Predicting Fluctuation in Ethereum Cryptocurrency Using Twitter Sentiment Analysis

NIKETH SHETTY, Indiana University, nhshetty@iu.edu
AKSHAY REDDY, Indiana University, reddyak@iu.edu

ABSTRACT

We know from research in the past that Twitter sentiment analysis can be used to predict movement in the stock market [1]. Research has also been done on successfully predicting the price of Bitcoin cryptocurrency using Twitter sentiment analysis [2]. With our study we intended to prove that it is possible to predict changes in the price of Ethereum cryptocurrency (also known as Ether, denoted by ETH) by analyzing Twitter sentiment. We first attempted to establish a linear relationship between Twitter sentiment and the change in the price of Ethereum. Having failed to do so, we used three labels for the change in price of Ethereum, increase (if price change is greater than zero), decrease (if price change is lesser than zero), same (if price change is exactly zero) and attempted to classify the trend according to these labels. We use supervised machine learning algorithms such as Naïve Bayes Classification, k-Nearest Neighbors Classification and Neural Networks and used 4-fold cross validation on our models. While the other classifiers produced an accuracy of around 53%, our Naïve Bayes classifier gave us an accuracy of around 57%. Given that the majority class in the labels had a percentage of is around 52%, we cannot say that we have a robust model and are forced to conclude that with the data in hand, and the resources at our disposal, it is not possible to establish a relationship between Twitter sentiment and the change in Ethereum price. But the score for the Naïve Bayes classifier does indicate that we cannot eliminate the possibility of finding a relationship in the future.

INTRODUCTION

"A cryptocurrency is a digital asset designed to work as a medium of exchange using cryptography to secure the transactions and to control the creation of additional units of the currency" ("Cryptocurrency", Wikipedia) [3]. Cryptocurrencies are traded on the cryptocurrency market (like foreign exchange trading) and can also be traded for US Dollars. Bitcoin is the most popular and well known cryptocurrency, and has the largest market cap (over 207 billion US Dollars as of 12/06/2017) [4]. Ether, more popularly known as Ethereum, the technology platform it is based on, is the cryptocurrency with the next largest market cap (over 42 billion US Dollars as of 12/06/2017) [4].

Our study focuses on finding out if there is a relationship between the fluctuation in the price of Ethereum and the current sentiment in Twitter. "Ethereum is a decentralized technology platform that runs smart contracts: applications running on a blockchain that run exactly as programmed without any possibility of downtime, censorship, fraud or third-party interference" (https://ethereum.org/) [7]. According to an article on Forbes, it is an exciting new technology that is being implemented increasingly around the world. As the technology becomes more widespread, the Ethereum cryptocurrency is more widely used [8]. Also, according to another article on Forbes, the cryptocurrency market is becoming bigger and more important each passing day, with an appreciation of more than 1200% in the past year [9]. The total market cap for the entire

cryptocurrency market is over 366 billion US Dollars [4]. Due to these reasons, we felt our study was important and could potentially have been highly beneficial to people interested in investing in these markets.

Pak and Paroubek tell us Twitter is a very good source to extract the sentiment of the general public [10]. We used a third-party sentiment analysis package for Python, that uses the AFINN lexicon to provide a sentiment score to any text. AFINN gives each English word a score of between minus five and plus five with more negative scores indicating a higher negative emotion and vice versa. We then used the average sentiment score per Tweet for each day to build our models.

RELATED WORK

Using Twitter sentiment data to perform financial time series analysis is not a new concept. As mentioned before, Bollen et. al. proved that Twitter mood could be used to predict changes in the stock market [1]. There are several papers and studies on the relationship between the Bitcoin market and social media data. Mai et. al. proved that sentiment analyses of Twitter and an internet talk forum could be used to predict certain variables in the Bitcoin market [5]. Matta et. al. analyzed the spread of Bitcoin's price in relation to web search trends and volumes of tweets. They were able to successfully establish a relationship between the two [6]. Colianni et. al. showed that Twitter sentiment analysis could be used to predict the market movement of Bitcoin [2]. But according our knowledge, there are no studies on how Twitter sentiment affects the price of Ethereum. The goal of our study was to bridge this gap. Having established the importance of Ethereum in the emerging blockchain scene, we felt this hitherto unexplored area was a good choice to examine more closely.

DATA DESCRIPTION

We started with two datasets. The first, a collection of Tweets, and the second, details of the price of Ethereum as it has varied each day. Both datasets ranged in date from 08/08/2015 to 10/02/2017. The Twitter API does not allow retrieval of tweets that are more than 7 days old. Hence, we used an open source GitHub repository called GetOldTweets-Python (https://github.com/Jefferson-Henrique/GetOldTweets-python) to extract older tweets. We extracted tweets that contained the text and/or hashtags 'ETH', 'ethereum' and 'ethereum smart contract'. We obtained over 163,000 tweets in the date range. The dataset had ten features: username, date, time, retweets, favorites, text, mentions, hashtags, id and permalink. The features we were interested in were date (date the tweet was tweeted) and text (content of the tweet). We then performed sentiment analysis on each tweet and found the average sentiment analysis for each date.

We obtained the details of the price of Ethereum from a Kaggle repository (https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory/data). The dataset had seven features: Date, Open, High, Low, Close, Volume and MarketCap. We added a feature, NextDayChange, which was the change in the price of the following day for each date. We obtained this by subtracting the opening price (Open feature) of the next day from the closing price (Close feature) of the next day. We then added another feature, Change, which was a factor variable with three labels: 'Increase' (if NextDayChange was positive), 'Decrease' (if NextDayChange was negative) and 'Same' (if NextDayChange was zero).

We then collated the two datasets to obtain a dataset with the following features: Date, Sentiment (average sentiment score per tweet), Open, Close, Volume, MarketCap, NextDayChange and Change. While researching the history of Ethereum, we ran across a New York Times article by Nathaniel Popper that spoke about the creation of the Enterprise Ethereum Alliance (EEA), a group of 30 companies including corporate giants like Microsoft and J. P. Morgan, that would "create a standard version of the Ethereum software that businesses around the world can use to track data and financial contracts." [11] We noticed that since the creation of EEA in March 2017, the price of one unit of Ethereum had risen from around 15 US Dollars to around 446 US Dollars (as of 12/06/2017). So we decided to create a subset of our data with data from 03/01/2017 to 10/02/2017, in case it could give us any additional insights.

METHOD AND RESULTS

There were two major steps to our analysis: sentiment analysis and statistical analysis.

Sentiment Analysis

As mentioned previously, we used the AFINN lexicon to perform sentiment analysis. The AFINN python package gives a sentiment score for a piece of text by summing up the sentiment scores for each word in the text. We ran this function on our dataset and obtained a sentiment score for each tweet in it. We then summed up the sentiment scores for each date divided this sum by the total number of tweets obtained for that date. This gave us an average sentiment score per tweet for each date. To provide an idea of the context of the sentiment scores, we applied the same procedure to six-day periods around two major emotional events this year: Thanksgiving Day (November 23rd, 2017) to indicate positive emotion (Figure 1), and the date of the Las Vegas shooting tragedy (October 1st, 2017) to indicate negative emotion (Figure 2).

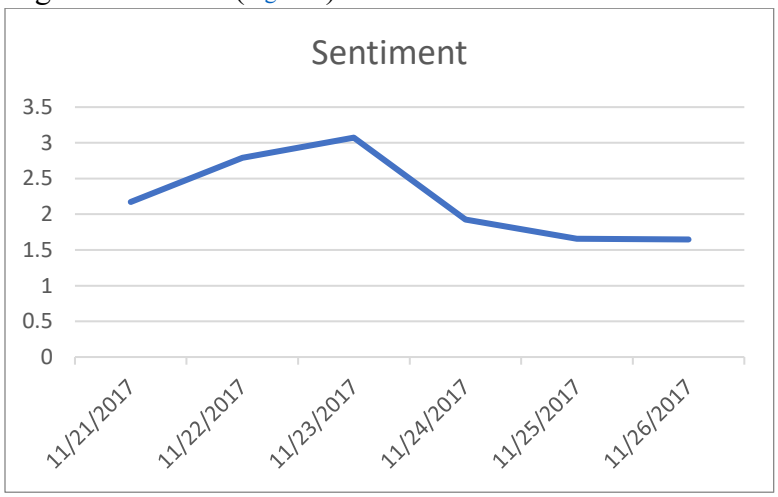

Fig. 1. Average sentiment analysis around Thanksgiving Day

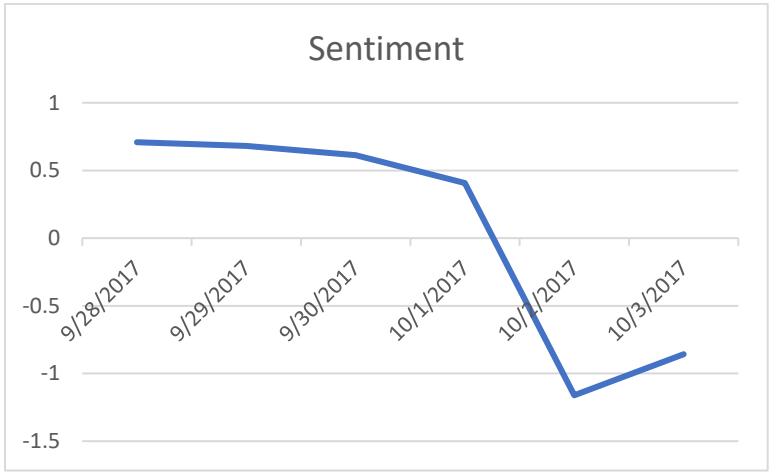

Fig. 2. Average sentiment analysis around Las Vegas Shooting

It can be clearly seen in Figure 1 that the sentiment score is highest (little over 3) on Thanksgiving Day, while Figure 2 shows that the sentiment score is lowest (little under -1.1) on the day after the shooting (the incident took place on the night of October 1st).

Having obtained the average sentiment score per tweet for each date, we then proceeded to

combine this data with the dataset containing the Ethereum price data. We got rid of the features we did not need and had our final dataset.

Statistical Analysis

Our intention was to establish a relationship between change in Ethereum price and Twitter sentiment. In terms of our features, we wanted to establish a relationship between NextDayPrice and Sentiment. We started off by plotting a scatter plot with NextDayPrice on the y-axis and Sentiment on the x-axis.

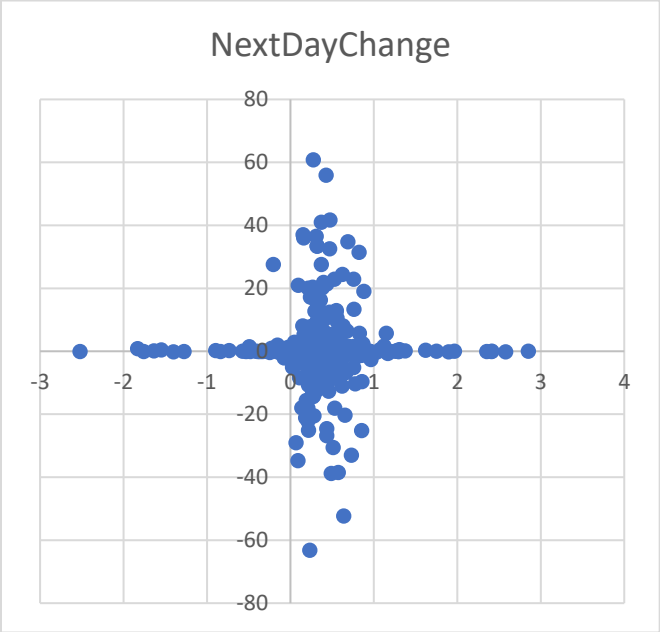

Fig. 3: Scatter plot of NextDayChange vs Sentiment

From the plot, it is pretty clear that there is no linear relationship between NextDayChange and Sentiment. Our next step was to evaluate if there was any relation between the predictor (Sentiment) and the response (NextDayChange). We calculated the correlation between the two and obtained a Pearson's correlation coefficient of 0.0017. This meant there is practically no correlation at all between the predictor and the response.

We then tried to use other features along with Sentiment to see if we would be able to build a linear regression model. Linear regression is a predictive learning technique that attempts to find the line that best explains the distribution between the response (y) and one or more predictors (X). But the results of our linear model showed that Sentiment had no significant linear effect on NextDayChange.

So, we decided to look at our problem as one of classification. We decided to try and train a classifier to predict the labels of the feature Change. This feature is a factor variable with three classes: 'Increase', 'Decrease' and 'Same'. There is a total of 787 rows, of which 399 are 'Decrease', 384 are 'Increase' and 4 are 'Same'. This gave us a baseline of 399/787 = 50.70%. Our target was to build a classifier that gave us an accuracy significantly greater than this number.

We decided to use the Python package scikit-learn, that has pre-defined functions for most common machine learning and data analysis algorithms. The classification algortims we chose to implement are: Naïve Bayes, k-Nearest Neighbor, Support Vector Machines, and the artificial neural networks class Multi-Layer Perceptron.

We ran the algorithms on the data and performed k-fold cross validation to obtain the accuracy scores of our classifiers. We chose k=4 for our cross validation since the label "same" only occurred four times in the dataset. We chose a train-test split of 80%-20%. The results of our cross-validation accuracy scores are displayed below in Table 1.

Support Vector Machines	Naïve Bayes	k-Nearest Neighbor	Multi-Layer Perceptron (Neural Network)
50.70%	49.30%	48.03%	50.57%

Table 1: 4-fold Cross Validation accuracy of entire data set

As we can clearly see, the accuracy is about the same (or just below) the baseline. This told us that the classifiers we used were not picking up any kind of signal at all. Thus, we were unable to establish any type of predictive relationship between Twitter sentiment and the fluctuation in Ethereum price. We ran the same classifiers on the subset we had created containing just the data from March to October 2017. The results of these cross-validation accuracy scores are displayed below in Table 2.

Support Vector Machines	Naïve Bayes	k-Nearest Neighbor	Multi-Layer Perceptron (Neural Network)
52.32%	56.97%	52.85%	52.32%

Table 2: 4-fold Cross Validation accuracy of data from March 2017

This subset had 216 rows, of which 103 are 'Decrease', 113 are 'Increase' and none are 'Same'. This gave us a baseline of 113/216 = 52.31%. As we can see, except the Naïve Bayes classifier, all the classifiers had scores that were about the same as the baseline. Even the Naïve Bayes classifier did not have a score much higher than the baseline. While this result by no means proves any relationship between the response and the predictor, it does indicate the possibility of a signal (albeit a minute one).

CONCLUSION AND FUTURE WORK

Even though our model failed, it does not mean it is impossible to predict cryptocurrency prices from Twitter data. As we have mentioned in previous sections of this paper, there have been several studies that performed a task similar to this quite successfully. Bollen et. al. proved that it is possible to predict financial time-series values using Twitter sentiment [1]. Colianni et. al. proved that it is possible to predict cryptocurrency prices from Twitter data by successfully predicting the price of Bitcoin [2]. There is also the fact that our Naïve Bayes classifier scored an accuracy of nearly 5% above the baseline for a subset of our data. Thus, there is indication of the possibility of a route existing to find a relationship between Ethereum prices and Twitter sentiment. There are several possible reasons for our model failing and this is what future collaborations can improve upon.

Data Collection: We collected our Twitter data by using the GitHub repository GetOldTweets (got3). Basically, this program collects tweets by searching for a query tag in the 'Search' bar of the Twitter website. Since this is not an official Twitter API, we cannot be sure that it accurately collected all the tweets we required. In fact, upon examining the raw dataset, we found several tweets that had been extracted from outside the date range we had entered. Therefore, it is also possible the program emitted a significant number of tweets within the date range. Twitter's API can be used to collect tweets that are more than 7 days old, but requires payment. We received no funding for this study, and thus had to rely on got3 for our data.

Noise and Data Cleaning: If one were to search for the term "Ethereum" on Twitter's search bar, many of the tweets that appear in the search results are going to be advertisements selling Ethereum. As one of the few cryptocurrencies (along with Bitcoin and Litecoin) that can be bought directly with US Dollars (and many other major

currencies), Twitter is a very popular platform for individuals and companies to advertise the sale and purchase of Ethereum. Thus, a lot of the tweets in our dataset are just advertisements, and have either zero, or very negligible, sentiment. It was beyond our ability and resources to figure out a way to get rid of (or even substantially reduce) this 'noise' in our data. But future collaborators with more resources and knowledge can possibly eliminate this noise and keep the tweets that are most useful to their model.

Taking Advantage of Important Twitter Users: A big problem with our model is that we give equal importance to every single tweet. This put us at a disadvantage not just in terms of noise in our data, but also from the aspect of not giving more weight to tweets from important users. There are bound to be regular users of Twitter who have made money on the cryptocurrency market and whose tweets have a greater influence on the market. Their tweets could be given additional weightage, either while calculating sentiment score, or with an additional feature in the final dataset. One way to identify them could be by tracking the usernames whose tweets had a high number or retweets and/or had been favorited by other users a high number of times.

TEAM MEMBER CONTRIBUTION

Akshay Reddy:

Akshay's strengths are coding (especially in Python) and he took on the responsibility of almost all the coding our project required, right from modifying the got3 program to suit our needs, to finding and applying Python packages for the statistical and machine learning algorithms we needed. Akshay was also the editor for any writing Niketh did, proof-reading the documents and suggesting changes where necessary.

Niketh Shetty:

Niketh's strengths are with statistical analysis and content writing. He performed the major analysis and manipulation of the data in R and Excel and obtained the results. Akshay would then translate those steps into Python code. Niketh also did most of the writing required for this project.

REFERENCES

[1]: Bollen J., Mao H., Zeng X. (2010). Twitter Mood Predicts the Stock Market. Retrieved from https://arxiv.org/pdf/1010.3003.pdf

[2]: Colianni S., Rosales S., Signorotti M. (2015). Algorithmic Trading of Crypto-currency Based on Twitter Sentiment Analysis. Retrieved from http://cs229.stanford.edu/proj2015/029_report.pdf

[3]: Cryptocurrency. (n.d.). In *Wikipedia.* Retrieved on December 5, 2017 from https://en.wikipedia.org/wiki/Cryptocurrency#cite_note-2

[4]: Cryptocurrency Market Capitalizations. (n.d.). In *CoinMarketCap.* Retrieved on December 6, 2017 from https://coinmarketcap.com/

[5]: Mai F., Bai Q., Shan Z., Wang X., Chiang R. (2015). From Bitcoin to Big Coin: The Impact of Social Media on Bitcoin Performance. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2545957

[6]: Matta M., Lunesu I., Marchesi M. (2015). Bitcoin Spread Prediction Using Social And Web Search Media. Retrieved from https://www.researchgate.net/publication/279917417_Bitcoin_Spread_Prediction_Using_Social_And_Web_Search_Media

[7]: The Ethereum Foundation. (n.d.). *Ethereum*. Retrieved on September 24, 2017 from https://ethereum.org/

[8]: Rahul Singireddy. (October 18, 2017). Vinay Gupta On Why Ethereum Is The Future. *Forbes*. Retrieved from https://www.forbes.com/sites/rahulsingireddy/2017/10/18/vinay-gupta-on-why-ethereum-is-the-future/#5a7f64d256f2

[9]: Charles Bovaird. (November 17, 2017). Why The Crypto Market Has Appreciated More Than 1,200% This Year. *Forbes*. Retrieved from https://www.forbes.com/sites/cbovaird/2017/11/17/why-the-crypto-market-has-appreciated-more-than-1200-this-year/#30fb13d36eed

[10]: Pak A., Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

[11]: Nathaniel Popper. (February 27, 2017) Business Giants to Announce Creation of a Compu-ting System Based on Ethereum. *New York Times*. Retrieved from https://www.nytimes.com/2017/02/27/business/dealbook/ethereum-alliance-business-banking-security.html