Bitcoin Price Prediction Using Twitter Sentiment

10th December 2021

Vyoma Desai
MS in Computer Science
University of Missouri -
Kansas City MO USA
vhdmmh@umsystem.edu

Asim Javed
MS in Computer Science
University of Missouri -
Kansas City MO USA
ajbng@umsystem.edu

Ergin Bostanci
MS in Computer Science
University of Missouri -
Kansas City MO USA
ebp52@umsystem.edu

Zeeshan Saleh Qamar
MS in Electrical Engineering
University of Missouri -
Kansas City MO USA
zsqbfc@umsystem.edu

Molan Zhang
MS in computer Science
University of Missouri -
Kansas City MO USA
mz9kk@umsystem.edu

ABSTRACT

Bitcoin is the new emerging payment system that is getting worldwide acceptance, and unlike other currencies this doesn't require to be under any administrative control and it is decentralized. Just like stocks, bitcoin has a price history which helps to predict its prices and trends, as history repeats. The project aims to find fusion based solutions where we can predict the value of bitcoin based on different methods

Like stocks, bitcoin prices are also affected by the sentiments of the users, this relation and the predictive power of machine learning is being used by researchers and scientists to come up with a system that is able to completely use twitter's predictive ability based on occurrence of events thereby the level of financial market increases day by day. This paper proposes an algorithm to predict bitcoin prices based on available tweets. Tweets are extracted based on generated keywords with translation into English Language for better understanding and accuracy. Sentiment analysis scores and bitcoin prices of the past have the capacity to predict new bitcoin prices. In this project we are focusing on bitcoin price prediction using twitter sentiment.

KEYWORDS

Keywords - BTC, BITCOIN and other Languages such as Chinese, Japanese keywords etc. Social media, Twitter tweets, Sentiment analyzing, Text mining, data preprocessing, machine learning, linear regression, time series analysis, model training, prediction analysis, time period range and interval - 4 hours, 10 days, 50 days, 100 days price prediction outcome.

1 INTRODUCTION

Before Bitcoin came into the world, there were a number of other digital cash technologies named as ecash systems of Chaum and Brands. In 1992, the first proposal by Dwork and Naor proposed solutions to computational puzzles based on value. The currency was launched in 2009 and was outsourced, bitcoins are rewarded or created as a result of mining.

After its development Bitcoin became one of the decentralized electronic currency systems which thereby introduced a big shift in the financial system[1]. Bitcoin is however not controlled by government or bank or other policies due to its decentralized existence and absolutely virtual digital procedure. Firstly, the main goal of Bitcoin is promotion of service transactions. Bitcoin is growing rapidly and it has attracted most of the users around the globe. With time, Bitcoin price is now constantly moving in real-time, the same as that of the stock exchange. Based on these fluctuations, modem world machine models can create their own model to predict the bitcoin price in real-time using a lot of social media data. These can be advantageous for people in finance and different organizations etc when the accuracy level of prediction is higher and if the predictions can be made at a higher level of accuracy.

The Internet has grown massively in recent years. It's been so easy and user-friendly to exchange information, share knowledge and gain experience with the emergence of social media networks and many communication networks.

Social media websites like twitter are generating huge amounts of data, called tweets which are in great volumes, we can get nearly 4.5 million tweets in a day. This immense

information can be helpful in the capture and analysis of bitcoin patterns through machine learning, and training our model for price prediction. For Bitcoin, It is a complete challenge project to find the perfect model to get higher precision using these social media data. There might be many use cases and machine learning research based on analyzing and time series prediction, however there is tremendous poor research directly in the field of bitcoin cryptocurrency. Our study and analyses do fill this gap by developing and training our model with previous price and sentiment for successful future price prediction

We have downloaded Jan 2020 to September 2020 Twitter tweets data, surmounts to 4.5 million tweets per day, bitcoin related tweets based on our keywords - 'BTC', 'BITCOIN', 'ビットコイン', '比特币', '比特幣' are extracted, which are further subjected to vader sentiment analysis which generates polarity average scores of each tweet row and can generates actual value of each tweet, it can state whether the user tweets is positive or negative or neutral, allowing us to study what the user wants to convey, based on our analysis there is exist a relation between sentiments and bitcoin prices, which are algorithm has been successful in capturing.

This paper is as follows. Section II shows the related work performed by other researchers. The methodology and analysis phase of the proposed model will be shown in Section III and Section IV, respectively. Finally, Section IV concludes our work.

2 RELATED WORK

There is no actual source of information when it comes to cryptocurrency, because of that everything can be used. This means news media and social media. Among those that dominates the others is social media. And among social media twitter is a very strong source of information and it is being widely used. Besides Twitter there are many sources of information that have been used to make sentimental analysis to see the future of the price graph and to predict the volatility in all cryptocurrencies and mainly in Bitcoin.

Main goal for the researchers is to figure out the effects of social media when it comes to predicting the Bitcoin prices and the trading volume in a short time frame like one day or one week, and also a long 30 to 90 days span[3].

Moreover, most cryptocurrency investors are having the tendency to overreact to news about the market trend. When it happens the market moves initially corresponding to the sentiment and then it is being slowly fixed over time.There are some examples and many experiments in which twitter sentiment analysis were used in the price prediction trend changes for the Bitcoin and other cryptocurrencies. By using SVM and various regression models, Twitter sentiment analysis was conducted by Georgoula, utilizing a Support Vector Machine (SVM) for predicting the price changes of cryptocurrency. As a result of this study % 89.6 accuracy was obtained and it was also found a short-term relation between the price of Bitcoin and positive Twitter sentiment. There is another example of a research performed by Pagolu. In that research social media was used microblogging to predict stock market prices. Depending on the result, It accurately reflected the common idea and the general opinions of the people about the current events.

3 PROPOSED TECHNIQUES

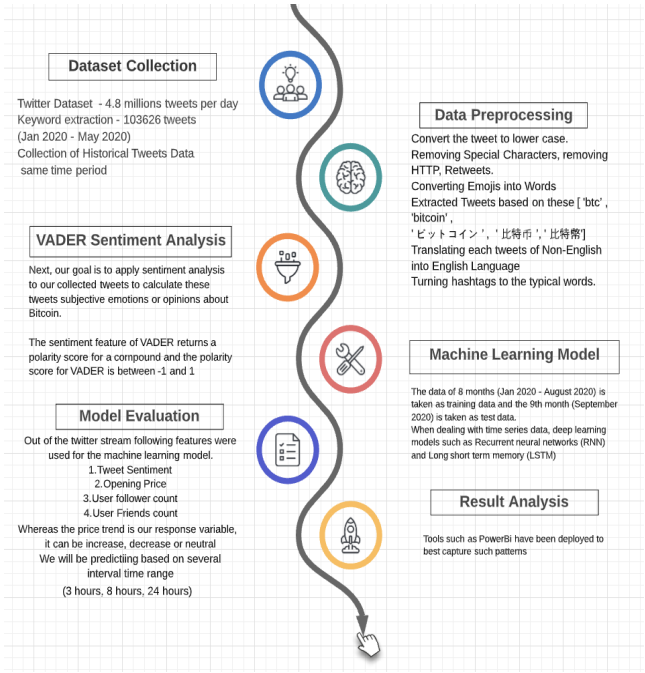

Figure 1 : Proposed Steps of project

Figure 1 states the proposed steps and sections are as followed. Data collection, data pre-processing, Vader Sentiment Analysis, deep learning model and evaluation.

3.1 Data Preprocessing

Once we get the sentiment score of each preprocessed tweet, we place timestamps on the bitcoin prices and sentiments this creates a time series . Various preprocessing steps have utilized vader sentiment analysis for all of our tweets. Starting from Jan 2020 - September 2020, tweets were collected and preprocessed. Within the time period of Jan 2020 - September 2020, per minute interval the tweets volume that create the dataset are 107320. The reason for the selection of the following time period is that a lot has occurred during this time, high rise and high falls, good news and bad news. Artificial

intelligence algorithms have proved to work when a certain pattern or trend occurs, Data has been resampled per month instead of minutes for better representation. Figure 2 represents bitcoin prices movements over the time period.

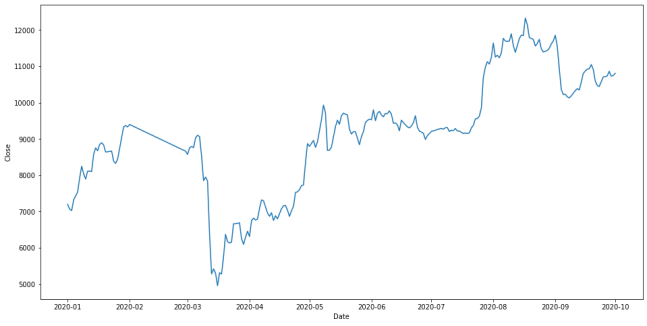

Figure 2 : Price Movements of Bitcoin (Jan 2020 - Sep 2020)

The dataset consists of combinations of expressions, emoticons, symbols, URLs. Specific features like retweets, emoticons, user references. To be easily learned by different classifiers, we have worked and preprocessed our dataset and created our own structured data based on all preprocessing steps. Raw twitter data requires to be structured in such a manner that is learnable by different classifiers, for data standardization and scaling following pre-processing steps are employed.

We have implemented MySQL Database for storing csv files and query analysis. Used spark, pyspark, pandas, numpy and multiprocessing for pre-processing based on keywords. Used emoji, TextTranslate, NLTK Vader Sentimentanalyzer, regular expression and Removed slang words to clean tweets.

These are the most important steps for data preprocessing so as to train our model more efficiently and after researching and working on a test sample of 20 rows, we applied these steps on overall data.

- Convert the tweet to lower case.
- Removing Special Characters, removing HTTP, Retweets.
- Converting Emojis into Words
- Extracted Tweets based on keywords ['btc' , 'bitcoin' , ' ビットコイン ' , ' 比特币 ', ' 比特幣']
- Translating each tweets of Non-English into English Language
- Turning hashtags to the typical words.

3.2 Sentiment Analysis

In order to obtain tweets sentiment or emotions, we have deployed a Vader sentiment analysis algorithm. As data was extracted from a social media platform, we got more positive sentiments. The output of The vader algorithm is a polarity score[4]. This score has a range between - 1 and 1, moving from -1 to 0 there are negative sentiments and in the middle, 0 has neutral sentiments and above that in between 0 to 1 we have positive emotions or sentiment. We have used the NLTK Vader Sentiment analyzer to calculate compound scores based on the words. Score above 0.05 is considered as positive sentiment and score below 0.05 is considered as negative sentiment and rest neutral.

3.3 Data Preparation

Data which is being dealt with is a time series data which is ordered by time stamps. In order to feed the data into the model it needs to be converted to a supervised dataset, before that few preprocessing steps have been performed, It is found that some samples are missing from per minute interval data, thus data is resampled at 8 hour intervals. In the next step we proceed as follows: we take data of three consecutive timestamps as our features and the fourth one has our response and similarly so on. Each feature at a time stamp has all the data of that particular timestamp. Thus a dataset is prepared. Figure 3 represents the data preparation

Feature 1	Feature 2	Feature 3	Feature 4	TARGET
2020-09-01 00:00:00	2020-09-01 08:00:00	2020-09-01 16:00:00	2020-09-02 00:00:00	2020-09-02 08:00:00
2020-09-01 08:00:00	2020-09-01 16:00:00	2020-09-02 00:00:00	2020-09-02 08:00:00	2020-09-02 16:00:00

Figure 3 : Data Preparation of every 8 hours

3.4 Model Selection

The data of 8 months is taken as training data and the 9th month is taken as test data. When dealing with time series data, deep learning models such as Recurrent neural networks (RNN) and Long short term memory (LSTM) are preferred, since these models deal with data sequentially, and possess memory enabling us to predict the future using the past, that is our requirement. Hence the selection.

Following features have been utilized.

1. Opening Price
2. Closing Price
3. User follower count
4. User Friends count
5. Sentiment score

3.5 Data Inference and Visualization

As mentioned there exists a correlation between the bitcoin closing price and sentiments, our inference based on our dataset plots suggests that as soon as the public's negative sentiment increases bitcoin prices start falling. One deduction from the data is that when bitcoin falls most steeply, members with huge twitter followers become active and induce positive sentiment into twiters, results can be seen as an increase in price.

Tools such as PowerBi have been deployed to best capture such patterns. Figure 4 represents the plots showing correlation.

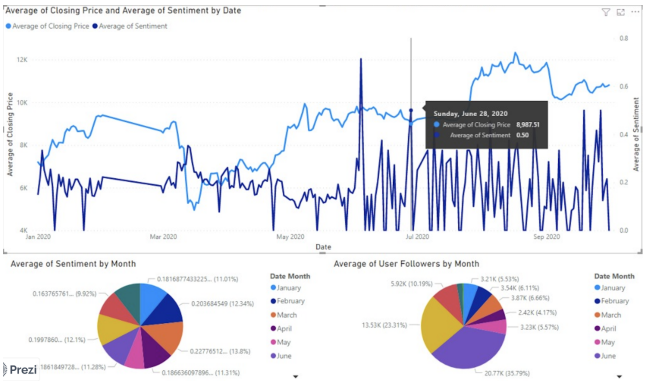

Figure 4 : Data Visualization from January - September 2020

The dataset we had at the start had tweets per minute and the number of tweets were too many and some of the data was either duplicated or missing for some minutes. Also to graph these datasets we need to have a unique date. So to tackle this issue we had to reassemble the dataset from minutes to hours then weeks and finally to per day. So finally we had an average price for that day and average negative sentiment for that day to graph. Using Power BI features such as date hierarchy the data was grouped in months and we built a dashboard that gave us the correlation between the sentiments and the price for each month and we were able to show the pattern between the two. For instance, if we look at the graph of May, we can see that the price has decreased when the negative sentiment has increased. In May, we saw the halving event for the bitcoin as the rewards for Bitcoin miners were reduced to half and the negative sentiment in peoples tweet was increased and a sharp decline was seen in the bitcoin prices. We can see such patterns throughout our dataset. The pattern was also confirmed with the news and a major price change was noticed around that time in real time.

Now while comparing the dataset of May to January, March & April, see that the graph is showing us the same pattern as the May graph i.e the inverse relationship between the closing price and the negative sentiments. The price has increased whenever the negative sentiments have decreased.

This result was achieved by taking the closing price for a particular day, like for instance the bitcoin price reached the highest price for a day and also to the lowest price and then closed in between. So our model will calculate the average of the price for the whole day and will give the value for that day. So we have a unique value for closing price for the whole dataset. To keep in mind, one must know that it is assumed that whenever the negative sentiment is decreased or touched zero it means that the positive sentiment has increased.

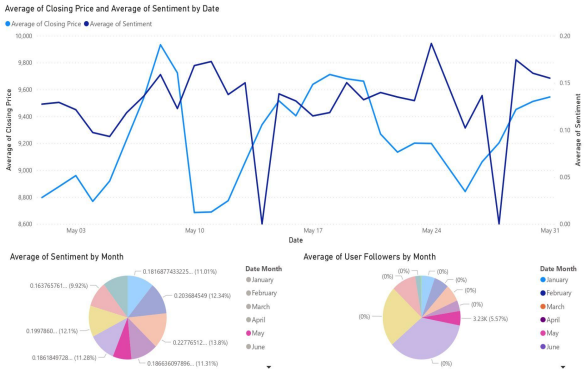

Figure 5 : Data Visualization of May

The graph also shows the Average User Follower of the person who tweeted, which is a really important factor as it depicts that the tweets with the most followers are affecting the price and not just the other person's tweet is moving the price.

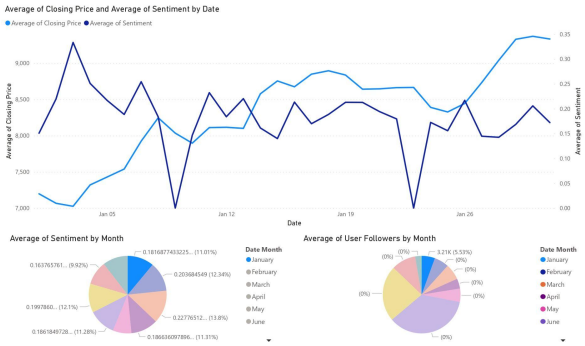

Figure 6 : Data Visualization of January

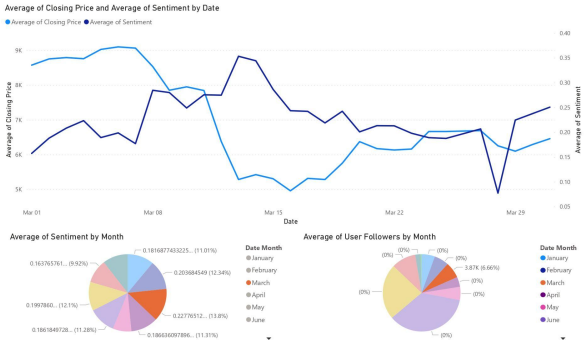

Figure 7 : Data Visualization of March

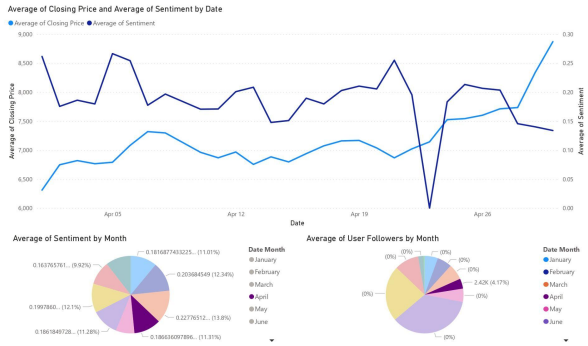

Figure 8 : Data Visualization of April

4 PRELIMINARY RESULTS

4.1 Model Evaluation

Deep learning models such as RNN and LSTM have been deployed, performance of RNN has been used as base line, LSTM has been observed to perform better, reason being RNN tendency to forget which is due to vanishing gradients

Metrics such as root mean squared error (RMSE) and Adjusted R square (R2) have been used to differentiate between the models in terms of their performance.

The R2 of RNN was found to be 0.35 as compared to 0.65 of LSTM. Models were trained over 60 epochs with a batch size of 64. Phenomenon of exploding gradient was observed in LSTM, which was improvised with the use of Monte Carlo dropout in every layer except the last one

4.2 Results and Conclusion

Figure 5 represents the final output of our model, the prices have been represented as moving averages. Blue one represents the original price whereas the cyan color represents the predicted price.

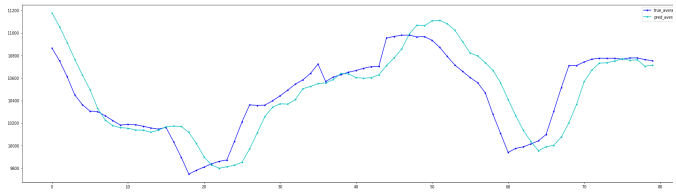

Figure 09 : Final output of LSTM Model

This is a fairly good prediction, despite the noise the model has proved to be effective. Thus our conclusion is that the bitcoin prices are highly influenced by the tweets, and if along with using the prices of the past, tweets are used, somewhat accurate predictions can be obtained.

5 AUTHOR CONTRIBUTIONS

We are 5 members in one team. And our tasks are divided as follows :

Vyoma Desai - Study and analyze Twitter Data Collection and Data Preprocessing.

Asim Javed - Data preprocessing, inference and implementation of time series model

Molan Zhang - Comparison of Linear Regression model and time series prediction

Ergin Bostanci and Zeeshan Qamar - Data Visualization using Power BI

5 REFERENCES

[1]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8701992

[2]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9351527

[3]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8404760

[4]Nakamoto S., “Bitcoin: A Peer-to-Peer Electronic Cash System”, 2008.

[5]Y. Kim, J. Lee, N. Park, “When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation”, PloS ONE 12(5), May 2017.

[6]Xie, Peng. “Predicting digital currency market with social data: Implications of network structure and incentive hierarchy”. Diss. Georgia Institute of Technology, 2017.

[7]Karalevicius, Vytautas, N. Degrande, J. De Weerdt. "Using sentiment analysis to predict interday Bitcoin price movements." The Journal of Risk Finance (2018).