Social Media Influence on Ethereum Value

By

Liam Newton

Juan Sillero

Kazi Abdul Muktadir

**Abstract**

With Cryptocurrencies high return and volatility it is an important topic of discussion on forums like Reddit. In this social media study, we will use sentiment analysis, Granger causality, and Naive Bayes to determine a relation between positive comments and Cryptocurrency value. Since there are so many types of Cryptocurrencies, we have decided to focus on one, Ethereum. Ethereum is an open-source, public, blockchain-based distributed computing platform and operating system.

**Introduction and Study Statement**

The crypto market is one of the most volatile and unpredictable markets in the world. Cryptocurrency volatility is reason enough for the curiosity of Data Scientists to try and figure out patterns and trends in the market to potentially understand what causes these seemingly unpredictable changes. Interestingly, social media seems to be one place to learn about these changes. Social media is a platform for sharing public opinions on-line and; therefore, has become an integral part of a cryptocurrency success as we see that “messages on Internet forum have stronger impacts on future bitcoin market measures at a daily frequency” according to the paper [*The Impacts of Social Media on Bitcoin Performance*](http://elibrary.aisnet.org/Default.aspx?url=https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1393&context=icis2015). For our research, we will look at how positive mentions on Reddit affect the value of a cryptocurrency price over a range of time. We will use a data set containing Reddit comments from the cryptocurrency subreddit to study the patterns. There are quite a few challenges in making the Reddit data available for our use. Firstly, we do not have enough computation power to effectively use all the Reddit comment data. In an attempt to reduce the amount of data used, we will initially look at Ethereum over a small period of time. The time period will be determined by looking at Ethereum’s price graph and determining a time range with a significant change. The next challenge will be to clean out missing data and noise. Noisy data is meaningless or uninterpretable data. To clean the data, we will remove empty comments or comments that are shorter than one character long. Since we have a large data set, removing some entries will not affect our results too much.

The problem of predicting stock value and in our case Ethereum values is a sought after question, which is normally solved using historical trend data from stock or cryptocurrency, such as in the Moving Average Convergence Divergence or MACD formula. In the article by Kim et al. [2016], they attempt to use social media to predict cryptocurrency price fluctuations. Our goal is to determine whether there is a relation between Ethereum and positive Reddit comments which then can be used to predict Ethereum price fluctuations.

**Past Work**

An important source of reference for the project is *Twitter mood predicts the stock market* by Johan Bollen, Huina Mao, and Xiaojun Zeng [2011] which was published in the Journal of Computational Science Volume 2 Issue 1. This paper looks at whether the moods derived from the sentiment analysis of large amounts of twitter data are correlated to the value of the Dow Jones Industrial Average (DJIA) over time (Bollen, p. 1). The researchers use two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). Next, Granger causality analysis and a Self-Organizing Fuzzy Neural Network are used to test how the moods are predictive of changes in DJIA closing values. An accuracy of 86.7% is found in the predictions of the daily up and down values of the DJIA.

Another source of reference for the project is *Forecasting Price of Cryptocurrencies using Tweets Sentiment Analysis* by Jain, Tripathi, Saxena and Dwivedi [2018]. The paper focuses on predicting the prices of the cryptocurrencies, Bitcoin and Litecoin, by considering their mentions on Twitter. The data was collected for 30 days. The method that was used to predict these fluctuations on the currencies is Multiple Linear Regression which forms a relationship between two or more independent variables and a dependent variable by fitting a linear equation to a trained set. The performance of the model is evaluated by using R2\_Score and accuracy. The R2\_Score of predicting Bitcoin is 59% and 44% for Litecoin.

Another reference that is considered important for the current research and analysis is the research paper *Sentiment analysis of Twitter data for predicting stock market movements* [Pagolu, 2016]. In this paper, the authors decided to use Microsoft to track how tweets affected its price on the stock market. More than 2.5 million tweets containing words related to Microsoft were collected. N-gram representation was used to match the corpus of the text being studied. N-gram was used with a random forest algorithm and it showed an accuracy of 70.5% on predicting stock price changes of Microsoft. Using the Word2vec representation, an accuracy of 70.2% was obtained using a random forest algorithm as well.

In the article *The impact of social and conventional media on firm equity value: A sentiment analysis approach* written by Yang Yu, Wenjing Duan and Qing Cao [2013], the authors compare conventional media to social media to see which type of media has more of an impact on firm equity value. To determine whether social media has an impact on frm equity the authors first randomly selected 824 companies and obtained their financial data. They then acquired their social media information from blogs, twitter, news, and forums. On the collected data, they ran an automated sentiment analysis program which graded each article, blog, etc… from -1 to 1 where -1 is a negative view and 1 is positive. The collected data was processed both from social media sources and conventional media sources. To measure accuracy of the sentiment analysis, the f-measure was calculated using a set of test data from the Cornell movie-review data set. An accuracy on average of 80% was achieved on the data set. Finally Naive Bayes was run to determine the probability of an event occurring given some other event; in this case, to predict whether equity changes based on the sentiment analysis. The results found that social media has a stronger effect on stock price than conventional media, but they both have some impact on stock returns.

Kim et al. [2016] *Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies* attempt to determine if there is a correlation between positive mentions in social media and cryptocurrency value fluctuation. Like the article written by Yu et al. [2013], sentiment analysis was run on the comment data and graded from -1 to 1, where 1 is a positive comment. An additional constraint was added: “...x, -1< = x < -0.6, -0.6< = x < -0.2, 0.2 < = x <0.6, and 0.6< = x < = 1.0...” (Kim et al. 2017) where the ratings varied from very negative to very positive, respectively. The graph of the normalized sentiment analysis and normalized bitcoin price over time was then compared resulting in a statistically significant correlation between positive and negative comment and Bitcoin price fluctuation. The Granger causality test was performed on each crypto currency with a time lag of 1 to 13 days. The results for Ethereum showed that the price fluctuation could be predicted with a time lag of two to six days with an accuracy of 71.823%. With a time lag of 11 to 13 days, the accuracy was 66.129%. The authors found that tracking social media can be useful in predicting cryptocurrency fluctuation, but could be even more accurate if economic principles are also applied.

Lastly, the *Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis* by Stuart Colianni, Stephanie Rosales, and Michael Signorotti looks at real-time Twitter data and attempts to predict market movement of securities and other financial instruments. The paper attempts to prove that mentions in Twitter leads to a positive impact on the prices of cryptocurrency. The main Cryptocurrency followed in this study is Bitcoin. By applying supervised learning algorithms, such as logistic regression, Naive Bayes, and support vector machines, the researchers gained a prediction accuracy of 90%. The researchers performed rigorous error analysis in order to ensure that accurate inputs are utilized at each step of the model which helped increase results by 25%.

**Methodology and experimental design**

The data that will be collected is within the time frame of May 2017 to July 2017.Our data includes two data sets, one with the price of the Ethereum in the given timeframe, and another one with comments from Reddit filtered with keywords related to cryptocurrency in the given timeframe as well.

The Reddit comments dataset is collected using Google BigQuery, where SQL statements are used to trim the data to the bare minimum needed. The data that was collected using SQL includes the attributes: subreddit, body, score, author and created\_utc. The resulting data from BigQuery is downloaded in CSV format and imported into R.

The cryptocurrency price history is obtained from Yahoo finance. After obtaining the price history data, we convert the price data to a nominal value. To convert to nominal data, we compare the current price to the day before. If the price increases, then the attribute increase is set to “yes” and vise versa for a decrease in prices. The price data is converted into nominal data for calculating Naive Bayes. For days where there is an increase in value of Ethereum, we will only be able to see if there was an increase or a decrease which will make the analysis easier.

After the data is preprocessed, we perform sentiment analysis on the data to convert the comments to specific moods. Sentiment Analysis is done using R with the following libraries from R: tidyverse, tidytext, glue, and stringr. We are able to identify what words have a bigger influence in a positive or negative Ethereum price, as it will be explained more in depth throughout the paper.

When doing sentiment analysis, we will focus on total positive comments by adding up the total positive sentiment for each day and subtract total negative sentiment. PosMinusNeg is the resulting attribute from this calculation which we use. We then create a Naive Bayes model to make predictions on cryptocurrency value. We use positive minus negative as the attribute used to predict if there was an increase in the price of the cryptocurrency.

To determine whether there is a relation between positive Reddit comments and Ethereum closing value, the Granger Causality test is run. The Granger test is run with three attributes: day, close, and PosMinusNeg, where day is a list of numbers from 1 to 93, close is the closing price of Ethereum, and PosMinusNeg is the total positive comments from the sentiment analysis minus the total negative comments per day.

We determine the accuracy of our model by using a different time frame and predicting the prices of Ethereum in those time frames based on Reddit comments. We compare the predictions to the actual price of Ethereum, and the results are used to determine the accuracy of our model.

**Experiments Results**

Over the course of the study, 108,147 comments from Reddit were collected. These comments were specifically selected from the Cryptocurrency subreddit. The time frame of the collected data is from May 2017 to July 2017. The time period was selected because it was the first appreciable spike and drop in Ethereum value. The data was collected from Google BigQuery using an SQL statement.

In this section, we show the results of our sentiment analysis, and the results obtained from our prediction model. Then we compare them to the actual prices of Ethereum in the same time frame that we predicted. Based on the predicted outcome, we calculate the ability our model has to predict an increase in Ethereum value.

**Sentiment Analysis**

Figure 1 shows our initial results from sentiment analysis. In Figure 1, we show total positive comments for each day and in figure 2 is the closing value of Ethereum for each day from May 1st to July 31st. In figure 1, we can see a general trend in the positive direction as the Ethereum price spikes as does total positive comments. From figures 1 and 2, we can determine if there is enough of a relation to warrant additional investigation.

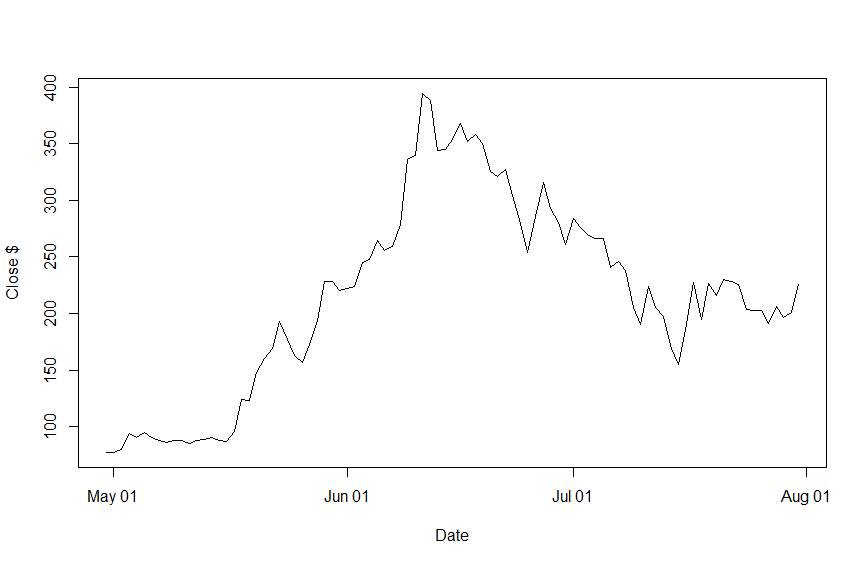
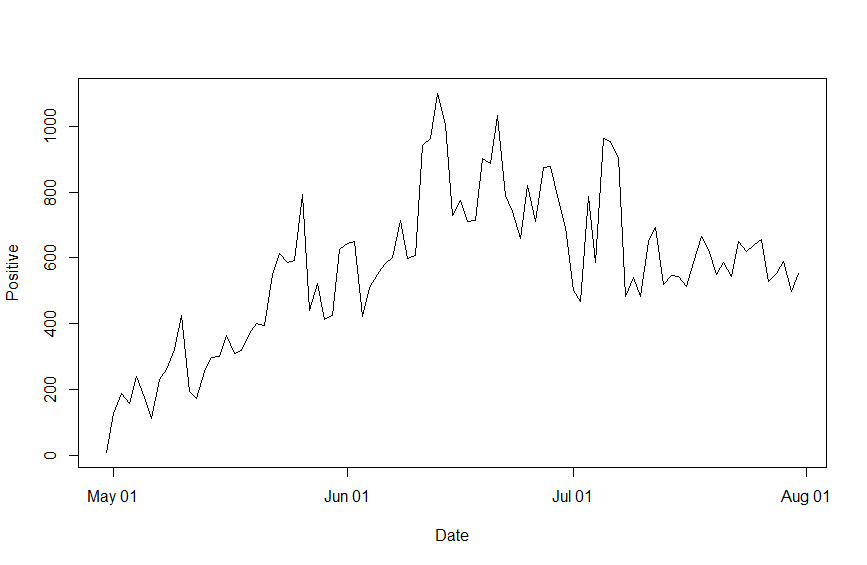
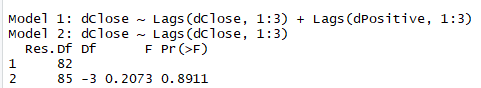
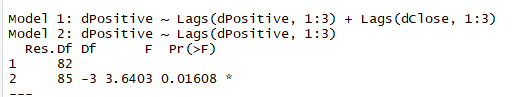
****

Figure 1: Total positive comments over time Figure 2: Ethereum value in Dollars over time

**Granger Causality**

To determine whether what we saw in figures 1 and 2 was actually a relation, we ran the Granger Causality test. The test was run using the forecast libraries in R and the total positive comments were compared to the closing value of Ethereum. Figure 3 shows the Granger causality of total positive comments causing Ethereum close value. The resulting value from the test is significant; however, we must test the reverse case. In figure 3, we check whether Ethereum closing value Granger causes positive comments and we find an insignificant value. Thus we can conclude that total positive comments Granger causes Ethereum closing values. We obtained the best results with a time lag of two to six days. In figures 1 and 3, we used a time lag of 3 days where Granger Causality shows its best results.





|  |  |
| --- | --- |
| Figure 3: Shows Granger Causality: positive comments causes Ethereum close price, time lag of 3 days. | Figure 4: Shows Granger Causality: Ethereum close price causing positive comments, time lag of 3 days |

**Naive Bayes**

To determine whether our hypothesis was correct or not, the Naive Bayes classifier was used to predict if there was an increase in the prices of Ethereum based on our sentiment analysis. The first attribute that was used to classify was positive minus negative as this attribute indicated the total amount of positive and negative comments for one day and adds them together. As can be seen in Figure 6 below, this attribute is not very good at predicting the price of ethereum. One thing that was noticed is the attribute is better at predicting if there will be an increase on the price of ethereum than to predict if there will be a decrease. Like it is shown in Figure 5, we ran the model to predict whether the price of Ethereum would go up or down, and we found out that when predicting a decrease the model can predict 40% of the values correctly. On the other hand, when predicting if there will be an increase, the model is correct at predicting 71.74% of values.

Like it was mentioned before, the accuracy for predicting an increase is better than for predicting a decrease.

|  |  |
| --- | --- |
|  |  |
| Fig 5 Naive Bayes results of predicting if there will be an increase, predicted results and actual results | Figure 6. Naive Bayes plot with total positive minus negative comments where yes is an increase in Ethereum value |

We used other sentiments to try and see if they would be better at predicting the fluctuations in the price of Ethereum, but we did not find one that was significantly better than positive minus negative comments. The sentiments we tried were very similar and definitely not the best at predicting the increase or decrease on the price of the cryptocurrency.

**Discussion**

In our study, we analysed the sentiments of Reddit comments. We then analysed whether these sentiments have any correlation with the stock price of Ethereum. Our study reveals that there is a correlation between the Reddit comments and price of Ethereum. We concluded the correlation by comparing the graphs from sentiment analysis with the graph showing price changes of Ethereum over the same period of time. To prove the relational trend in the graphs, we used the Granger causality test, which showed that positive comments Granger cause closing value.

Moreover, we implemented a Naive Bayes model to use the results from sentiment analysis to predict the price changes of Ethereum. However, the results produced were not accurate enough because we did not have enough data and the time range used was not long enough. We believe that with larger datasets, our analysis would improve.

**Future Research**

In future research, an extend time frame of the comments analyzed to see if the results when predicting fluctuations are the same. Extending the time frame would allow us to see if there is any difference in our models results. Having a more granular model that classifies a large increase in price and small increase in prices as well as highly negative and highly positive comments would improve our results. We also would like to analyze different Cryptocurrencies, specifically cryptocurrencies that have had bigger fluctuations in their price so that we might be able to have our model predict more efficiently the fluctuations of the currency.

**References**

* Bollen, J., & Mao, H. (2011). Twitter Mood as a Stock Market Predictor. *Computer,44*(10), 91-94. doi:10.1109/mc.2011.323
* Jain, Arti & Tripathi, Shashank & DharDwivedi, Harsh & Saxena, Pranav. (2018). Forecasting Price of Cryptocurrencies Using Tweets Sentiment Analysis. 1-7. 10.1109/IC3.2018.8530659.
* Go, Alec, Lei Huang, and Richa Bhayani. Twitter sentiment analysis . Entropy 17 (2009).
* [**arXiv:1805.00558**](https://arxiv.org/abs/1805.00558) **[physics.soc-ph]**
* **Colianni, Stuart et al. “Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis.” (2015).**
* Pagolu, Ventaka Sasank, et al. “Sentiment Analysis of Twitter Data for Predicting Stock Market MovementsSentiment Analysis of Twitter Data for Predicting Stock Market Movements.” SCOPES, 28 Oct. 2016, pp. 1–6., arxiv.org/pdf/1610.09225.pdf.
* Yu, Duan, and Cao. "The Impact of Social and Conventional Media on Firm Equity Value: A Sentiment Analysis Approach." *Decision Support Systems* 55.4 (2013): 919-26. Web. [DOI](https://doi.org/10.1016/j.dss.2012.12.028).
* Kim, Young Bin, et al. “Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies.” *PLOS ONE*, Public Library of Science, 17 Aug. 2016, journals.plos.org/plosone/article?id=10.1371/journal.pone.0161197.
* Mai, Vai, et al. “The Impacts of Social Media on Bitcoin Performance” Thirty Sixth International Conference on Information Systems, 2015.
* Data-set compiled by [/u/into\_the\_matrix](https://old.reddit.com/user/Stuck_In_the_Matrix) on this [post](https://old.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/)
* Uploaded to Google BigQuery by [/u/fhoffa](https://old.reddit.com/user/fhoffa) on this [post](https://old.reddit.com/r/bigquery/comments/3cej2b/17_billion_reddit_comments_loaded_on_bigquery/)
* [Google BigQuery Dataset](https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.2015_05)