Understandng and Predicting the Dogecoin price based on Twitter sentiments

Ruchil Barya[1]

[1] Harrisburg University of Science and Technology

Author Note

Department of Analytics, Harrisburg University of Science and Technology

Correspondence concerning this article should be addressed to Ruchil Barya, Harrisburg University. E-mail: rbarya@my.harrisburgu.edu

Understandng and Predicting the Dogecoin price based on Twitter sentiments

Abstract

As social media is becoming an essential part of life, it has increasingly started influencing many complex systems which were not considered before. Recent events like AMC, GameStop, show that social media has the power to enhance or reduce the value of financial assets. In this paper, the relationship between cryptocurrency prices and Twitter is investigated. The primary goal of this study is to understand the effect of tweets on Dogecoin prices. This research focuses on the prediction of Dogecoin prices with the sentiment score via VADER sentiment analysis. This study uses an advanced deep learning technique called LSTM (Long Short Term Memory) to forecast the prices for a 15 days window.

Keywords - Cryptocurrency, Twitter, LSTM, Sentiment Analysis, Dogecoin

Multiple technological developments, such as artificial intelligence, internet of things, and digital payments, have occurred in recent decades.(Rüßmann, 2003)These advancements have an impact not only on how people interact with the world but also on how they exchange money for services. Banking, transaction, and other e-commerce activities are becoming more advanced as a result of new technologies.(Lahmiri & Bekiros, 2019) With these advancements, a new form of currency known as cryptocurrency has evolved as a result of the trend of switching from printed to virtual money.

Since its emergence, cryptocurrency has seen a significant increase in use. After a major crash in 2018, cryptocurrencies have rebounded strongly in 2020. Digital currency such as Bitcoin, Ethereum has increased by approximately 1500% since mid-March of last year. The number of blockchain wallet users has nearly quadrupled from 2012 to 2021. (Chan, Chu, Nadarajah, & Osterrieder, 2017) Currently, many companies such as Tesla, Microstrategy, Twitch have started using cryptocurrency as a medium of exchange. Many educational and financial institutions are in the process of implementing cryptocurrency wallets. The increasing acceptance of cryptocurrencies has led to the rise in price and popularity of cryptocurrencies.

Cryptocurrencies have unique features that are not shared by conventional forms of currency exchange. Their prices are highly difficult to forecast since they do not act like conventional currencies. Unlike institutional controls, cryptocurrency price fluctuations are determined by people's perceptions and opinions. Social Media is the primary source of information about cryptocurrency prices. Based on the opinions on several platforms like Twitter, Reddit, Youtube Videos people have started investing in cryptocurrencies. Twitter acts as a marketing tool for cryptocurrencies because people discuss their views on them. Another platform used for this purpose is YouTube, a popular social media site for people looking for entertainment and information on various topics.(Dabas, Kaur, Gulati, & Tilak, 2019) It is widely used for learning about cryptocurrency investments around the world.

The estimated market capitalization of all cryptocurrencies is approximately USD 1.7 trillion as of March 2021. The top 15 currencies accounted for over 97% of the market capitalization, with seven currencies accounting for 90% of the total. (Chan, Chu, Nadarajah, & Osterrieder, 2017) Bitcoin went from having no value in 2008 to $63,588 on April 13, 2021 the highest price ever reported in the history of cryptocurrency.("Bitcoin hits another record," 2021) The worth of well-known digital currencies has been hitting new highs every day as a growing number of positive investors bet big on the cryptocurrency's potential to become a mainstream asset. Since cryptocurrencies are still relatively new, little research has been done on the tools that can assist people in investing in them. (Cohen, 2020) Therefore, the price prediction of cryptocurrency has become an interesting research topic all over the world.

Literature Review

In 2008, a researcher named Satoshi Nakamoto wrote a paper Satoshi Nakamoto (2008) on the concept of Peer-to-peer electronic cash transfer without the involvement of any intermediary financial institutions. He illustrated the concept of a shared chain of legitimate transactions (blockchain) that is distributed among all network peers. To validate the transaction and create a new block in the chain, a PoW algorithm is used.(Gupta et al., 2020) Since no third party is involved in the transaction execution, this P2P distributed system removes the confidence and accountability problems. This is lead to the creation of a modern type of digital currency, called cryptocurrency.(Narayanan, Bonneau, Felten, Miller, & Goldfeder, 2016)

Cryptocurrency is a decentralized, network-based, digital currency that uses advanced encryption algorithms such as Secure Hash Algorithm 2 (SHA-2) and Message Digest 5 (MD5) to regulate its currency units.(Patel, Tanwar, Gupta, & Kumar, 2020) It provides secure, traceable, and transparent transactions. Bitcoin was the first cryptocurrency to be created based on the paper of Satoshi Nakamoto. After Bitcoin, a slew of new cryptocurrencies has been launched for a variety of uses. The other popular cryptocurrencies are Ethereum, Cardano, Litecoin, Tether, Ripple, Binance Coin. The blockchain, the promising technology that underpins cryptocurrencies, makes it likely that they will continue to be used in some capacity and that their use will expand.

The price of cryptocurrencies is heavily influenced by public sentiment. The way people think about money is crucial. Behavioral economists such as Daniel Kahneman and Amos Tversky established that emotion, not just value, influences decisions, including those with financial implications.(Heukelom, 2009) Pang, Sundararaj, and Ren (2019) stated that there is a relationship between Bitcoin prices and social sentiment data, which is recommended for predicting the prices for shorter intervals of time (i.e. in minutes) in their future study section.

According to Li and Wang (2017) research, Bitcoin can be used as a hedging mechanism against global volatility and as a portfolio diversifier for a variety of indices, currencies, and commodities. Urquhart (2017) discovered price clustering for Bitcoin, and (Chu, Chan, Nadarajah, & Osterrieder, 2017) researched the predictability and volatility of cryptocurrencies using GARCH-modeling. The use of proxies for general interest, such as Google Trends or tweet volumes, is recommended in the paper by (Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan, 2018). The paper also shows that the search volume index, like tweet volumes, is strongly associated with cryptocurrency values, both when they rise and when they fall.

Jay et al. (2020) used a stochastic neural network model to forecast the price of Cryptocurrencies, and the results indicate that the proposed hypothesis was not only accurate but also successful in intercepting market volatility. Smuts (2019) which stated that Google trends and telegram data can predict cryptocurrency prices, but stopped short of demonstrating its pattern remains constant over time. Sin and Wang (2017) used a time series of daily Bitcoin closing prices between 2012 and 2018 to estimate Bitcoin prices using machine learning techniques such as linear regression (LR) and support vector machine (SVM). Cheuque Cerda and L. Reutter (2019) used historical prices and influencer sentiments as RNN inputs to try to boost performance.

In comparison to the rest of the cryptocurrency industry, Vidal-Tomás, Ibáñez, and Farinós (2021) supports the view of Bitcoin as a means of exchange rather than an asset. Given its high liquidity, the findings from Vidal-Tomás, Ibáñez, and Farinós (2021) suggest that the launch of Bitcoin's futures market improved its risk-return trade-off in comparison to the majority of cryptocurrencies. The research article Knežević, Babić, and Musa (2020) shows that young people under the age of 35 in the United States of America are more prone to working with cryptocurrencies than the ones in Croatia. The findings of Rehman and Apergis (2019) suggest that cryptocurrency returns should be used to forecast commodity market returns and volatilities, although the reverse could also be useful. These results have major consequences for risk management decisions and portfolio strategy creation.

Mohanty, Patel, Patel, and Roy (2018) forecasted Bitcoin price fluctuations using daily produced data such as price, block size, several transactions per block, and other 26 features of the Bitcoin blockchain, as well as Twitter data. When cryptocurrency prices rise or fall, the search volume index, like tweet volumes, is strongly associated. Teker, Teker, and Ozyesil (2019) depicted the relationship between regular Bitcoin, Tether, Ethereum, and Litecoin prices and shifts in gold and oil prices. Li et al. "The best of two worlds" (2018) used Generalized Autoregressive Conditional Heteroskedasticity (GARCH) in combination with Support Vector Regression to forecast Bitcoin, Ethereum, and Dash prices and found substantial progress.

To determine if Twitter sentiment is a cryptocurrency price driving factor, other cryptocurrency price driving factors must be investigated. Researchers have looked at these factors in detail for several variables. (Kraaijeveld & De Smedt, 2020) Mittal, Dhiman, Singh, and Prakash (2019) looked for a connection between tweet volume, Google search patterns, sentiment in tweets, and Bitcoin prices.

(Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan, 2018)Jain, Singh, and Vatsa (2018) used tweets to conduct cryptocurrency sentiment analysis and make predictions. Karalevičius (2017) used sentiment analysis of social media forums to forecast intraday Bitcoin prices and found that short-term price volatility could be predicted with some precision, but that this accuracy decreased as time went on. Patel, Tanwar, Gupta, and Kumar (2020) stated that cryptocurrencies can be accurately predicted using neural network algorithms, but left a gap in the analysis of sentiment analysis from youtube videos and news.

This study will predict the prices of Dogecoin and understand if Twitter sentiments affect its prices. The underlying assumption is that the sentiments and cryptocurrency prices are correlated. This research will try to find the best method to predict the Dogecoin prices with and without sentiment analysis score. This will help to understand the impact twitter sentiments make on cryptocurrency prices.

Methods

This research is based on both quantitative and qualitative perspectives. The qualitative perspective will analyze the sentiments of the tweets from Twitter tweets while the quantitative perspective will focus on the price prediction of cryptocurrencies. This is correlational research as it is based on the assumption that there is a correlation between the sentiments of social media and cryptocurrency prices.

Procedure

The data for this research has been gathered primarily from two sources i.e. Public comments from Twitter and actual cryptocurrency prices from Yahoo Finance from Jan 1, 2021 to April 30, 2021. The highest recorded price was recorded for a particular day and analyzed in dollars.

The tweets mentioning the hashtag for Dogecoin such as #DOGE and #Dogecoin was analyzed. Due to the recent limitation on the extraction of data from Twitter, the tweets can only be downloaded for the last 7 days. That is why the Scweet library ("Twitter streaming API," n.d.) is used to extract the tweets from Jan 1, 2021 to April 30, 2021. The idea behind the Scweet library is that it tries to save the text of each tweet on each page in the specified language and scrolls the page until it does not find any tweet of that specified word. . The tweets were filtered in the English language.

With the usage of the Twitter Scweet library, 8208 tweets were downloaded in the English language without having any images and any special characters along with time stamps. Data from social media is typically unstructured and noisy, and it must be cleaned up before it can be analyzed. Punctuations and stop words such as comma, question marks, numbers, special characters, emojis were removed. Tokenization was used to split the sentence, paragraph into smaller units such as words. Case folding was used to make the letters lower case. Lemmatization was implemented to remove affixes such as 'ly,' 'ed' etc. Lemmatizing was used to change the basic words into standard ones with the same meaning.(Irawaty, Andreswari, & Pramesti, 2020)

After the tweets are cleaned, sentiment analysis was used to get the sentiment score for the tweets. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool which is mainly used to analyze social media sentiments. The VADER sentiment analysis generates the positive, negative and neutral sentiment scores. The study uses the compound score which is the average of positive, negative, and neutral scores for each tweet. Later on, the tweets are averaged for each day to label the price. Figure.6. shows the sentiment score of tweets plotted against time which shows that most of the tweets are positive.

The final model will be based on the LSTM (Long short-term memory) to predict the price of the respective cryptocurrencies with and without sentiment score.(Hochreiter & Schmidhuber, 1997)

Measure

Sentiment analysis is a technique that employs natural language processing (NLP), text analysis, and a variety of other techniques to determine the writer's attitude toward a given subject from the text. VADER (Valence Aware Dictionary and Sentiment Reasoner3) is a sentiment analysis tool that is highly sensitive to text nuances like punctuation, capitalization, negation, and lexicon meaning amplification.(Hutto, 2014) It classifies text as positive, negative, or neutral, also measures the intensity, or polarity, of words used. It works best with social media data as the sentiment of a text can vary based on emotion.(Irawaty, Andreswari, & Pramesti, 2020)The compound sentiment score was used in the analysis.

LSTM is a modified variant of recurrent neural networks that makes it easier to recall past data. (Mittal, Dhiman, Singh, & Prakash, 2019) The deep learning LSTM neural networks solve the issue of vanishing gradients in recurrent neural networks (RNN) by replacing nodes in the RNN with memory cells and a gating mechanism. (Lahmiri & Bekiros, 2019) The gates manage the flow of information. They pass relevant information in a long chain of sequences to make predictions. LSTM has a variety of parameters, including batch size, unit count, and epochs. It produces different effects depending on the parameters. These parameters were decided based on different factors such as error, accuracy, precision. (Mittal, Dhiman, Singh, & Prakash, 2019)

Patel, Tanwar, Gupta, and Kumar (2020) stated that cryptocurrencies can be accurately predicted using neural network algorithms which lead to the usage of LSTM(Long short-term memory) networks for the prediction of cryptocurrency prices in this research. Zhu, Wang, Xu, and Li (2008) also suggested using neural networks with trading volumes as they enhance the prediction efficiency. Bollen and Zeng (2011) applied a self-organizing fuzzy neural network with Twitter and emotion as an input to forecast market shifts in the Dow Jones Industrial Average and obtained an accuracy of more than 85%.

The model performance will be measured using mean squared error (MSE) by comparing the actual cryptocurrency prices with the prediction from May 16, 2021 to May 31, 2021. The mean square error (MSE) is the sum of the squares of the difference between the actual and predicted values. We have not attempted to put more effort into flawless forecast since the cryptocurrency market is more volatile, and this might result in a loss to the user.

Analysis

Once the data is cleaned and the sentiment score is generated for the corresponding date, the LSTM model is applied. The first model makes a prediction based on the date and prices whereas the second model predicts based on the sentiment scores along with the past prices. The data was split into train and test data before the model application. In both the model's date was used as an index, the LSTM model was applied and the prediction was made for the next 15 days from May 16, 2021 to May 31, 2021. The data is normalized before it is applied to the model. The model mentioned below is designed based on the best mean square error value and the parameters such as epochs, unit count, and batch size are decided on that.

The first model was applied with only one open price of Dogecoin with the date as an index. Linear stack architecture is applied which means layers are used sequentially. So, the open prices are passed into a sequential architecture that consists of 4 LSTM hidden layers along with the input layer and output layer. All the pre-processed data is first passed through the input layer. The second layer consists of 50 units with 20% Dropout. At each update of the training phase, Dropout operates by changing the outgoing edges of hidden units (neurons that make up hidden layers) to 0. The second, third and fourth layer consists of 60, 80, 120 units with 30%, 40%, and 50% Dropout respectively. All the layers are activated by ReLU which stands for Rectified Linear Unit. The ReLU function is mainly used over other activation functions as it does not activate all the neurons at the same time. All the layers have set return sequences to access the hidden state output for each input time step. It is used when stacking LSTM layers so that the second LSTM layer has a three-dimensional sequence input. At last, there is an output dense layer with a single unit because the output is continuous.

The above model uses Adam optimizer for compilation. For deep learning model training, Adam is a substitute for stochastic gradient descent. Adam combines the finest features of the AdaGrad and RMSProp optimization methods to create an algorithm that can handle sparse gradients in noisy situations. The model was trained using epochs as 20 and batch size as 50 with a validation split of 10%. Model performance was measured using mean squared error (MSE).

The input sequence to the LSTM model should be properly defined for the proper functioning of the model. The input of the LSTM model is always a 3D array (batch size, time steps, and units). Batch size corresponds to the number of training samples to utilize in a single iteration. Time steps correspond to the time series data time difference between each data point and unit defines the number of features. The model uses 136 as batch size for training units and 15 as batch size as the test units. The number of features differs for both the models as the first model is based on the univariate LSTM model and the second model will be a multivariate model.

The second method uses the multivariate LSTM technique in which all the input features can be predicted. This model only predicts the Dogecoin price. The model used sequential architecture but with different parameters. This model has two features open price for Dogecoin and sentiment score generated with the sentiment analysis on tweets on that particular date. This model also consists of 4 LSTM hidden layers along with the input layer and output layer. All the pre-processed data is first passed through the input layer. All the hidden layers consist of 50 units with a 20% Dropout. All layers use the default activation function Tanh and return sequence Tanh is a hyperbolic tangent that is commonly utilized in recurrent neural networks for applications such as natural language processing and speech recognition. The model used Adam optimizer for compilation. The model was trained using epochs as 60 and batch size as 15. Model performance was measured using mean squared error (MSE). r cite_r("r-references.bib")' was used for all our analyses.

Results

The results show that the model with a sentiment score provides the least mean square error. The mean square error is 0.05 is for a model with a sentiment score and 0.073 without a sentiment score. This shows that the twitter sentiments predict the Dogecoin prices with less error. After the model is trained and complied with the Adam optimizer, it generates the predicted value. The predicted value along with the actual value can be used to calculate the mean squared error. Figure.1 Shows the training and validation loss of the LSTM model without sentiment score.

The input data was divided into training and testing set i.e. 90% of the data was in the training set and 10% of the data was in the testing set. When the data was fed to the model without a sentiment score the training loss was 0.1 after the 25 epochs it was reduced to 0.073. The batch size was 15 and the total number of epochs was 40. The testing data was imputed in place of validation data.

Initially, the validation loss was greater than the training loss which led to the overfitting of the model. Overfitting occurs when the model fits on the training set but it does not generalize for the new or unseen data. In this situation, the model does not improve and can be identified by metrics such as loss or accuracy. In this paper, overfitting was identified by loss. One way to reduce overfitting is to get more training data but that was not possible for this study as Twitter data can only be extracted for a limited period.

Regularization is another way to handle the overfitting it can be applied as the weight regularization to the model. For larger weights regularization tried to add a cost to the loss function which will make the model simpler and will allow only the relevant pattern to the model. L1 regularization will try to add a cost about the absolute value of the parameters while L2 regularization will add a cost with regards to the squared value of the parameters.

This model used Dropout to handle the overfitting which tried to modify the network itself by dropping the neurons during training in each iteration. At each update of the training phase, the outgoing edges of hidden units (neurons that make up hidden layers) are set to 0 at random. This made the validation loss follow the training loss and at the end of the 25 epoch, both the validation loss and training loss can be seen overlapping. This confirms that this is a fitted model.

The validation loss has some spikes at the end of the epochs that could be due to the small batch size and large learning rate. When advancing to the minimum of a loss function, the learning rate defines the step size at each iteration. This high learning rate might lead to a transition from one side of the hypothetical valley to the other while sinking to the global minimum. That is, the number of steps you take to update the weights is rather big. That should be eased up a little, resulting in a slower learning pace. The model uses the default learning rate of 0.01 for training the model. Batch size is the number of training samples utilized in one iteration. This study uses the batch size as 15 which corresponds to mini-batch gradient descent which sums the gradient over the mini-batch which further reduces the variance of the gradient. Mini-batch gradient descent aims to strike a compromise between stochastic gradient descent's resilience and batch gradient descent's efficiency.

A correlation was run to see whether there was a link between Twitter sentiments and Dogecoin price movements. The "Pearson R" and the p-value are the two major measures used to establish this. The Pearson R is a measure of the correlation's strength. Its value is between -1 and 1. A positive number indicates that the two variables are positively linked, or that rise in one variable produces an increase in the other (this is a correlation, not causation, so we can't say that one variable causes the other to change, only that there is a relationship). A negative value, on the other hand, indicates that the two variables are negatively linked, or that an increase in one variable's value is associated with a drop in the other. The p-value indicates how likely these correlation measurements would have been discovered by chance. The Pearson R of the correlation is 0.516 with a p-value of 0.020.

Figure.2 Shows the training and validation loss of the LSTM model with a sentiment score.The input data was divided into training and testing set i.e. 90% of the data was in the training set and 10% of the data was in the testing set. When the data was fed to the model with a sentiment score the training loss was 0.1 after the 20 epochs it was reduced to 0.05. The batch size was 50 and the total number of epochs was 60. The testing data was used as validation data while compilation. This model also uses Dropout to reduce overfitting. There are no spikes in the validation loss as in the previous model. The validation loss overlaps the training loss after the 20 epochs. Early stopping was used to stop the model when the best model metrics are attained. Early stopping is stopping early during compilation when there is no improvement on the validation set. Early stopping does not lead to overfitting and also reduces training time. Due to the least mean square error, overlapping training, and validation losses this model can be considered as the best model to predict the Dogecoin prices.

Figure.3 shows the prediction of Dogecoin Prices from 16 May 2021 to 31 May 2021 for a window of 15 days without a sentiment score. The difference in the value of actual and predicted prices can be in the plot. The actual Dogecoin prices are higher as compared to the predicted price. As the model is based on LSTM (which uses the past behavior of series to predict the future pattern) is noticeable. Figure.4 shows the prediction of Dogecoin Prices from 16 May 2021 to 31 May 2021 for a window of 15 days with a sentiment score. The difference in the value of the actual and predicted price is negligible. The predicted prices follow the trend of actual prices. The actual Dogecoin prices are lower as compared to the predicted price. Figure.5 shows the prediction of Dogecoin Prices from 1 Jan 2021 to 30 Jun 2021 for a window of six months with a sentiment score.

Discussion

This study presents significant evidence that Dogecoin prices are affected by twitter sentiments. Many previous research papers have already shown this with different cryptocurrencies such as Bitcoin, Ethereum. In this paper, first, the tweets were extracted for 5 months and after cleaning them VADER sentiment analysis was applied. The sentiment analysis provides the positive, negative, neutral sentiments. Compound sentiments were used as the mean of positive, negative, neutral sentiments. The correlation between the compound sentiment score and the Dogecoin was significant. Two different models were used to predict the prices with and without sentiment scores using the LSTM network. The price predicted with sentiment score has less mean square error.

The study found that the Twitter sentiments affect the Dogecoin prices as we found a positive correlation between the Twitter sentiments and Dogecoin price. Although most of the tweets seem to be positive even when the price is decreasing, this might be due to the fact people might be tweeting even if the prices go down. Another important factor to consider is the Twitter bots which were not considered in this study. In previous studies, there was a huge presence of Twitter bots and they might try to affect the price when the price goes down. This suggests that tweet volume, rather than sentiment, is a preferable indicator to use because the number of individuals talking about cryptocurrencies on Twitter may change with market values.

Cryptocurrency prices have a wide spectrum of volatility. The prices resemble a random walk process in that they are time-independent to some extent. The way people think about a currency has a big impact on its value. Prices are likely to be higher if the currency is more popular. Pang, Sundararaj, and Ren (2019) stated that there is a relationship between Bitcoin prices and social sentiment data, which is recommended for predicting the prices for shorter intervals of time (i.e. in minutes) in their future study section.

The limitation of this study is that it only investigates a single cryptocurrency. These results cannot be generalized to other cryptocurrencies. To understand the nature of other cryptocurrencies analysis should be performed separately as per their nature. This study was limited to a short period and data for training was not sufficient as there was not much price movement before this year in the price of Dogecoin. If other cryptocurrencies like Bitcoin were considered in this analysis then there would have been more data for training. The different pricing indicators like 'high,' 'close,' 'low' along with volume can be considered for the price prediction. The market volume seems to have a high correlation with cryptocurrencies which should be analyzed.

Conclusion

Due to cryptocurrencies' dramatic value increases and losses, their market has recently attracted a lot of interest. While the potential of cryptocurrencies extends far beyond their values, this study looked at how well public Twitter sentiments may be utilized to predict Dogecoin prices. This study was able to find the effect of Twitter sentiments on Dogecoin prices. The correlation between the sentiment score and prices was significant; the study predicted the Dogecoin prices with Twitter sentiments more accurately. The study also focuses on the application of the latest deep learning techniques in the analysis including LSTM, Dropout, Early stopping, Regularization to get more accurate results.

Cryptocurrencies are a new and unexplored research subject with a plethora of potential study topics. By implementing a robust cryptocurrency-specific lexicon-based sentiment analysis approach along with the latest neural network techniques, the study was able to forecast the future price of Dogecoin. By offering a literature overview on the issue, a cryptocurrency-specific analysis tool, and study beyond the boundaries of popular cryptocurrencies such as Bitcoin, Ethereum, this work has contributed to the small quantity of cryptocurrency research.

As suggestions for future research, one could apply this research to a larger set of cryptocurrencies, extend the period of observation, and experiment with various levels of granularity. Another topic for future research would be to apply ARIMA, AR models, and other supervised machine learning approaches. One possibility is to investigate ways to more precisely identify Twitter bots and then assess their influence on Twitter sentiment and/or cryptocurrency pricing. Finally, one might assess the research's repeatability by attempting to reproduce the findings of applying the findings to forecast, for example, price returns. These insights can be utilized to build a trading strategy in a more advanced environment.

Some studies have used YouTube comments and video transcripts for the prediction of cryptocurrency prices and their results we considerable. Jay et al. (2020) forecasted the price of Cryptocurrencies using a stochastic neural network model, and the findings show that the presented hypothesis was not only correct but also successful in predicting market volatility. Smuts (2019), which claimed that Google trends and Telegram data may forecast cryptocurrency values, but didn't go so far as to show that the pattern is consistent over time.

Although cryptocurrencies are growing in popularity, they are still prohibited in several nations. There is no regulating authority in place to keep them in check. Even in areas where it is legal, it comes with a slew of legal ramifications. People's perceptions, and hence cryptocurrency values, are influenced by these variables.

References

Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan. (2018). *Cryptocurrency price prediction using tweet volumes and sentiment analysis* (Vol. 1, p. 21). SMU Data Science Review. Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss3/1/

Bitcoin hits another record. (2021). https://doi.org/https://www.cnbc.com/2021/02/12/how-much-youd-have-today-if-you-invested-100-in-bitcoin-in-2009.html

Bollen, M., J., & Zeng, X.-J. (2011). Twitter mood predicts. *Journal of Computer Science*, *2*, 1–8.

Chan, S., Chu, J., Nadarajah, S., & Osterrieder, J. (2017). A statistical analysis of cryptocurrencies. *Journal of Risk and Financial Management*, *10*(2). Retrieved from https://www.mdpi.com/1911-8074/10/2/12

Cheuque Cerda, G., & L. Reutter, J. (2019). Bitcoin price prediction through opinion mining (pp. 755–762). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3308560.3316454

Chu, J., Chan, S., Nadarajah, S., & Osterrieder, J. (2017). GARCH modelling of cryptocurrencies. *Journal of Risk and Financial Management*, *10*(4). https://doi.org/10.3390/jrfm10040017

Cohen, G. (2020). Forecasting bitcoin trends using algorithmic learning systems. *Entropy*, *22*(8), 838. https://doi.org/10.3390/e22080838

Dabas, C., Kaur, P., Gulati, N., & Tilak, M. (2019). Analysis of comments on youtube videos using hadoop. In *2019 fifth international conference on image information processing (ICIIP)* (pp. 353–358). https://doi.org/10.1109/ICIIP47207.2019.8985907

Gupta, R., Tanwar, S., Al-Turjman, F., Italiya, P., Nauman, A., & Kim, S. W. (2020). Smart contract privacy protection using AI in cyber-physical systems: Tools, techniques and challenges. *IEEE Access*, *8*, 24746–24772. https://doi.org/10.1109/ACCESS.2020.2970576

Heukelom, F. (2009). Kahneman and tversky and the making of behavioral economics.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hutto, E. E., C. J. & Gilbert. (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.

Irawaty, I., Andreswari, R., & Pramesti, D. (2020). Development of youtube sentiment analysis application using k-nearest neighbors (nokia case study). In *2020 3rd international conference on information and communications technology (ICOIACT)* (pp. 39–44). https://doi.org/10.1109/ICOIACT50329.2020.9332151

Jain, A., Singh, R., & Vatsa, M. (2018). On detecting GANs and retouching based synthetic alterations. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)* (pp. 1–7). https://doi.org/10.1109/BTAS.2018.8698545

Jay, P., Kalariya, V., Parmar, P., Tanwar, S., Kumar, N., & Alazab, M. (2020). Stochastic neural networks for cryptocurrency price prediction. *IEEE Access*, *8*, 82804–82818. https://doi.org/10.1109/ACCESS.2020.2990659

Karalevičius, V. (2017). Using sentiment analysis to predict interday bitcoin price movements. *The Journal of Risk Finance*, *19*, 00–00. https://doi.org/10.1108/JRF-06-2017-0092

Knežević, A., Babić, T., & Musa, Z. (2020). Cryptocurrency as the currency of the future: A case study among algebra university college students. In *2020 43rd international convention on information, communication and electronic technology (MIPRO)* (pp. 782–787). https://doi.org/10.23919/MIPRO48935.2020.9245265

Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, *65*, 101188. https://doi.org/https://doi.org/10.1016/j.intfin.2020.101188

Lahmiri, S., & Bekiros, S. (2019). Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos, Solitons & Fractals*, *118*, 35–40. https://doi.org/https://doi.org/10.1016/j.chaos.2018.11.014

Li, X., & Wang, C. A. (2017). The technology and economic determinants of cryptocurrency exchange rates: The case of bitcoin. *Decision Support Systems*, *95*, 49–60. https://doi.org/https://doi.org/10.1016/j.dss.2016.12.001

Mittal, A., Dhiman, V., Singh, A., & Prakash, C. (2019). Short-term bitcoin price fluctuation prediction using social media and web search data. In *2019 twelfth international conference on contemporary computing (IC3)* (pp. 1–6). https://doi.org/10.1109/IC3.2019.8844899

Mohanty, P., Patel, D., Patel, P., & Roy, S. (2018). Predicting fluctuations in cryptocurrencies' price using users' comments and real-time prices. In *2018 7th international conference on reliability, infocom technologies and optimization (trends and future directions) (ICRITO)* (pp. 477–482). https://doi.org/10.1109/ICRITO.2018.8748792

Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S. (2016). *Bitcoin and cryptocurrency technologies: A comprehensive introduction*. Princeton

University Press.

Pang, Y., Sundararaj, G., & Ren, J. (2019). Cryptocurrency price prediction using time series and social sentiment data (pp. 35–41). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3365109.3368785

Patel, M. M., Tanwar, S., Gupta, R., & Kumar, N. (2020). A deep learning-based cryptocurrency price prediction scheme for financial institutions. *Journal of Information Security and Applications*, *55*, 102583. https://doi.org/https://doi.org/10.1016/j.jisa.2020.102583

Rehman, M. U., & Apergis, N. (2019). Determining the predictive power between cryptocurrencies and real time commodity futures: Evidence from quantile causality tests. *Resources Policy*, *61*, 603–616. https://doi.org/https://doi.org/10.1016/j.resourpol.2018.08.015

Rüßmann, L., M. (2003). Industry 4.0: The future of productivity and growth in manufacturing industries, *1*, 54–89. Retrieved from https://www.scopus.com/record/display.uri

Satoshi Nakamoto. (2008). *Bitcoin: A peer-to-peer electronic cash system.* Bitcoin.org. Retrieved from https://bitcoin.org/bitcoin.pdf

Sin, E., & Wang, L. (2017). Bitcoin price prediction using ensembles of neural networks. In *2017 13th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)* (pp. 666–671). https://doi.org/10.1109/FSKD.2017.8393351

Smuts, N. (2019). What drives cryptocurrency prices? An investigation of google trends and telegram sentiment, *46*(3), 131–134. https://doi.org/10.1145/3308897.3308955

Teker, D., Teker, S., & Ozyesil, M. (2019). Determinants of cryptocurrency price movements.

The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with support vector regression. (2018). *Expert Systems with Applications*, *97*, 177–192. https://doi.org/https://doi.org/10.1016/j.eswa.2017.12.004

Twitter streaming API. (n.d.). Retrieved from https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data

Urquhart, A. (2017). Price clustering in bitcoin. *Economics Letters*, *159*, 145–148. https://doi.org/https://doi.org/10.1016/j.econlet.2017.07.035

Vidal-Tomás, D., Ibáñez, A. M., & Farinós, J. E. (2021). The effect of the launch of bitcoin futures on the cryptocurrency market: An economic efficiency approach. *Mathematics*, *9*(4). https://doi.org/10.3390/math9040413

Zhu, X., Wang, H., Xu, L., & Li, H. (2008). Predicting stock index increments by neural networks: The role of trading volume under different horizons. *Expert Systems with Applications*, *34*(4), 3043–3054. https://doi.org/https://doi.org/10.1016/j.eswa.2007.06.023

Appendix A

Figures

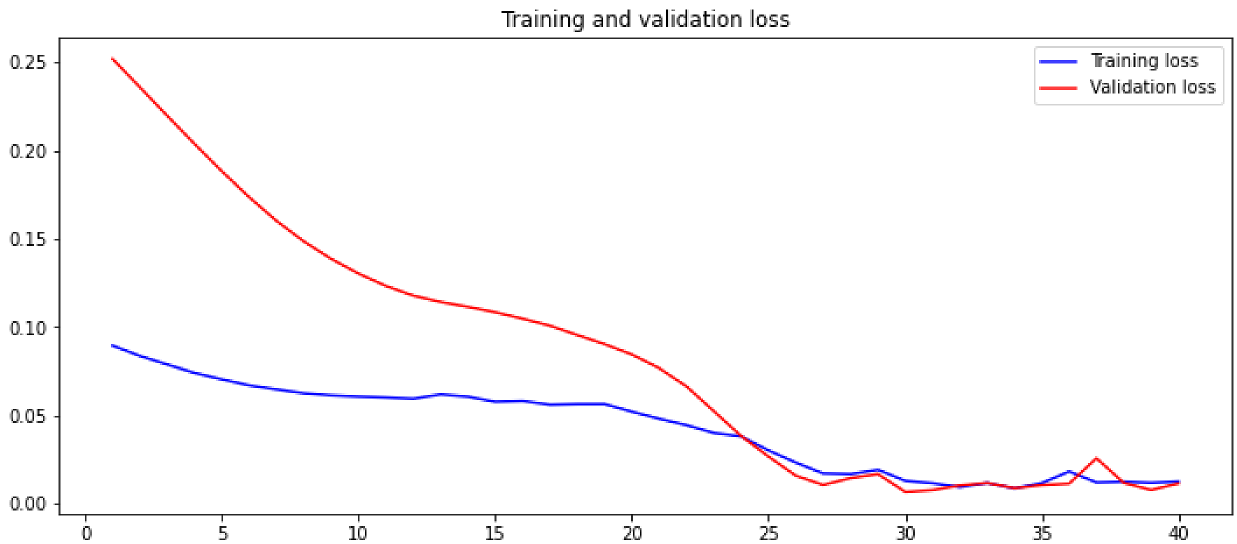

Figure A1. Figure.1. Training and validation loss for model without sentiment score

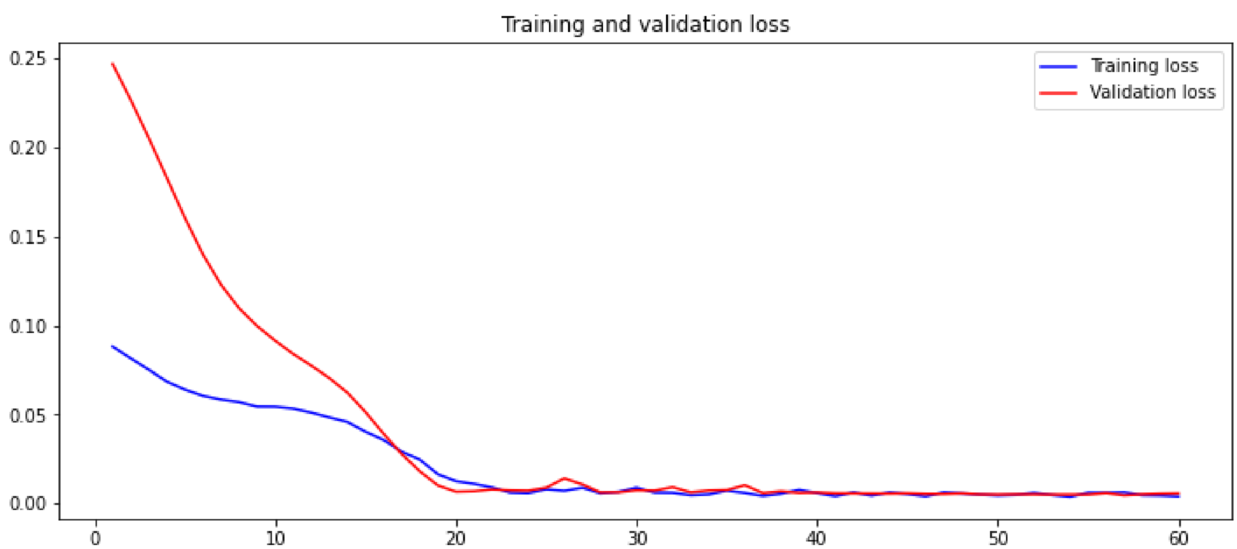

Figure A2. Figure.2. Training and validation loss for model with sentiment score

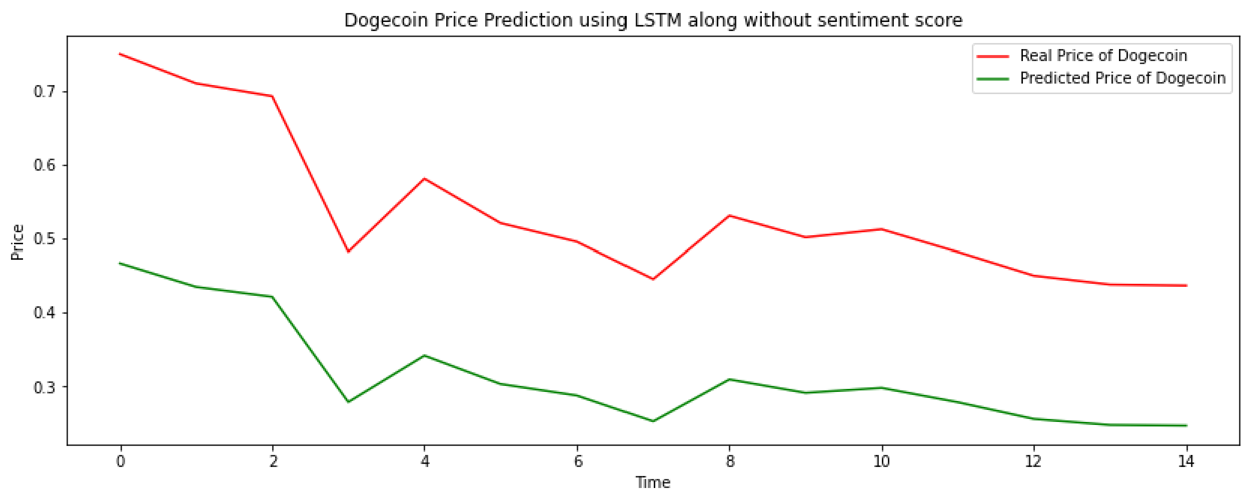

Figure A3. Figure.3 Prediction of the Dogecoin prices without sentiment analysis

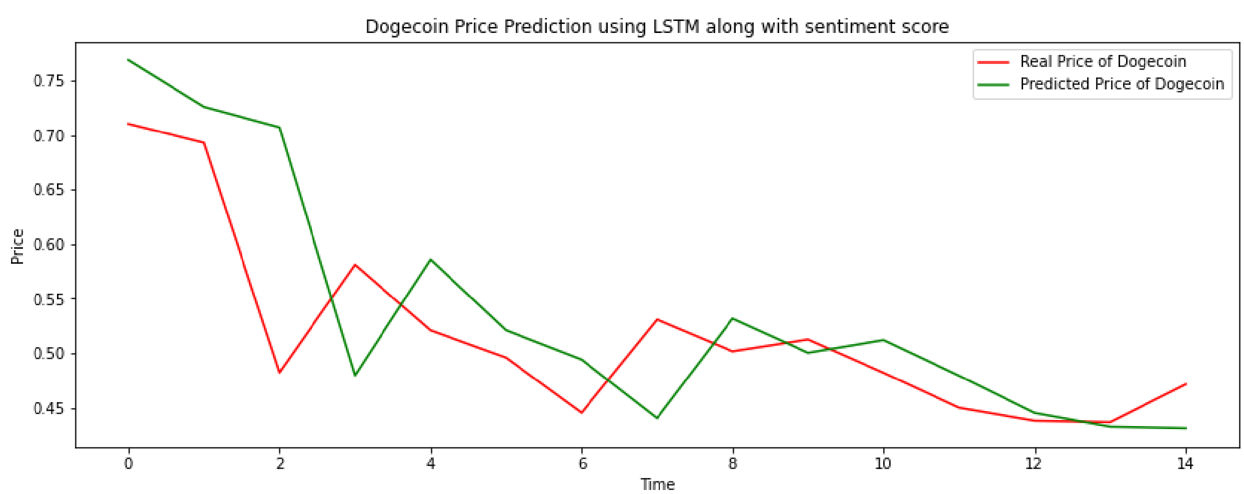

Figure A4. Figure.4. Prediction of the Dogecoin prices with sentiment analysis

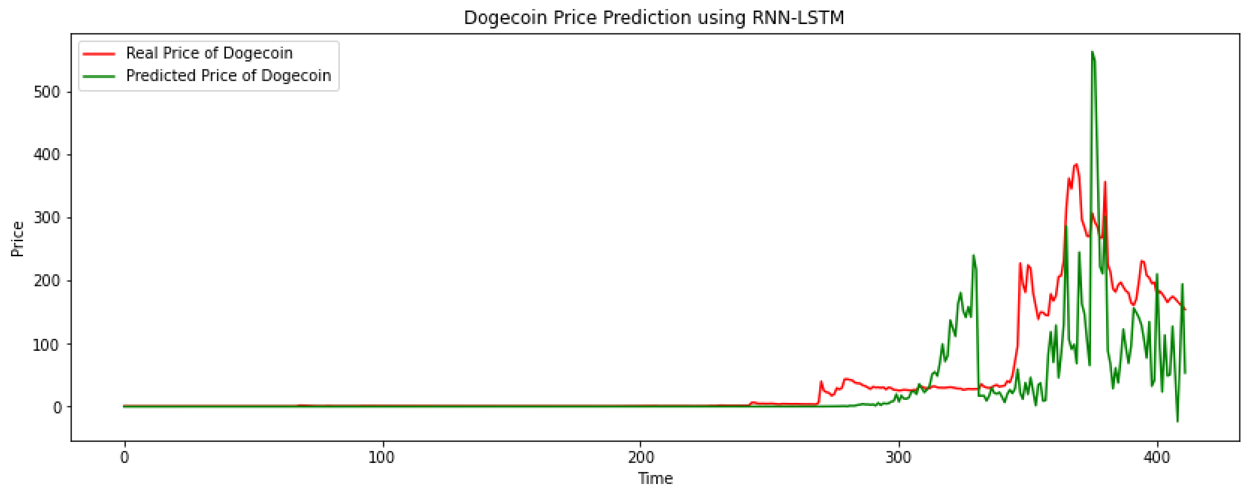

Figure A5. Figure.5. Prediction of the Dogecoin prices with Sentiment Analysis from Jan 1,2021 to June 30,2021

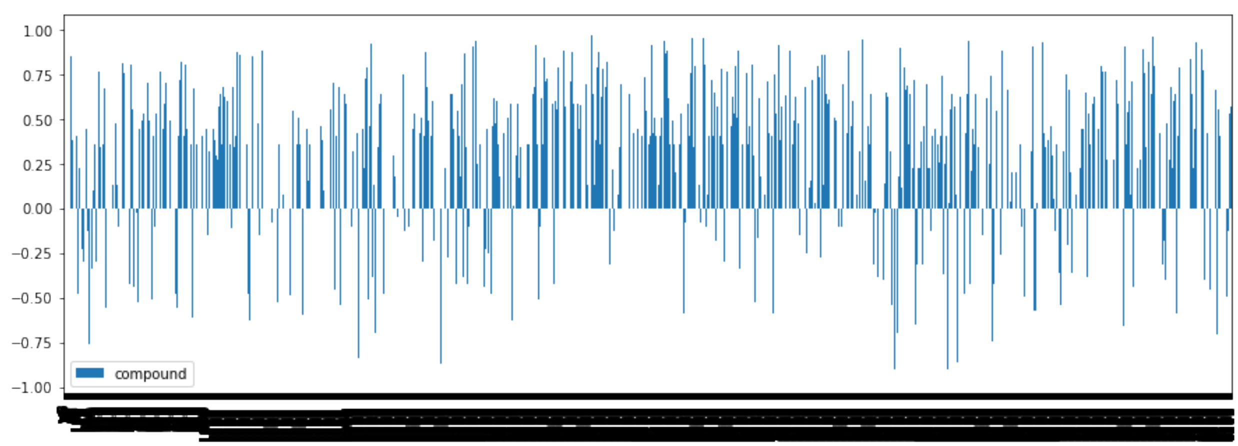

Figure A6. Figure.6. Sentiment score of tweets

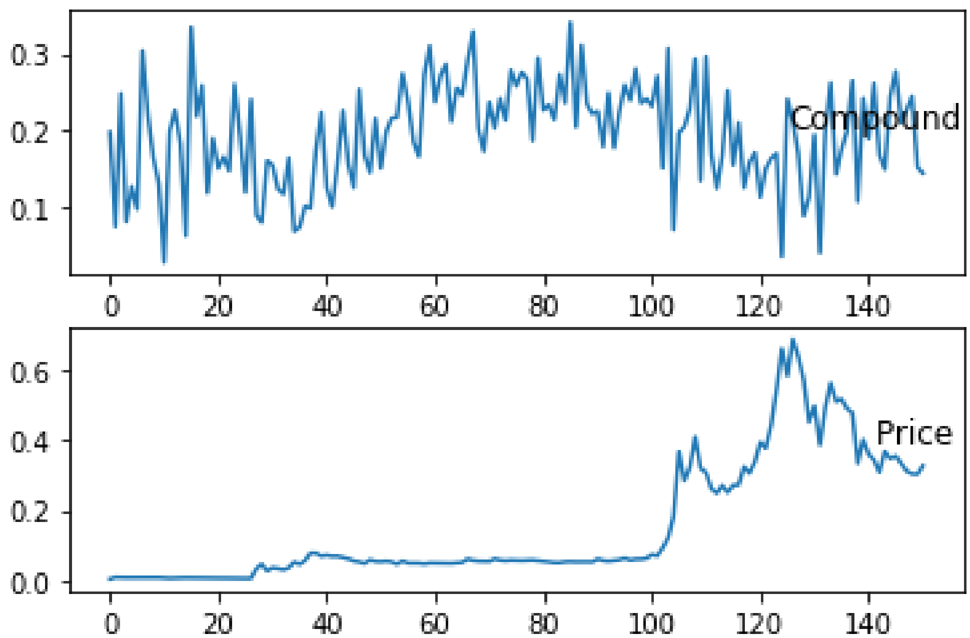

Figure A7. Figure.7. Compound Sentiment score and Dogecoin price

Appendix B

Python code