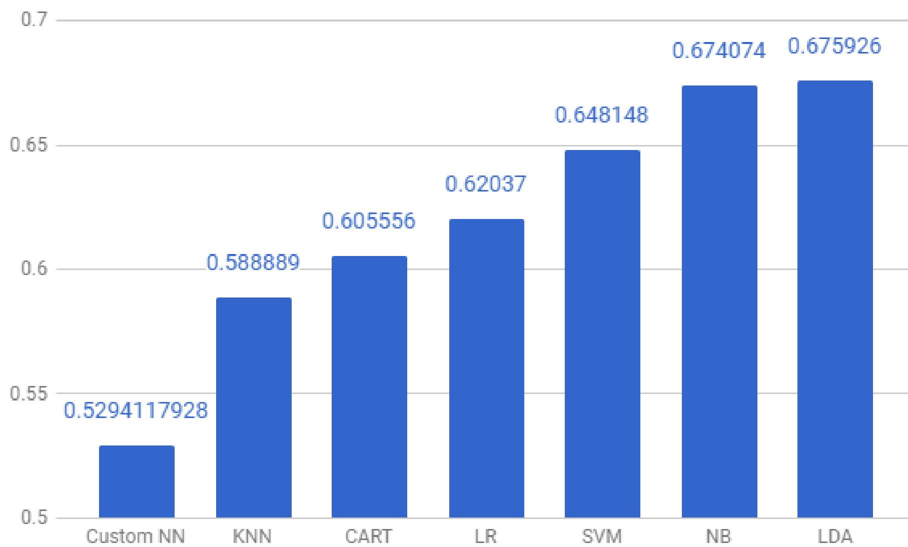

Fig. 7: Mean Accuracy scores of the algorithms tested

5.1 Result analysis

Our analysis revealed that our model had higher accuracy in predicting the direction of Bitcoin price movements in contrast to those used in previous studies, which had the accuracy between 51-61% [25, 29, 30]. This accounts for the fact that our model managed to incorporate user sentiments in the prediction process. This in turn allowed us to additionally assess the results produced by LSTM RNN. We acknowledge, however, that further adjustments can be applied to refine our system.

Moreover, our model showed that sentiment does not have an immediate effect on the market or the currency. That is, the negative or positive user sentiments usually take some time to impact Bitcoin price and that a day of timestep works more appropriately for this type of model. Finally, treating Bitcoin as a commodity provided us with an opportunity to design a model that somehow reflected the reality of Bitcoin's behaviour as a rare commodity rather than a mere cryptocurrency.

6 Conclusion/future work

We created a model that predicts Bitcoin's price fluctuation. Initial observations revealed that changes in Bitcoin price movements were mainly influenced by the sentiment of the community and social media. Running the system we generated for this project allowed us to validate this hypothesis and gave us a better understanding on how Bitcoin behaves in the market. Overall, cryptocurrencies particularly Bitcoin are still in their infancy making it hard for users to predict how they will evolve in the future. The model we developed in this paper, therefore, sheds light on this issue as it does not only predict the price fluctuations of Bitcoin but also has the potential to evaluate and anticipate the market behaviour of other cryptocurrencies as well.

References

1. Bitcoin, https://github.com/bitcoin/bitcoin
2. Bitcoin avg. transaction fee chart, https://bitinfocharts.com/comparison/bitcoin-transactionfees.html#3m
3. Bitcoin block explorer, https://blockchain.info/
4. Blockchain warns of duplicate bitcoin addresses on android, https://www.financemagnates.com/cryptocurrency/news/blockchain-warns-of-duplicate-bitcoin-addresses-on-android/
5. Coindesk, http://www.coindesk.com
6. Cryptocurrency market capitalizations, https://coinmarketcap.com/all/views/all/
7. Keras documentation, https://keras.io/
8. Major ddos attacks hit bitcoin.com - bitcoin news, https://news.bitcoin.com/ddos-attacks-bitcoin-com-uncensored-information/
9. Quandls api, https://www.quandl.com/
10. Scikit-learn, http://scikit-learn.org/stable/
11. Selenium with python, http://selenium-python.readthedocs.io/index.html
12. Source code, http://gitlab.computing.dcu.ie/kinderm3/practicum_2017_Bitcoin_currency_fluctuation/tree/master
13. Tensorflow, https://www.tensorflow.org/
14. Textblob, textblob 0.13.0 documentation, http://textblob.readthedocs.io/en/dev/
15. Twitter, https://twitter.com/
16. What is ajax and where is it used in technology?, https://www.seguetech.com/ajax-technology/
17. Amjad, M., Shah, D.: Trading bitcoin and online time series prediction. In: NIPS 2016 Time Series Workshop. pp. 1–15 (2017)
18. Badev, A.I., Chen, M.: Bitcoin: Technical background and data analysis (2014)
19. Chu, J., Nadarajah, S., Chan, S.: Statistical analysis of the exchange rate of bitcoin. PloS one 10(7), e0133678 (2015)
20. Chuen, D.L.K.: Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data. Academic Press (2015)
21. Dozat, T.: Incorporating nesterov momentum into adam (2016)

22. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12(Jul), 2121–2159 (2011)
23. Dyhrberg, A.H.: Hedging capabilities of bitcoin. is it the virtual gold? Finance Research Letters 16, 139–144 (2016)
24. Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D.N., Giaglis, G.M.: Using time-series and sentiment analysis to detect the determinants of bitcoin prices (2015)
25. Greaves, A., Au, B.: Using the bitcoin transaction graph to predict the price of bitcoin (2015)
26. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
27. Madan, I., Saluja, S., Zhao, A.: Automated bitcoin trading via machine learning algorithms (2015)
28. Mai, F., Bai, Q., Shan, Z., Wang, X., Chiang, R.: From bitcoin to big coin: The impacts of social media on bitcoin performance. SSRN Electronic Journal (2015)
29. Matta, M., Lunesu, I., Marchesi, M.: Bitcoin spread prediction using social and web search media. In: UMAP Workshops (2015)
30. McNally, S.: Predicting the price of Bitcoin using Machine Learning. Ph.D. thesis, Dublin, National College of Ireland (2016)
31. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008)
32. Park, H.K.: How many people in the world own bitcoin or ethereum?, `https://hankyulpark.wordpress.com/2017/03/24/how-many-people-in-the-world-own-bitcoin-or-ethereum/`
33. Yli-Huumo, J., Ko, D., Choi, S., Park, S., Smolander, K.: Where is current research on blockchain technology?a systematic review. PloS one 11(10), e0163477 (2016)
34. Żbikowski, K.: Application of machine learning algorithms for bitcoin automated trading. In: Machine Intelligence and Big Data in Industry, pp. 161–168. Springer (2016)
35. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)