ETHEREUM PRICE PREDICTION USING NUMBER OF TRANSACTIONS AND VOLUME OF TWEETS

AREND GROOT BLEUMINK

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2026189

COMMITTEE

dr. Giacomo Spigler
dr. Merel Jung

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 19, 2022

ACKNOWLEDGMENTS

Some room for acknowledgements.

ETHEREUM PRICE PREDICTION USING NUMBER OF TRANSACTIONS AND VOLUME OF TWEETS

AREND GROOT BLEUMINK

Abstract

In the recent years, cryptocurrencies and blockchain technology have gained mass attention from both a technical as well as a speculative aspect. A lot of media attention has gone towards the losses and gains from the cryptocurrency market, while companies are looking into how they apply blockchain within their industry. The largest cryptocurrency and blockchain projects include Bitcoin and Ethereum, with a respectable combined Market Value of over 781 Billion U.S. Dollars (CoinMarketCap, 2022) amongst Bitcoin and Ethereum. While Bitcoin is more known amongst the public. Ethereum has established itself as a solid blockchain project with a utility application and platform and is the second-largest cryptocurrency. A large community of blockchain and cryptocurrency enthusiasts can be found on Twitter, including Ethereum developers, users, and speculators. One multivariate and three univariate LSTM models were developed to answer the question if an LSTM model is able to correctly predict the price movement of Ethereum based on Tweet Volume and Number of Transactions. One baseline model was developed to predict the future price change based on the previous price changes, while the other univariate model tested each predictor separately. The research found that the baseline model was not a good indicator for future price changes. Furthermore, it showed that even though the other models are not accurate, they do provide a better coefficient of determination than the baseline model. The research showed that using the Transaction count and Volume of Tweets as the input of our proposed LSTM model cannot accurately predict the future day-over-day percentage change.

1 INTRODUCTION

This research paper aims to investigate if it is possible to predict the price change of Ethereum using a blockchain statistic and the Volume of Twitter tweets that contain the Ethereum hashtag. By May 2022, the total cryptocurrency market capitalization (Market Cap) will be over 1.5 Trillion U.S. Dollar (CoinMarketCap, 2022). The two largest cryptocurrencies are Bitcoin, with a Market Cap of $640 Billion, and Ethereum, with a Market Cap of around $296 Billion. Even though the blockchain space is only Fourteen years old, it gained a lot of traction in the last years. Cryptocurrency and blockchain are often used interchangeably, but blockchain refers to the actual technology behind the cryptocurrencies, and cryptocurrency refers to the digital currency or medium of exchange.

1.1 *Blockchain and Bitcoin*

In 2008 an unknown person or group of people using the name Satoshi Nakamoto introduced the world to Bitcoin. Bitcoin was introduced as a peer-to-peer version of electronic cash that would allow online payments to be sent without going through a financial institution. (Nakamoto, 2009) Instead of using trusted parties, such as financial institutions, Bitcoin uses cryptography to establish a transaction. All transactions on the network are non-reversible, which can prevent fraud by buyers as they can not dispute payments. On the other end, buyers could be negatively impacted when ill intentions arise. However, by using escrow mechanics, this could be avoided. In the context of Bitcoin, two terms are used Bitcoin, which represents the Bitcoin blockchain network, and BTC, which is the medium of exchange, also called cryptocurrency. Each Bitcoin client generates so-called public and private keys, which provide ownership over the Bitcoin client, also called a wallet. The public key, often referred to as the bitcoin public address can be shared and is used as the receiving address. The private key is used to sign transactions and is the ultimate key to the BTC on a wallet. All transactions are stored on the blockchain, which can be compared to a distributed ledger. The blockchain is powered by nodes that enable transactions to go through. These nodes all house the same ledger. Before a transaction can go through, the node verifies with other nodes that the sending party has enough BTC to send to the receiving address. A consensus among the nodes has to be reached before it will perform the transaction. These transactions are bundled into a block, which is added as a whole to the blockchain after acceptance. The nodes in the Bitcoin network are then up-to-date with the latest mutations, after which the next block can be constructed.

1.2 *Ethereum and Vitalik Buterin*

In 2014 Vitalik Buterin published the paper Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform. Ethereum was launched with the intent of merging the concepts of scripting, alt-coins, and on-chain meta-protocols and allowing developers to create consensus-based applications while also allowing for scalability and interoperability at the same time. (Buterin, 2013) He developed Ethereum as he found the The bitcoin blockchain is too limited. He came up with the idea to implement applications on the blockchain with a built-in programming language (*Vitalik Buterin created one of the world's largest cryptocurrencies in his early twenties — here's how he did it and why*, 2019). He compares the Ethereum to a plot key calculator and a smartphone, where the calculator does one thing well, but people want to do multiple things instead. If someone has a smartphone with an app instead, that person can, for example, also play music on that smartphone. So instead of solely focusing on acting as a crypto payment system, he envisioned a platform where all types of applications could be developed. Ethereum introduced a way of writing smart contracts using a built-in Turing-complete programming language. A smart contract is a consensus-based application housed on the Ethereum platform, which allows for writing contracts with its own arbitrary rules for ownership, transaction formats, and state transition functions. In other words, it provides a way to create contracts, or a rule-based system, which executes specific actions when certain arbitrary rules are met. An example of a smart contract is a contract that executes when a company trespassed a certain profit value in a quarter. If the threshold trespasses, the board of directors automatically gets a set percentage of the total profit; if not, they will not receive the bonus. An essential condition of the execution is the availability of the required data, but if that is present, the smart contract would be able to execute on its own. Aside from monetary smart contracts, a wide variety of other types are available, including cloud computing, smart multi-signature escrow, decentralized marketplaces, and more. These smart contracts live on the blockchain, and the executions of the contracts are saved as transactions on the blockchain. New contract deployments are also considered transactions, as are regular transactions between addresses. Each transaction has a fee associated with it, which is dependent on the number of transactions, which is called a "gas" fee.

1.3 *Related work*

Since the introduction of Bitcoin and the gained interest in other cryptocurrencies, Twitter has housed a large community of blockchain/cryptocurrency enthusiasts. Mai, Shan, Bai, Wang, and Chiang (2018) found a similar effect of Twitter sentiment on predicting daily up and down changes with an accuracy of 86.7%. Although a significant number of researchers focus on predicting stock prices and Bitcoin, only a small number of research papers try to predict the price movement of Ethereum using blockchain statistics. Ethereum is lesser-known but is the second-largest cryptocurrency, with a market capitalization of over 780 billion U.S. Dollars at the time of writing (May 8, 2022). It is interesting to research if the Number of Transactions or Tweet volume could help predict the price change of Ethereum.

1.4 *Research Questions*

In this research, a Long Short Term Memory (LSTM) model will be used, which is a recurrent neural network capable of using persistent information. An LSTM model is very well suited for predictions based on time-series data, of which the Ethereum prices per day is an example. This research focuses on using the number of transactions and the number of Twitter messages, so-called Tweets, to predict the price movement of Ethereum using an LSTM model. Since it is not possible to segment the transactions into contract execution/initiations and transfers, the total number of transactions is considered.

RQ1 *To what extent is it possible to predict the short-term price of Ethereum based on the Volume of Tweets, and the Number of Transactions on the blockchain using a LSTM model?*

RQ2 *To what extent is it possible to predict the short-term price of Ethereum based on the Volume of Tweets using a LSTM model?*

RQ3 *To what extent is it possible to predict the short-term price of Ethereum based on the Number of Transactions using a LSTM model?*

2 RELATED WORK

Prediction of commodities, stocks, and currencies has been an often researched topic within academics as well as within trading companies and businesses. Many research papers have addressed the question if it is possible to predict these markets. The price of a cryptocurrency is determined by a supply and demand equilibrium, much like commodities and traditional currencies. A previous study (Nguyen, Shirai, & Velcin, 2015) has found that sentiment analysis can predict the price movement with more than 60% accuracy for a few stocks, with an overall accuracy of 54.41% in its prediction. However, the price prediction was limited to up and down movements instead of a price prediction in itself. The model used to predict the price movement was based on the Support Vector Machine with a linear kernel, while the sentiment data was acquired on Yahoo Finance Message Board.

As opposed to using the sentiment to predict stock prices, Nelson, Pereira, and De Oliveira (2017) used 175 Technical Analysis indicators using the TA-Lib library to predict the movement of several Brazilian stocks. Instead of the Support Vector Machine model, the researchers used an LSTM model with 190 features and a Sigmoid activation. The results showed an average of 55.9% accuracy in predicting the price movement, which was higher than the baseline they were comparing it to. The baselines used included Multi-Layer Perceptron, Random Forest, and pseudo-random models, which were trained and evaluated with the same input data. Since the indicators were not tested individually, it is unclear which of the Technical Analysis indicators were the best predictors of the price movement.

Aside from related research on the stock market, related work has been done on the cryptocurrency market as well. Research done by Kwon, Kim, Heo, Kim, and Han (2019) collected the prices of seven cryptocurrencies: BTC (Bitcoin), ETH (Ethereum), XRP (Ripple), BCH (Bitcoin Cash), LTC (Litecoin), DASH (Dash), and ETC (Ethereum Classic) on a ten-minute interval from June 9, 2017, to May 8, 2018. Five features were captured as part of the data collection: open price, close price, high price, low price, and volume every 10 minutes. Similar to the research of Kwon et al. (2019), the researchers used an LSTM model to classify the binary cryptocurrency price trend. Overall, the performance improved by about 7% compared to the baseline gradient boosting model. The highest F1 score of the ETH prediction found was around 0.65. Furthermore, a paper focused on predicting the BTC price using Twitter sentiment and an LSTM model (Pant, Neupane, Poudel, Pokhrel, & Lama, 2018) also found a moderate correlation between the sentiment and the Bitcoin price and had an accuracy of 77.62% in its predictions.

Little to no research has been done on the correlation between Tweet volume and transaction volume and its relation to the ETH price. One research paper focused on Bitcoin found Tweet Volume in combination with Google Trends as a predictor for the BTC price Abraham, Higdon, Nelson, and Ibarra (2018). The researchers found that rather than the sentiment of Tweets, the Google Trends and Tweet Volume were highly correlated with the BTC price. It excluded the sentiment of Tweets due to consistently high day-over-day price variability, while the sentiment stayed the same. The researchers used a linear model to predict the price using the Google Trends and Tweet volume data. Our research aims to use the Tweet volume data with an LSTM model since the LSTM model might be able to better fit the data due to the possibility of using looking back at older data. One research paper that did take into account various Ethereum blockchain statistics (Kim, Bock, & Lee, 2021) found that macro-economic factors and generic Blockchain information of Ethereum contributed to the accurate predictions of the ETH prices. The general blockchain information used in the research included: Transaction Volume, Transaction Count, Generated Count, Active Addresses, and Block Size, among others. Since the researchers did not specify the outcome of each of the statistics in regards to the price, it is not clear if the number of transactions by itself proved a credible predictor. So it is unclear what influence the Number of Transactions had on the results. Overall a combination of the macro-economic factors and blockchain information predicted the ETH price with an accuracy of 95.2%.

A research paper that used the price movement of Bitcoin to predict the price of Ethereum found that there was no improvement compared to the baseline model (Caldegren, 2018). Similar research in the stock market showed a significant influence of related stocks was found (Li, Y, & Y, 2018). The researchers used a Multi-Input LSTM model and major stocks under the CSI-300 index, which is comprised of the top 300 stocks traded on the Shanghai Stock Exchange and the Shenzhen Stock Exchange. The model's performance deteriorated when the number of related stocks used was 15 and 20.

3 RELEVANCE

Given the interest in predicting prices of digital currencies and assets, it is interesting to research if the Volume of Tweets and, or Number of Transactions could be viewed as an indicator to predict price movement. The market cap of Ethereum is bigger than public companies such as Since the market cap of Ethereum is larger than public companies such as Coca-Cola, Alibaba, McDonalds, and ASML, among others (*Companies ranked by Market Cap - CompaniesMarketCap.com*, 2022), combined. This makes it an attractive digital currency to track and attempt to predict price changes. A large number of research papers are focused on predicting stock prices with different types of algorithms and models, but aside from Bitcoin, a lot less research was conducted into other cryptocurrencies and blockchain projects. Since Twitter houses a large community of cryptocurrency enthusiasts, it is interesting to see if the Tweet volume is correlated with the Ethereum price change. Additionally, since Ethereum is quite unique in its blockchain technology, the transaction count is interesting to include as well. Even though non-relevant transactions are counted as well, which includes new

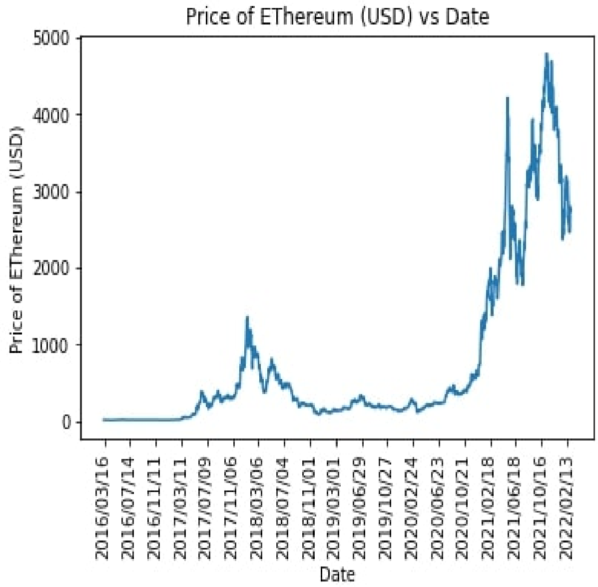

Figure 1: Historical Prices of ETH starting from March 2016

smart contracts deployed, and new smart contract interactions, among other transactions, an influx of transactions could be related to a price change in Ethereum. Since only a handful of research papers included the Tweet Volume and Number of Transactions, and no research was found that combined both variables, it is worth researching the extent to which these variables are correlated with the price of Ethereum.

4 METHODS

4.1 *Data set*

The data set used in this research paper was extracted from Bitinfocharts.com by using Python and Python Requests. The data on the Price, Number of Tweets, and Number of Transactions were all on separate pages and were extracted from JavaScript code in the HTML pages. The data was stored inside JavaScript as part of the source for the data inside the chart, which shows the regression of these metrics. This data source has been gathering information for many years on various blockchains and cryptocurrencies, including tweets containing a particular cryptocurrency hashtag. Other data sources found either used the Ethereum keyword or did not have the same size of historical data. In this research, the hashtag Ethereum was used instead of the keyword Ethereum as more historical data was available.

Data (per 24h)	Number of Rows	Missing Values	Min	Max
Number of Tweets	2176	54	543	62007
Transaction Count	2405	8	1329	1716600
Day over Day	2176	54	-25.72	33.45

Table 1: The descriptive information about the datasets used in this research. The Number of Tweets and Transaction Count are cummalative of 24 hours, Ethereum Price is closing USD price.

The date and value were then extracted from the Bitinfocharts.com using Regex, and then each dataset was saved to a CSV file for later use. After data exploration of each dataset, various null values were found. These null values were filled in using the average of the preceding and following values; in some cases, multiple consecutive null values were found. In those cases, the next available value was used instead. The average was chosen since it does not change the overall statistics of the data, and no abnormal data points were introduced, which might have introduced noise in the LSTM model. All values were then transformed into integers, except for the ETH price, since it was recorded with decimal values, and in the early beginning of Ethereum, the price was below 1 USD.

Additionally, the data was further cleaned to only include the data on the metrics after March 16, 2016, since that was the moment when the Number of Tweets started to get saved. One alternative method for acquiring the Twitter data was by using their API. However, this API was only available for Master/Ph.D. students and businesses, so, therefore, the Number of Transactions and ETH price before that day were omitted. After that, the three datasets, stored as a CSV file, were combined into the final dataset.

4.2 *Further Pre-processing*

Since our features highly vary in units and range, each feature was scaled using a MinMax Scaler from the Sci-Kit Learn package. The MinMax Scaler scales each feature to a value between 0 and 1 by using the minimum and maximum values and normalizing the features. The minimum is subtracted from all values, after which the values are divided by the difference between the minimum and maximum values. The number of Tweets and the Number of Transactions were normalized using the above method.

Since predicting the price in USD can prove difficult due to uneven distribution, a new column was added to the data frame, which contains the Day over Day price change in percentage. Using the Day over Day values allows the LSTM model to use the input variables to model a relative change instead of an absolute change of value.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

4.3 *LSTM Model*

Long Short-Term models are used in a variety of areas, from Speech Recognition, Image Captioning, and Music Generation to Sentiment Analysis and Time Series prediction, as in our research. LSTM can be applied wherever there is time-dependent information, and the data consists of sequences, commonly referred to as Time Series data.

Conventional "Back-Propagation Through Time" or "Real-Time Recurrent Training" may lead to oscillating weights or takes a long time, or does not work at all (Hochreiter & Schmidhuber, 1997). In the paper, by Hochreiter and Schmidhuber (1997) the "Long Short-Term Memory" model was introduced as a novel recurrent network architecture combined with a gradient-based learning algorithm. The advantages of using an LSTM model include its ability to handle noise, distributed representations, and continuous values. LSTM works well over a broad range of parameters, including the learning rate, input gate bias, and output gate bias. Another important advantage is that the algorithm updates complexity per weight and time step is essentially the same as the "Back-Propagation Through Time" while not having the same disadvantages. The application of LSTM has been widely researched in the context of Time Series Prediction. In research on traffic prediction and user-location forecasting, multiple LSTM models were compared to a Support Vector Regression (SVR), an Autoregressive Integrated Moving Average (ARIMA), and Feedforward Neural Network (FFNN) model (Hua et al., 2019). Three variations of the LSTM model are implemented, including an adjusted LSTM model, called the RCLSTM model, where the neurons are randomly connected rather than fully connected, as is the case in an LSTM model. Additionally, two LSTM models were constructed, one with a memory cell size of 30 and the other with a memory size of 300. To evaluate the performance of the models, the Root Mean Square Error (RMSE) was applied. The results of the experiment showed that the LSTM with a memory size of 300 was able to get the best prediction with both the user-location forecasting and the traffic prediction.

A survey of Forex and Stock Price Prediction using Deep Learning (Hu, Zhao, & Khushi, 2021) studied different deep learning methods among selected articles. The articles were selected by keywords which included

"CNN stock/Forex", "LSTM stock/Forex", "Deep learning stock/Forex", "RNN stock/Forex", and "Reinforcement learning stock/Forex" between 2015 and 2021, in total 86 papers were reviewed. The deep learning models reviewed included Convolutional Neural networks, LSTM, Deep Neural networks, Recurrent Neural networks, reinforcement learning, and other deep learning methods. The survey investigates different effects of different deep learning methods on the stock and Forex market. Of the 86 papers, 44% of the papers used an LSTM model. The paper with the highest accuracy used an LSTM model, while also 3 of the 4 papers with the best accuracy also were based on an LSTM model. At the same time, not all papers that used LSTM had high accuracy scores. However, in general, the LSTM models performed exceptionally well; 92% of the papers had an accuracy of over 50%. Considering the stock market is more mature in research, and both the cryptocurrency and the Stock/Forex market contain Time-Series data, it shows that LSTMs are an excellent choice for price prediction.

4.4 *Implementation*

The loading, preprocessing, and actual implementation of the multivariate and univariate LSTM model were all done in Python 3.9.12. For the LSTM model, Keras (version 2.8) uses the LSTM layer combined with multiple Dense layers, a Dropout layer, and an EarlyStopper callback. The EarlyStopper callback stops training when the monitored metric has stopped improving. The monitored metric in the case of this research is the loss; additionally, the best weight during the training is saved and used in the model. Pandas (version 1.4.2) was used to load in and manipulate the dataset and coupled with Numpy (1.22.3). Matplotlib (3.5.1) was used for the graphs, and Scikit-learn (1.0.2) was used for the scaling and metric functions. Lastly, the Adam optimizer was used from the Tensorflow (2.8) package. The implementation was inspired by an article and accompanying code by Lendave (2021).

The LSTM model is constructed using a Sequential model, which turns the LSTM model into a Bidirectional LSTM model. This means the model trains two LSTMs on the input sequence instead of one LSTM (Brownlee, 2021). The first recurrent layer is duplicated, so there are two layers side-by-side providing the input sequence to the first layer.

Additionally, four Dense layers are introduced, which are deeply connected neural network layers. These layers enable every neuron in one layer to be connected to every neuron in the next layer. Afterward, a regularization technique, Dropout regularization, is applied. Dropout is a technique where randomly selected neurons are ignored during training

Table 2: The RMSE and MAE results of each of the models, combined with additional information about model.

		Scores	
Layers	Models	RMSE	MAE
200	**LSTM (base)** Dense (20) Dropout (0.25)	5.376	7.165
200	**LSTM (RQ1)** Dense (20) Dropout (0.25)	4.379	4.892
200	**LSTM (RQ2)** Dense (20) Dropout (0.25)	4.327	4.838
200	**LSTM (RQ3)** Dense (20) Dropout (0.25)	2.327	2.943
	Best RMSE and MAE	**2.327**	**2.943**

(Brownlee, 2020). As neurons are randomly dropped, other neurons have to compensate for that as the neurons can not learn every detail of the training set. This technique is used to prevent over-fitting during the training of the model. The Dropout value of 0.25 indicates that 25% of the neurons are randomly deactivated during the training.

For the first LSTM layer, 200 layers were used; for the second layer, 150 layers were used. To save the weights of the trained model, a Keras ModelCheckpoint callback was used, which was later used in the fitting of the model. During the fitting, 150 epochs are used with 20 steps per epoch. The parameters were chosen at random while taking into account the respective documentation and the parameters of related LSTM models.

4.5 *Evaluation criteria*

In total, four models are trained, which includes a baseline model and three models which are modeled after each of the research questions. The baseline model solely uses the previous day's change as the input and uses that to predict the following percentage change. The R^2 metric is used to measure the goodness-of-fit of the models compared to the actual data and is measured as a value between 0 and 1. It is possible to have a value below 0, which indicates that the prediction is worse than using the mean value. The better the fit of the model, the closer the R^2 value will be to 1.

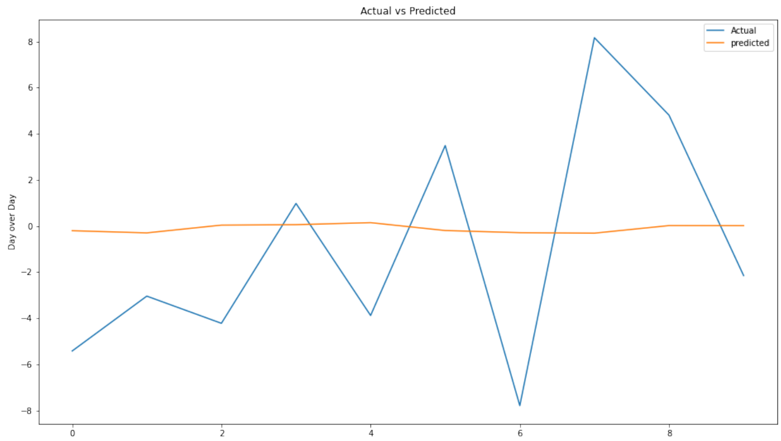

Figure 2: Actual vs Predicted with the RQ1 model

To compare the performance of the models among each other, the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are calculated, as both measures the error of the models. Both the RMSE and MAE can range from 0 to infinity. The lower the values, the better the model. Before being able to calculate the evaluation metrics, the inverse transformation of the scaled Y prediction is applied. Since our validation data is unscaled, the Y should be unscaled as well.

5 RESULTS

The results from all four models show that it is not possible to predict the short-term price change using the Number of Transactions and the Volume of Tweets 3. Of the four models, the best R^2 value is the combination of the Number of Transactions and Volume of Tweets as the LSTM input data. However, the model shows a low correlation between this data and the change in price, at a R^2 of 0.021. Therefore, the hypotheses of RQ1, RQ2, and RQ3 can be rejected.

Interestingly enough, the base model, which only takes into account the previous day's percentage change, performed the worst of all models explaining the variance of the dependent variable predicted from the independent variable (R^2 = -1.052). The negative R^2 value shows that the model performed worse than a straight line would. The same results can be found with the LSTM of the research questions; *To what extent is it possible to predict the short-term price of Ethereum based on the Volume of Tweets using an LSTM model?* and *To what extent is it possible to predict the short-term price*

Table 3: The results from the four models based on the research questions, combined with a selection of the parameters of the models.

Layers	Models	R^2
200	**LSTM (base)** Dense (20) Dropout (0.25)	-1.052
200	**LSTM (RQ1)** Dense (20) Dropout (0.25)	0.021
200	**LSTM (RQ2)** Dense (20) Dropout (0.25)	-0.018
200	**LSTM (RQ3)** Dense (20) Dropout (0.25)	-0.047
	Best R^2	**0.021**

of Ethereum based on the Number of Transactions using an LSTM model?. These results show that the models are not able to correctly predict the price change of Ethereum in the current setting.

During the fitting of each model around epoch 20, the models started to perform worse and were stopped by the EarlyStopping callback. After epoch 20, the training loss was higher or the same as the training loss at the first epochs of the four models. To compare the models themselves amongst each other, the RMSE and MAE were used. Of the four models **??**, the model, which only took into account the Number of Transactions, had the lowest RMSE and MAE scores (RMSE = 2.327 and MAE = 2.943). This means that that model was able to predict the value of change the best; however, it does not explain the variance. As can be observed in Figure 2 the model does not fit the data ever so slightly; similar behavior can be observed in the other models as well. The worst model to predict the value of change was the base model, which only used the previous Day over Day price change to predict future price changes (RMSE = 5.376 and MAE = 7.165).

6 DISCUSSION

The goal of this paper was to test whether the number of transactions and the number of tweets were helpful indicators to predict the short-term price of ETH. As the results show, a causal relationship can not be proved. As it turns out, the previous changes per day also do not have a causal relationship with the actual price change. The research questions and their hypotheses; *To what extent is it possible to predict the short-term price of Ethereum based on the Volume of Tweets and the Number of Transactions on the blockchain using an LSTM model?*, *To what extent is it possible to predict the short-term price of Ethereum based on the Volume of Tweets using an LSTM model?*, *To what extent is it possible to predict the short-term price of Ethereum based on the Number of Transactions using an LSTM model?*, were rejected in this paper due to missing a causal relationship.

The results were a bit surprising since at least the Number of Transactions ought to be related to buying, selling, and transferring of ETH, in addition to launching new smart contracts and further interactions on the Ethereum blockchain. Perhaps if the other interactions were removed, it would better fit; however that is unknown, and no such data source was available.

6.1 *Limitations*

Limitations of this study included a lack of hourly data on the Number of Transactions, the Volume of Tweets, and the price change. Since price fluctuations can occur at any time of the day, it would be possible that data on a 15-minute timescale could improve the accuracy since a large influx of transactions or tweets could hint at a change of interest or price. However, the data on the number of tweets was unavailable for Bachelor thesis research; only Master students and businesses can retrieve the Twitter archive and query it.

The Ethereum price data could be retrieved in such intervals; however, since the Twitter data was not available, that was not pursued. The larger data set could impact the overall accuracy as the current models only have 2176 data points to work with, which is relatively low for an LSTM model. Additionally, the introduction of look-back values could improve the accuracy as well as a higher look-back value enables the model to use more data to predict the subsequent time step. An alternative would be to use a Twitter extraction tool to listen for Tweets with a particular hashtag or keyword. However, that listener would need to run for a prolonged time as it would otherwise; the price data would be insufficient to develop a sound model. Additionally, more data could allow the LSTM model to better model sporadic changes in price during the day since the data currently used only considers the close value of Ethereum. Combined with hourly (or even shorter) interval data, the model could pick up on trends and better predict the changes due to it having more outliers. Combining outliers with other trends in the data could further improve the model's accuracy; however, the effect of this is unknown and should be further researched.

Since our model was performing worse with more epochs, it should be noted that multiple factors could cause the issue. In this paper, no optimizations were applied since the performance was significantly low, and with this data, no sound model was possible. However, in cases like ours, multiple approaches could prove helpful, including reducing the learning rate, providing more data, changing the Dropout rates, decreasing the batch size, and using a different optimizer.

6.2 *Contribution*

Previous research (Kim et al., 2021) found other aspects of the Ethereum blockchain statistics to be significantly related to the Ethereum prices. The researchers found that Ethereum-specific information, including the uncle block, gas price, gas consumption, and gas limit, were significantly associated with the Ethereum price. Similar to this research, they found no significant association between the Number of Transactions and the Ethereum prices; our research reinforces that finding. While most research is focused on predicting the actual price of the given asset, this research tries to predict the percentage change instead. Since the price development in Ethereum has a considerable variation in prices (Fig. 1), the day-over-day percentage change was chosen as it is normalized and not as erratic.

Additionally to the research mentioned above, our research investigated the question if the Volume of Tweets could predict the price change. In this research, no significant correlation was found. However, other researchers (Abraham et al., 2018) found that the model trained on the Volume of Tweets was able to predict the direction of price changes accurately. Since this research attempted to train the model to predict the percentage change instead of the actual U.S. dollar value, it is hard to compare the findings. Their research used a linear model instead of an LSTM model, which interestingly fit the data very well.

7 CONCLUSION

Using a Long Short Term Memory with the Number of Transactions and Volume of Tweets does not appear to produce accurate predictions of the Day over Day price change of Ethereum. The highest coefficient of determination for the proposed models was around 0.02 for the model, which represents Research Question 1, which shows the models explain at most 2% of the variation within the data. The least accurate model actually fit worse than a horizontal line, with a negative coefficient of determination of -1.052, which was trained on the previous day over day changes. It should be noted that with different optimizations, the coefficient of determination could increase; however, it is not very likely that the fit would be significant using the same data set.

The data set contains over 2176 data points of the Number of Transactions and Volume of Tweets; the data is the cumulative figure of one day. The price of Ethereum was extracted as the Close value of Ethereum of that day, which is transformed into a Day over Day percentage change. Of the four models, the model which represents the research question *To what extent is it possible to predict the short-term price of Ethereum based on the Number of Transactions using an LSTM model?*, this model was able to predict the change the best. In order to improve the accuracy and fit of the LSTM model on Day over Day price change, additional research and more data points would be needed. The extra data should be acquired by decreasing the time step to, for instance, hourly data of the Number of Transactions, Price Change, and Volume of Tweets.

Since multiple factors can have a significant impact on the price change of cryptocurrencies, including stock market performance, among other factors, more research should be conducted to identify meaningful metrics or events. Perhaps a combination of factors could improve the model significantly, although that is uncertain at this stage. Previous research on predicting the absolute value of Ethereum found that Tweet Volume can be an indicator of price movement; while this was not found in our research, more data could prove more accurate.

REFERENCES

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, *1*(3), 1.

Brownlee, J. (2020, 08). *Dropout Regularization in Deep Learning Models With Keras.* Retrieved from https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/

Brownlee, J. (2021, 01). *How to Develop a Bidirectional LSTM For Sequence Classification in Python with Keras.* Retrieved from https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/

Buterin, V. (2013). Ethereum white paper: A next generation smart contract & decentralized application platform. Retrieved from https://github.com/ethereum/wiki/wiki/White-Paper

Caldegren, A. (2018, 05). *The Influence Of Bitcoin On Ethereum Price Predictions.* Retrieved from https://www.diva-portal.org/smash/get/

diva2:1225270/FULLTEXT01.pdf
CoinMarketCap. (2022). *Cryptocurrency Prices, Charts And Market Capitalizations.* Retrieved from https://coinmarketcap.com/

Companies ranked by Market Cap - CompaniesMarketCap.com. (2022, 05). Retrieved from https://companiesmarketcap.com/

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hu, Z., Zhao, Y., & Khushi, M. (2021). A Survey of Forex and Stock Price Prediction Using Deep Learning. *Applied System Innovation*, *4*(1), 9. doi: 10.3390/asi4010009

Hua, Y., Zhao, Z., Li, R., Chen, X., Liu, Z., & Zhang, H. (2019). Deep Learning with Long Short-Term Memory for Time Series Prediction. *IEEE Communications Magazine*, *57*(6), 114–119. doi: 10.1109/mcom.2019.1800155

Kim, H.-M., Bock, G.-W., & Lee, G. (2021). Predicting ethereum prices with machine learning based on blockchain information. *Expert Systems with Applications*, *184*, 115480. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417421008915 doi: https://doi.org/10.1016/j.eswa.2021.115480

Kwon, D.-H., Kim, J.-B., Heo, J.-S., Kim, C.-M., & Han, Y.-H. (2019). Time series classification of cryptocurrency price trend based on a recurrent lstm neural network. *Journal of Information Processing Systems*, *15*(3), 694–706.

Lendave, V. (2021, 07). *How To Do Multivariate Time Series Forecasting Using LSTM.* Retrieved from https://analyticsindiamag.com/how-to-do-multivariate-time-series-forecasting-using-lstm/

Li, H., Y, S., & Y, Z. (2018). Stock Price Prediction Using Attention-based Multi-Input LSTM. *Proceedings of Machine Learning Research*, *95*, 454–469. Retrieved from http://proceedings.mlr.press/v95/li18c/li18c.pdf

Mai, F., Shan, Z., Bai, Q., Wang, X. S., & Chiang, R. H. L. (2018). How does social media impact bitcoin value? a test of the silent majority hypothesis. *Journal of Management Information Systems*, *35*, 19 - 52.

Nakamoto, S. (2009, 03). Bitcoin: A peer-to-peer electronic cash system. *Cryptography Mailing list at https://metzdowd.com*.

Nelson, D. M., Pereira, A. C., & De Oliveira, R. A. (2017). Stock market's price movement prediction with lstm neural networks. In *2017 international joint conference on neural networks (ijcnn)* (pp. 1419–1426).

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603-9611. Retrieved from https://www.sciencedirect.com/

science/article/pii/S0957417415005126 doi: https://doi.org/10.1016/j.eswa.2015.07.052

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018). Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. In *2018 ieee 3rd international conference on computing, communication and security (icccs)* (pp. 128–132).

Vitalik Buterin created one of the world's largest cryptocurrencies in his early twenties — here's how he did it and why. (2019, 02). Retrieved from https://www.businessinsider.com/vitalik-buterin-created-ethereum-one-of-the-worlds-three-largest-cryptocurrencies-2019-1/