Forecasting Bitcoin Prices Using Deep Learning Techniques and Sentiment Analysis

*Gitlab Link: https://gitlab.computing.dcu.ie/khaira2/ca683_assignment_bitcoin_price_prediction

Sashank Phatkare
School of Computing
Dublin City University
Dublin, Ireland
sashank.phatkare2@mail.dcu.ie

Aakash Khair
School of Computing
Dublin City University
Dublin, Ireland
aakash.khair2@mail.dcu.ie

Aditya Kulkanri
School of Computing
Dublin City University
Dublin, Ireland
aditya.kulkarni5@mail.dcu.ie

Aditya Vadagave
School of Computing
Dublin City University
Dublin, Ireland
aditya.vadagave2@mail.dcu.ie

***Abstract*—Bitcoin was the very first Blockchain - created by an individual called Satoshi Nakamoto in 2008. The value of bitcoin has increased to a ridiculous level. You must have seen news articles about " if I had brought $100 of bitcoin last decade, I would have had more than $100 million now". Everything aside, digital currency such as Bitcoin has a mining limit of 21 million, so in no way the total amount of cash in the system can be increased by any central bank. By nature , Bitcoin itself is scanty. While there has been extensive research done on Bitcoin, most of it has focused on the behavior of Bitcoin. In this paper, we propose the use of Deep Learning techniques along with Sentiment Analysis using Google news data to forecast future Bitcoin prices. We compare the use of LSTM, GRU and Bidirectional LSTM used in predicting Bitcoin prices while using elements from Google news as input features for Sentiment Analysis. The results show that it is possible to predict bitcoin prices using Deep Learning models such as LSTM and GRU when predicting long term trends followed by sentiment analysis on news articles to try predicting if they have any effect on future values of Bitcoin.**

***Index Terms*—CRISP-DM, Deep Learning, Sentiment, Long Short Term Memory**

I. Introduction

Bitcoin is a medium of exchange that is digital, encrypted and decentralized. Unlike the major currencies, there is no central owner or authority that manages and maintains the value of a cryptocurrency. Instead, these tasks are broadly distributed among cryptocurrency's users via the internet. The rapid evolution of digital cryptocurrencies over the last decade is perhaps one of the most tricky and ambiguous trends in the cutting-edge global economy. The structure of markets, monetary based industries and payment mechanisms are all changing as a result of increased competition. A surprising, swift occasion like this happened in 2011 when Bitcoin prices surged in a very short amount of time. Bitcoin's cost bounced from 1 USD in April of that year to a pinnacle of 32 USD in June, an addition of 3200% within just three months [1] [2]. The Crypto Research Report, a paper which helps numerous Bitcoin and blockchain analysts to learn the market's trends, forecasted a major increase in the next decade. According to a survey, field analysts estimate that the price of Bitcoin would hit 397,000 USD by 2030 which includes both tangible and intangible assets [3].

Price volatility is the significant issue with intangible digital assets, especially cryptocurrencies. The price of Bitcoin (BTC) for the period of October 1, 2017 to March 31, 2021 can be seen in Fig. 1. Over this time, Bitcoin values have become extremely volatile. The price rose by 1900% in 2017, only to lose 72 percent of its value in 2018 [4]. Prior to 2013, there was no public interest in Bitcoin, it was rarely used in virtual transactions, and its values was poor. Our models do not take that time frame into account. While BTC prices are extremely volatile, BTC as a digital commodity is very robust in that it can recover its value after large declines and even when market instability is high, such as during the COVID-19 pandemic [5].

In this study, we implement and compare various state-of-the-art Deep Learning methodologies like Long short-term memory (LSTM) model, Gated Recurrent Units (GRU) and Bidirectional LSTM. We also conducted a sentiment analysis on google news data to examine whether we can improve the bitcoin price predictions using the results from same. The sentiment data is very unpredictable in nature and can sometimes lead to higher risks of manipulation in the market. Overall, the performance of the Deep Learning-based prediction models were compared and evaluated in this study.

Research Question:

1. How effective can sentiment analysis be for predicting bitcoin prices as compared to the deep learning models?

This research contains following sections. Section II presents a discussion on the literature review of related work done. The exploratory data analysis on historic prices is elaborated in Section III. Section IV provides insights on methodology and feature engineering. Section V contains evaluation of the models along with sentiment analysis. Corresponding conclusions have been discussed in final section VI followed by references.

II. Related Work

This paper is built on the ideas from various researches and topics. Over the past decade Deep Learning and Machine learning techniques have been widely used in cryptocurrency price prediction. There are currently many price forecasting models available. The source [6] deals with time series data consisting of daily Bitcoin closing prices between 2018-2021 followed by performing Linear Regression (LR) and Support Vector Machine techniques. These models were then compared to see which one provides better and accurate results.

The author in [7] aims at discovering the most efficient and high accuracy model for predicting bitcoin prices using various machine learning algorithms. They have used 1-minute interval trading data from the Bitcoin exchange website named Bitstamp. They worked on regression models with scikit-learn and Keras libraries. Their best results showed that MSE was as low as 0.00002 and the R-square was as high as 99.2%. Price prediction models have been a huge interest but research on predicting the price using machine learning algorithms is specifically lacking. [8] implemented a Bayesian optimized Recurrent Neural Network (RNN) and a Long Short-Term Memory (LSTM) network. Their LSTM approach achieved a classification accuracy of 52% and a RMSE of 8%. They also implemented ARIMA model for time series forecasting as a comparison to their deep learning models. It is a matter to try, understand the factors that influence the bitcoin price formation. In [9] the author has used advanced Artificial Intelligence framework of fully connected Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) model to analyze the price predicitons of Bitcoin, Ethereum and Ripple. There is aldo work done using textual data from social media platforms and similar sources to try and predict the future prices [10] In this paper the author has utilized Twitter data and Google trends data for predicting changes in Bitcoin and Ethereum prices. They used a linear model which took tweets and google trends data as inputs for their configuration. Similarly, in [11] the author proposed the usage of common machine learning tools and available social media data for predicting the prices of Bitcoin, Ethereum, Ripple and Litecoin cryptocurrency market moments. They compared the utilization of neural networks, support vector machines and random forest along with using the elements from twitter and market data as input features.

However, one limitation of these studies is often small sample size as well as tendency of misinformation being spread through various social media platforms which in turn artificially inflates/deflates prices. As a result the market suffers from much higher risk of manipulation. For this reason, sentiment analysis from Twitter data is not considered viable further.

III. Dataset And Exploratory Analysis

Construction of dataset involved four major steps. Data was extracted from (1) Google News and (2) Cryptocurrency exchanges as part of the data mining phase. Sentiment analysis was performed on the text data from Google News in the third phase. To create the final dataset, the sentiment results were concatenated with historical cryptocurrency prices in the fourth phase. Entire data was obtained with a granularity of one-minute interval from December 31, 2011 to March 31, 2021. The only exception was Google News, where the data was extracted on a daily basis from January 1,2020 to March 31,2021.

A. Google News Data

a) Data Collection: To get data from Google News, a Python script is implemented that uses googleapi and other relevant libraries. The script makes use of the Python library 'requests' [12] to send HTTP GET requests with specially formatted URL. The API has a function that allows you to search for news using keywords and a date range. All the news articles were collected in order to conduct sentiment analysis, as precisely as possible. The 'newspaper' [13] Python library is used to retrieve the URLs of news articles during the initial Google News request. With the aid of this script the first ten news articles for each day are obtained.

b) Sentiment Analysis: The text from the news article is fed into pre-trained sentiment analyzer machine learning model, which generates numerical values. This is done with the entire dataset, and the final sentiment value is divided by the amount of articles parsed for the day to get the every-day's normalized sentiment value. Valence Aware Dictionary and Sentiment Reasoner (VADER) [14], TextBlob [15] and Flair [16] functions are used to perform sentiment analysis. Flair is a state-of-the-art natural language processing (NLP) model and allows named entity recognition (NER), part-of-speech tagging (PoS), sense disambiguation and classification. VADER is a lexicon and rule-based sentiment analysis tool that is specifically tailored to process social media sentiments. TextBlob is a simple NLP tool that allows to do things like PoS tagging, noun phrase extraction, sentiment analysis and classification.

B. Bitcoin historical Data

The data has been scrapped from Binance [17] and Coinbase [18] with the help of Python-based client APIs. Data collected by this script has 8 features namely timestamp, open, high, low, close, volume_btc, volume_currency, weighted_price.

C. Exploratory Analysis

The line graph depicted in Fig.1. shows the volatility of the bitcoin closing prices right from 2012 to 2021 in USD. We saw gradual rise in the price during the end quarter of 2013 and beginning of 2014. Post that, the price value remained constant throughout the time till the mid of 2016. While on this trend, we saw a characteristically huge spike in 2018 which was a gradual result of the steady increase in the beginning of 2017. This was the first time in the history of Bitcoin where it's price touched a 20000 USD figure. After this, we saw a significant slump in the price values which continued till the beginning of

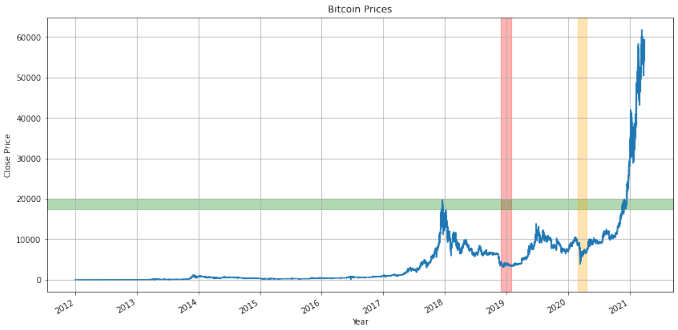

Fig. 1. Distribution of Closing Prices

2019 where it dropped with a major blow touching a longtime low of around 5000 USD.

The graph shown in figure 2 below carefully depicts the value range for Bitcoin after the slump of 2018 till the beginning of 2019 in January. The prices seem to have taken a roller-coaster ride wherein they shifted from a market value of approximately 6000 USD to roughly 2200 USD. The trend in between was a constant cycles of high and low closing values.

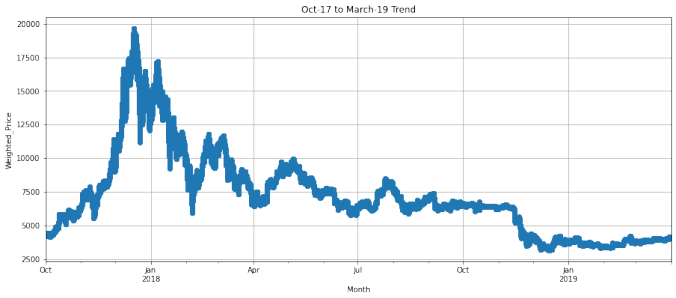

Fig. 2. October-2017 to March-2019 Trend for Closing Prices

With a good spike of roughly around 15000 USD per unit, the prices continued a normal trend until reaching the beginning of 2020. It was followed by a good rise in the values with an increased margin of 4000 USD. March 2020 was the period when the per unit price saw a big drop to 3000 USD. This was the period when the market was hit by the COVID19 pandemic and the subsequent lockdown. We analyse, expect that this drop was majorly because of low investments by industrialists, investors who suffered huge losses on businesses and stocks. The prices continued a low continued normal until the end of 2020 and the first patch of 2021 where the bitcoin unit price saw the biggest increase, prices shooting upto 30000 USD followed by 55000 USD and reaching the all time highest of 63,729.5 USD.

The relationship between the parameters has been studied with the help of a correlation matrix. The heat map below helps us understand the correlation coefficient between the features. The matrix of correlation states that the connection between given attributes is in the range from -1 to +1. There was a weak and neutral relationship between all data attributes. The ”Divide and Conquer” approach is therefore used here

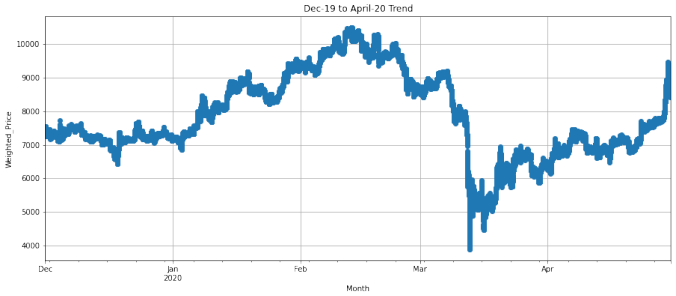

Fig. 3. December-2019 to April-2020 Trend for Closing Prices

to analyze the situation in depth. The weighted price is adjusted for the volume. Open, high, low, close are linked with weighted price directly.

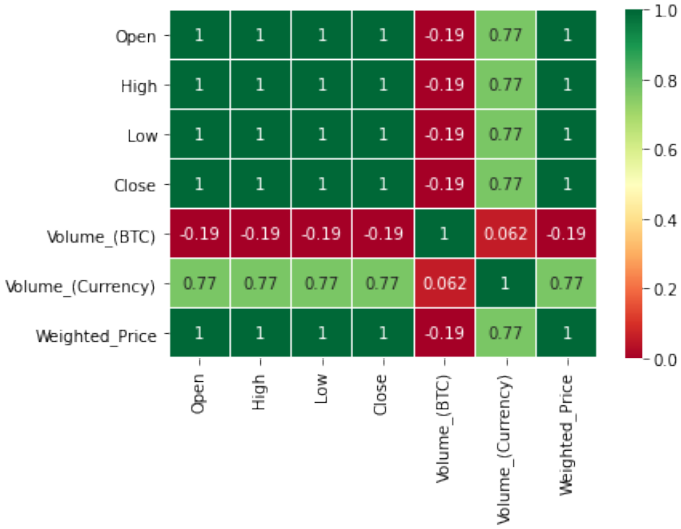

Fig. 4. Heat map of Correlation Matrix

IV. Methodology

This paper follows the CRISP data mining methodology.The motivation for CRISP-DM over the more traditional KDD [19] revolves around the business setting of the prediction task. This methodology has been divided into two sections:- The first implementation evaluates three distinct deep learning models used in forecasting the bitcoin prices thereby defining and analyzing relevant features. After applying the three models for price prediction, the model with highest accuracy has been determined for the future fulfillment of target by selecting appropriate parameters to obtain more better performance. In this assesment, we propose deep learning mechanisms of LSTM, Bidirectional LSTM and GRU. Since Bitcoin is the most popular cryptocurrency, the price volatility issue should be looked into solved by the coming period of time. [20].

In the second implementation, sentiment analysis is done along with LSTM neural network to predict the trends in the bitcoin prices. The system pre-processes the google news feeds to sentiment analyser. The sentiment analyser gives sentiment

percentage of each day which is fed to the LSTM predictor along with the historical price of bitcoin. The model then finally predicts the price.

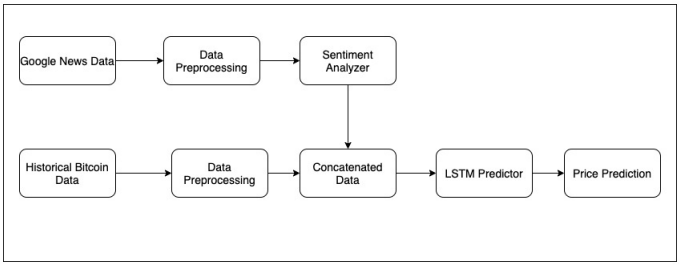

Fig. 5. System Design

A. Data Normalization

Min-Max scaler was used to adjust the data into a -1 to 1 range by using the data's min and max [9]. Equation (1) can be used by MinMaxScaler in scikit-learn to transform data of a selected function.

$$x_{scaled}^{(i)} = \frac{x^{(i)} - min(x^{(i)})}{max(x^{(i)}) - min(x^{(i)})} \quad (1)$$

Since the hyperbolic tangent (tanh) function is the default activation function for machine learning models, and it outputs a range of -1 to 1, the values have been scaled to match this range, and MinMaxScaler has been used to do so. The fit() function was used to fit the scalar to the training set. The transform() function was then used to scale the training set, which was then applied to the test set.

B. Splitting of Data

The dataset is divided into an 80:20 ratio, with 80% of the data used for training and 20% for research. Splitting of the dataset is done to gauge to the two most important aspects used in the predictions, over-fitting and under-fitting. Over-fitting occurs when the model is trained too well and the predictions are too similar, while under-fitting occurs when the model does not match as closely as it should.

C. Modeling

The three models used in this project includes implemented stacking of layers to allow encoding of complex function from input to output. The reason for choosing 2 layers in all the mentioned models is that on running multiple trials, the accuracy and predictions were better as compared to any different configuration.The Fig. 6. represents a stacked neural network model where x is the input fed to the system and h(x) is the output evaluated from the predictive analysis.

1) LSTM: Long short-term memory (LSTM) is a type of a recurrent neural network which is implemented in the lines of deep learning. This model uses feedback connections otherwise feedforward neural networks using forward connections. The common architecture of LSTM

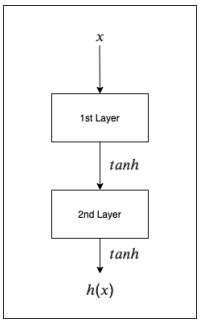

Fig. 6. Stacked Layer Representation

model includes an input gate, an output gate and a forget with a cell. These are mostly used for making predictions, classifying on projects with time-series as the prime data. In the proposed model we have used 2 layers of LSTM, 64 units and 50 epochs while framing the base skeleton.

2) Bidirectional LSTM: The second model is a Bidirectional LSTM model. The crux of this model is that it consists of two LSTM models both working in the opposite directions of forward and backward. It effectively boosts the amount of data available to the neural network giving a more better, proper context to the algorithm for working. Mean Squared Error(MSE) has been used as the evaluation metric.The configurations used in this model are similar to the one used for LSTM.

3) GRU: Gated Recurrent Unit (GRU) implements a gating mechanism in any recurrent neural network it is implemented on. It is exactly similar to the LSTM model with a forget gate but handles lesser number of parameters then the former, as it lacks the output gate. GRU has shown better performance results on various smaller and frequent datasets as compared to similar approaches like LSTM or multi-layer LSTM. Adam optimizer has been chosen to handle the sparse gradient in all the three mentioned models.

V. Evaluation And Results

a) Implementation I: Visual inspection reveals that both the predicted scores from the LSTM and BiLSTM closely match the test data until the sharp rise in values near the end of the data set.

The GRU model, on the other hand, appears to be the most well-fitting in overall. It appears to be better at following the demographics present in the data, even if it does not accurately reflect the true data.

We evaluated performance of the models by using mean squared error (MSE) and mean absolute error (MAE). The MAE seems to be more in accordance with the tight fit seen in figure 3 above. The MAE values of both the BiLSTM and the LSTM are lower, indicating a more accurate model by considering this metric. The GRU, on the other hand, has a lower RMSE, implying that it is a more accurate model using

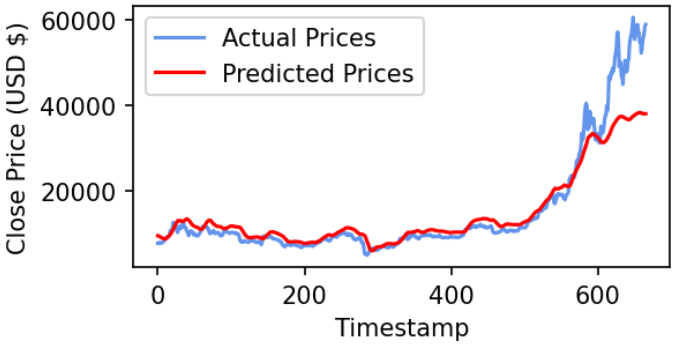

Fig. 7. Actual Future vs Prediction for Bidirectional LSTM

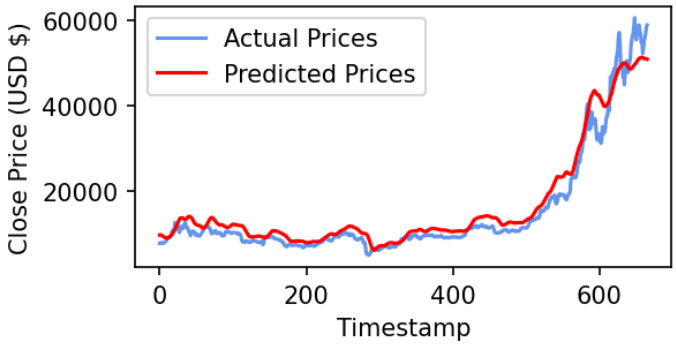

Fig. 8. Actual Future vs Prediction for LSTM

this evaluation metric. Suppose 1 Bitcoin will be worth 50,000 USD tomorrow, for example, is useless without context. The predicted upward trend would be more informative if it is known that 1 Bitcoin is worth 40,000 USD today. The GRU model performed better as an indicator of trend, rather than price.

b) Implementation II: The second implementation is sentiment analysis of news articles using the LSTM model to predict the bitcoin prices. After analyzing the models at the earlier stage, we find that the Bi-directional model under-fits, the GRU model over-fits but the LSTM model is close to consistent over the time. Hence we chose to apply the

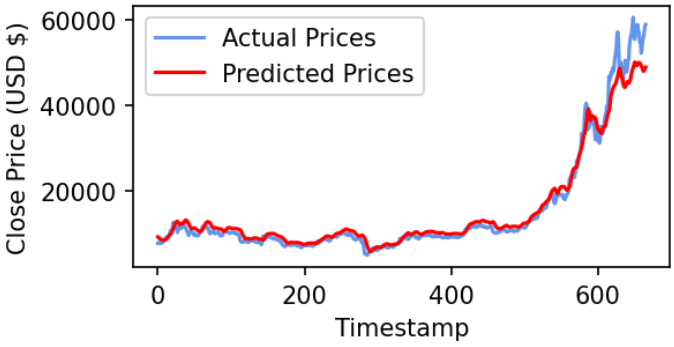

Fig. 9. Actual Future vs Prediction for GRU

TABLE I
IMPLEMENTATION I

Model Head	Evaluation Metrics		
	MAE	***RMSE***	***Accuracy***
LSTM	2075.64	2761.17	95.35
Bi-LSTM	2462.00	4802.38	85.92
GRU	1355.47	2216.99	97.00

same LSTM technique to implement the Sentiment Analysis of the news articles. This technique aims to showcase the relationship between the dependent variable, which is Bitcoin price in USD and the independent variable, which is the compound sentiment score provided by VADER algorithm for each chosen time interval. We present the model's prediction in form of a line chart below in Fig. 10.

TABLE II
IMPLEMENTATION II

Model Head	Evaluation Metrics	
	MAE	***RMSE***
LSTM	25858.77	27110.46

The RMSE and MAE scores calculated are way too high thereby indicating a huge training loss which clearly depicts that the model is under-fitting.

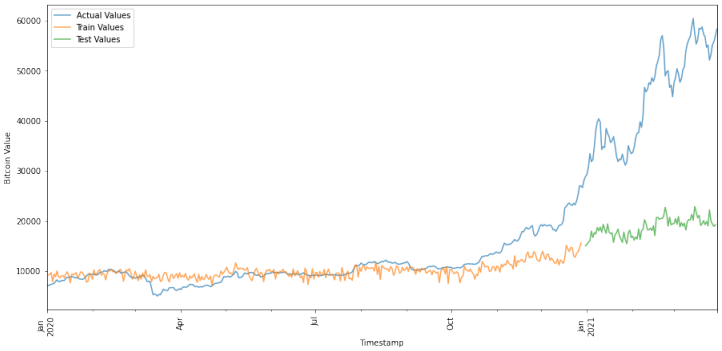

Fig. 10. Actual vs Train vs Test for LSTM using sentiment analysis

VI. CONCLUSION

Deep learning models like the LSTM and GRU perform better on training data, with the LSTM being better at identifying longer-term trends. Conversely, with such a high variance analysis, it is hard to articulate prediction patterns and values. There is a fine line to maneuver between overfitting a model and best fit. To prevent the model from overfitting, an ensemble technique can be used with appropriate dropout and layers. The results of our analyses demonstrate that sentiment analysis is less effective for predicting bitcoin prices when the values are experiencing a sudden spike or drop. This is due to the fact that cryptocurrency news can be less subjective at times.

REFERENCES

[1] J. Edwards. Bitcoin's price history. [Online]. Available: https://www.investopedia.com/articles/forex/121815/bitcoins-price-history.asp

[2] Bitstamp. The original global crypto exchange. [Online]. Available: https://www.bitstamp.net/markets/btc/usd/?action=buy

[3] *Crypto Research Report*, ser. Modeling Bitcoin's Price with Irving Fisher's Equation of Exchanges, 2020.

[4] S. D. Saloni Shukla. Bitcoin beats coronavirus blues. [Online]. Available: https://economictimes.indiatimes.com/markets/stocks/news/bitcoin-beats-coronavirus-blues/articleshow/75049718.cms

[5] D. Morris. Bitcoin hits a new record high, but stops short of $20,000. [Online]. Available: https://fortune.com/2017/12/17/bitcoin-record-high-short-of-20000/

[6] S. Karasu, A. Altan, Z. Saraç, and R. Hacioğlu, "Prediction of bitcoin prices with machine learning methods using time series data," in *2018 26th signal processing and communications applications conference (SIU)*. IEEE, 2018, pp. 1–4.

[7] T. Phaladisailoed and T. Numnonda, "Machine learning models comparison for bitcoin price prediction," in *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 2018, pp. 506–511.

[8] S. McNally, J. Roche, and S. Caton, "Predicting the price of bitcoin using machine learning," in *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*. IEEE, 2018, pp. 339–343.

[9] W. Yiying and Z. Yeze, "Cryptocurrency price analysis with artificial intelligence," in *2019 5th International Conference on Information Management (ICIM)*. IEEE, 2019, pp. 97–101.

[10] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, "Cryptocurrency price prediction using tweet volumes and sentiment analysis," *SMU Data Science Review*, vol. 1, no. 3, p. 1, 2018.

[11] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, "Price movement prediction of cryptocurrencies using sentiment analysis and machine learning," *Entropy*, vol. 21, no. 6, p. 589, 2019.

[12] Requests python library. [Online]. Available: https://2.python-requests.org/en/master

[13] Newspaper python library. [Online]. Available: https://newspaper.readthedocs.io/en/latest/

[14] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1638–1649.

[15] C. Hutto and E. Gilbert, "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*, vol. 8, no. 1, 2014.

[16] Textblob python library. [Online]. Available: https://textblob.readthedocs.io/en/dev/

[17] Binance: Buy sell crypto in minutes. [Online]. Available: https://www.binance.com/en

[18] Binance: Buy sell crypto in minutes. [Online]. Available: https://www.coinbase.com/

[19] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[20] S. Anderson. Why bitcoin has a volatile value. [Online]. Available: https://www.investopedia.com/articles/investing/052014/why-bitcoins-value-so-volatile.asp