Predicting Fluctuations in Cryptocurrencies' Price using users' Comments and Real-time Prices

Pavitra Mohanty[1], Darshan Patel[2], Parth Patel[3], Sudipta Roy[4]

[1,2,3]Department of Computer Science & Engineering, Ganpat University, Kherva, Mehsana, 384012, Gujarat, India
[4]Washington University in Saint Louis, MIR department, 510 South Kingshighway Blvd., Campus Box 8225, MO 63110, USA
[1]pavitramohanty111@gmail.com,[2]djsmarty99@gmail.com,
[3]parthcri1997@gmail.com, [4]sudiptaroy01@yahoo.com

Abstract: **This paper shows the prediction of fluctuation in the future price of cryptocurrencies. Users' comments and tweets from twitter using Apache Flume and Price data was fetched from exchanges. Bitcoin first documented by allies Satoshi Nakamoto, the first decentralized currency payment system has gained a considerable attention in the financial system, economics, social media and computer science due to its combination of peer-to-peer nature, encryption technology, and monetary unit. Predicting the price of Bitcoin and other cryptocurrencies is a great challenge because it is immensely complex and dynamic in nature. In this paper, we have tried to predict the future price of cryptocurrencies like Bitcoin using LSTM (Long Short-Term Memory) and used Twitter data to predict public mood. By combining both market sentiment and social sentiment because bitcoin price shows mixed properties. We also have selected some other important features from the blockchain information which has a major impact on Bitcoin's supply and demand and using them to train model that improves the predictive power of the future Bitcoin price. We have performed a deep study of how data from social media affect the price of Bitcoin and so we have included the twitter data in model training. Our model shows that how well LSTM predict the price of Bitcoin considering the high volatility. The precision given by our model is 60% and accuracy is 50%. More focus is not given to accuracy, in this case, considering the highly volatile market.**

Keywords: **Bitcoin, Cryptocurrency, Long Short-Term Memory (LSTM), Prediction Method, Recurrent Neural Network (RNN).**

I. INTRODUCTION

In this technical era, the new form of currency called digital currency, more popularly known as cryptocurrency has gained people's attention from the starting of the year 2017. Cryptocurrency is a decentralized and secure form of digital transactions which make the world more attracted towards this technology. As the name itself shows their characteristic of using encryption technique for the creation of coins, doing transactions. Among all the cryptocurrencies, Bitcoin has captured almost all of the people attention, this virtual currency was created in the year 2009 by an alias named Satoshi Nakamoto and it is the first decentralized (peer-to-peer) network cryptocurrency- an "electronic payment system based on cryptographic proof" and has gained a larger following from the media [1]. Bitcoin is built on the top of blockchain technology, as after the establishment it emerges as the leader of cryptocurrencies and it is increasing gradually. Unlike, traditional way of verifying the transaction by a trusted third party, this currency is based on cryptographic proof and protocols (i.e. a complex mathematic algorithm is used to verify the transaction and usually done by group of people known as miners) and no middleman is present while performing transaction, due to this it provides various advantages over traditional payment methods (such as Visa and MasterCard) like low transfer cost, anonymous transaction, typically fast transaction, more secure, irreversible transaction and it is not ruled over by a government, meaning the money is yours and not controlled by a central government body [2].

Due to the various advantages mentioned above and by the development of blockchain, the technology of distributed ledgers, known as the backbone of Bitcoin and other cryptocurrencies (after internet and cloud technology, blockchain is considered to be the next fruitful invention) due to such reasons, thus investor from all-around the globe sees more potential in such new emerging technology and consider it a safe investment. In the year 2017 according to CoinDesk, data approximately 5 billion dollars have been raised from 343 ICOs (It is the process of raising money by a cryptocurrency start-up firm through an Initial Coin Offering). And in the 2018 year up to date [1st April] 3 billion is raised by 92 ICOs and still more time remaining [3]. Bitcoin, Ethereum, Ripple has not affected various economic downturns and are accepted larger amount and these digital assets are used in various systems. For example, a ripple is a payment protocol based on blockchain which is accepted by various banks worldwide and used in daily operation. The spike up of Bitcoin value lays the foundation for its adoption. Ethereum, which works on smart contract used in the decentralized application (DAPP) [4].

Cryptocurrencies like Bitcoin has its own benefits as an international payment, but it is due to its volatile price it may still suffer from the problem of traditional currencies. So, it could be considered as a currency exchange rate. Since 2009, numerous cryptocurrencies have been developed, almost 1600 has been developed till 1 April 2018[5] and the dominance of Bitcoin is about 43%. Total market capitalization of all

978-1-5386-4692-2/18/$31.00 ©2018 IEEE

cryptocurrencies is nearly $260billion in US Dollar (as of March 2018). People are showing great interest see as a great opportunity for trading and trading of these coins made easy due to the launching of various exchanges like binance, bittrex, kucoin and many more. Economic and financial theory are not able to explain the high fluctuation in the price of bitcoin [6]. There are various factors such as interest rates and inflation doesn't exist as there is no central bank. Due to this Bitcoin is traded on different exchanges usually at some price differences, and correlation is seen against the US dollar and other major currencies. Apart from this all things the price somehow depends on the social platform like Twitter, Reddit, etc. By incorporating social media as an additional channel of information we can better emulate the speculative patterns of the traders. For example, the fall in prices of bitcoin with the Chinese regulator's decision to ban all virtual currency and its subsequent rise when the government decided to soften its stands. So, it becomes essential to consider the human reaction, feedbacks for a particular post in social media and gains insight from it which helps a lot while predicting the market. And crypto whales are posting all their signals for buying and selling so using these tweets the future price that whether it will fall or rise can be predicted. Our model is good enough which has a good precision of 60% and accuracy is 50%. More accuracy in this model may be not good in this volatile market but the precision is noticeable as the predicted data is 60% and rest analysis should be done to do more accurate trading.

II. METHODS

A. Data Collection

We used two different Dataset in our Project. Firstly, dataset1 consist of daily data like price and additional 26 feature about the blockchain of bitcoin and market, described in Table1. These data are collected from Blockchain Info [10]. One daytime series minimize the noise problem concerns from per minute volatility and this helps to determine which feature is good for predicting the bitcoin price. Features include information's like bitcoin total market cap as well as the bitcoin to US dollar conversion volume and vice-verse. The data needed to decide the human emotion is collected from Twitter as it provides API [11] to collect tweets. We have collected those tweets which contain hashtags #Bitcoin, #CoinListingNews, #cryptocurrency, #cryptopumpdump, #exchangehacked which is very much important because due to such news the price of bitcoin rise and fall according to the news and reaction of people to the news.

The second dataset consists of bitcoin price data with the interval of 10 second and 10 minutes. We collected 10 minute and 10-second data by developing an automatic real-time web scraper which uses Bitfinex API [12] over the period of multiple weeks. Bitfinex is a cryptocurrency trading platform based on Hong Kong.

Web scraper script runs on an instance of Amazon EC2 and the scrapped data is stored in a NoSQL database via Amazon Dynamo DB. Using this we have collected high-granularity bitcoin price data and around 70,000 price points are used in our Project.

B. Feature Selection

We have considered 26 independent features related to market and network of bitcoin. Out of these 26 features, we have selected 19 features. These features where chosen manually considering important of them and our own research. After performing forward stepwise selection and backward elimination methods, we selected which features may be best for our model.

TABLE I: SELECTED FEATURES WITH DEFINTION

Feature	Definition
Market capitalization	(Total amount of Bitcoin in market) * (current price).
Revenue of miners	(Reward per block) + (processing fees)
Number of Orphaned Blocks	blocks mined in a day
Number of TXN per block	Mean No. of transactions per block
Number of TXN	Sum of all uniquely done bitcoin transaction per day
Average Confirmation Time	Mean period to accept the transaction in block
Number of unique addresses	Number of unique Bitcoin addresses used per day
Total Bitcoin	Bitcoin mined till date
TXN Fees	Total BTC value of transaction fees miners earn/day
Block Size	The average size of the block
Hash Rate	Bitcoin blockchain network tera hashes
Trade Volume	USD trade volume from the top exchanges
Cost per transaction percent	Miners charge a small amount for a transaction
Difficulty	difficulty in finding a new block
Estimated Transaction Volume	Total volume of output
Ratio of Transaction and trade	Relationship between BTC transaction volume and USD volume.
Opening price	A 1st opening candle can be used to know the opening price
Closing price	the last candle of per hour,6 hours, day, month, a year can be used to know the closing price
Sentiments	Users' emotion related to the cryptocurrencies when some news comes.

Our dataset consists of these 19 variables collected daily over the course of the past 4 months. For training the model we have used 70% data of our dataset and the remaining 30% data is kept for testing purpose of the model.

C. Sentiment Analysis

Sentiment analysis is a method that uses natural language processing (NLP), text analysis and many other methods to identify attitude of the writer from the text with respect to the specific topic. It is also called as an Opinion Mining because the opinion of a person can be recognized from the text. Use of sentiment analysis is increased drastically for the data obtained from social media, surveys, and reviews. In our project, we used Sentiment analysis to determine person's opinion towards cryptocurrencies whether it is positive or negative from the tweets. If resultant sentiment is positive then price of cryptocurrencies is likely to increase and if the sentiment is negative, the price is going to decrease.

Word2vec algorithm

Word2vec is a group of models which are shallow, two-layer neural networks. Word2vec algorithm converts text into a vector of scores. Word2vec algorithm's input is a large set of data in text form and it produces a vector space. Each unique word from the input is assigned a vector in a space. Those word vectors are placed in a vector space as the words sharing common context are placed in close proximity to each other in vector space [13].

D. LSTM (Long Short Term Network)

LSTM [14] networks are a different version of RNN. LSTM networks consist of a cell, input gate, output gate and a forget gate. LSTM cells are memory blocks used to store a value. The value can be stored for a long period or short period of time. Forget gate is used for forgetting the irrelevant information i.e. removing the value from the cell state. The value is removed via multiplication of a filter. This helps to optimize the performance of an LSTM network. Forget gate is taking 2 inputs. i.e. hidden state and from the input and previous cell at that time. After multiplying the inputs with weight matrices and adding bias, the activation function is applied to the obtained result which is shown in equation II. Input gate is working in a similar way as of forget gate. Applying the sigmoid function to all the values from previously hidden layer values to current input values as shown in equation I. Now, we are creating a vector of all values which are possible from the hidden layers and from the current input and adding them to the cell state. tanh function is applied to them and hence the values are in the range from -1 to 1. Now, we are multiplying the value obtained from the sigmoid gate with the vector. The multiplied value is then added to the cell state. The output gate is giving an output of the relevant information which in our case is the predicted price. A filter is applied using a sigmoid function on the previously hidden values and current input and these values are passed to the next cell as hidden state values. The equations [15] I, II and III shown below are the equations of input gate, output gate and forget gate.

$$i = sigm(h_{t-1}U_i + x_tW_i) \ldots \text{I}$$

$$f = sigm(h_{t-1}U_f + x_tW_{f)} \ldots \text{II}$$

$$o = sigm(h_{t-1}U_o + x_tW_o) \quad \ldots \text{III}$$

$$f(x) = x \ldots \text{IV}$$

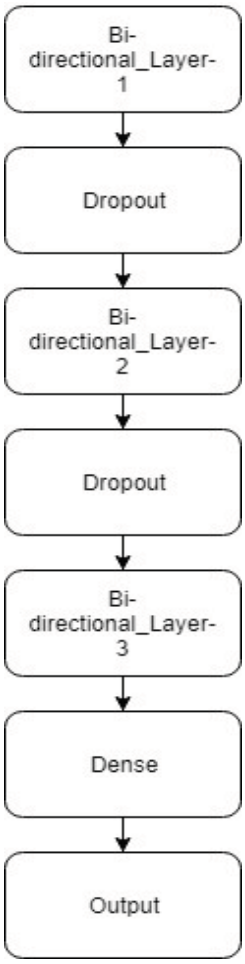

Fig. 1. Linear Stack Architecture

Linear stack architecture is used here. It means that layers are stacked above one other which is shown in figure 1. Hence, the data cleaned is passed into this linear stack architecture which consists of 3 layers of bi-directional LSTM. The first layer is the input layer and hence the pre-processed data is passed first to this layer. Second and third layers are the hidden layer. More features are extracted from the data due to the stacking of 3 LSTM layers. Also, dropout regularization method is used at first two LSTM layers. 20% Dropout is used here at first two layers. The activation function used is linear and its equation is shown above as IV. Adam optimizer is also used. The learning rate is 0.001. We have taken epochs equal to 100 because the

accuracy is increased due to more number of epochs. Also, if the numbers of epochs are more than 100 then computational time will increase. Our model is built in keras. The pseudo code of bi-directional LSTM is given below:

1. Forward pass for forward states from time 1 to T and backward states for time T to 1.
2. Do forward pass for the output layer.
3. Do backward pass for the output layer.
4. Do backward pass for forward states from time T to 1 and backward states from time 1 to T.
5. Update Weights.

Where forward pass includes resetting all activations to 0, taking inputs and updating the values of activations. The forward pass also includes storing all values of hidden layers and output activations at every timestamp. Backward pass includes resetting all partial derivations to 0and propagating the output errors backward. Also values of input gate, output gate and forget gate are updated as per the equations are shown above in I, II and III in the forward pass as well as backward pass.

III. RESULTS

Training the Model:

- The model was fitted to the training data which was in the batch size of 1024 because it is leading to a smaller number of iterations for training.
- 100 epochs were performed on training dataset to give time to the model to adjust its biases and weights.
- 70% of the whole dataset was used as training data and rest was used as a testing data.
- The model was trained by minimizing the loss of training data using mean squared error because it is one of the common measures used for evaluation.

Testing the Model:

- The model was given N values of testing data and computed normalized prices from it.

Comparison of Predicted data and Real data:
The confusion matrix is generally used to show the performance of a classifier on such data whose true values are known. Where TP (True Positive): TP is a case where classifier predicted true when they are actually true.TN (True Negative): TN is a case where classifier predicted false when they are actually false. FP (False Positive): FP is a case where classifier predicted true when they are false, it's also known as Type-1 error. FN (False Negative): FN is a case where classifier predicted false when they are true, it's also known as Type-2 error. The following table shows Confusion Matrix of predicted data

TABLE 2: CONFUSION MATRIX

Confusion Matrix	P'(Predicted)	N'(Predicted)	Total
P(actual)	True Positive=7239	False Negative =4871	12110
N(actual)	False Positive=4629	True Negative=3261	7890
Total	11868	8132	20000

Accuracy was found by an equation, $Accuracy = \frac{TP+TN}{total} = 0.5$

Precision was found by an equation, $Precision = \frac{TP}{TP+FP} = 0.6099$

Recall was found by an equation, $Recall = \frac{TP}{TP+FN} = 0.597$

F1-score was found by, $F1 - Score = \frac{2*precision*recall}{precision+recall} = 0.6033$

MSE (Mean Squared Error) was found by, $MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - Y^i)^2 = 0.043$

- Precision was used to calculate how many times it got true positive in comparison to how many times it got positive (addition of true positive and false positive)
- A recall is used to calculate how many times it got true positive in comparison to how many times it should give positive.
- F1-Score is a weighted average of recall and precision.
- Mean Square Error (MSE) is an average of squares of the difference between actual and predicted values. As the cryptocurrency market is more volatile keeping this in mind we have not tried to put more effort on perfect prediction as this might cause a loss to the user.

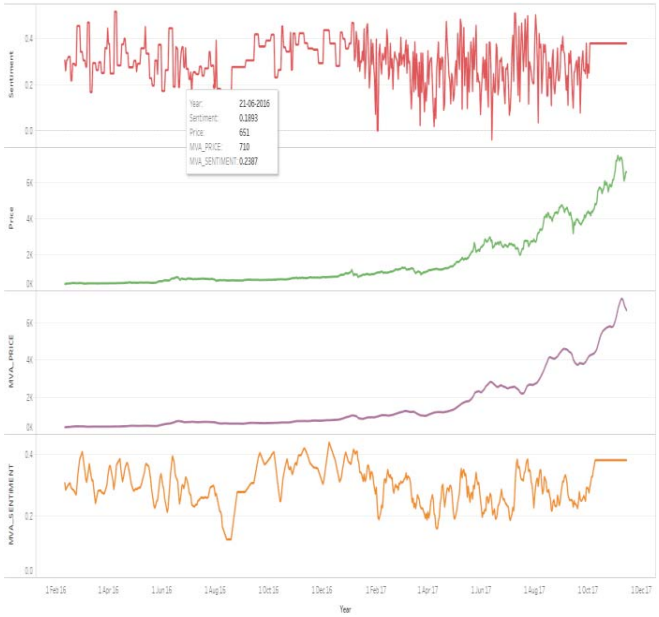

Fig. 2. Bitcoin price and Twitter sentiment on a different period of time

The above figure describes the graph of Bitcoin Price, Sentiment analysis from Twitter, moving average price (MVA_PRICE) and moving average sentiment(MVA_SENTIMENT) from February 2016 to December 2017.

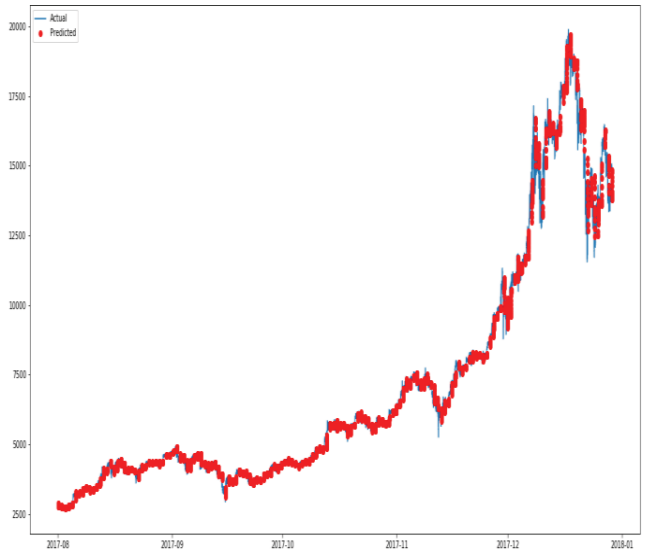

Fig. 3. Bitcoin price over time

Above graph represents the price of bitcoin over days, the x-axis denotes time while y-axis denotes price of bitcoin in USD. Actual price is represented using blue line and predicted price is represented using a red line. It denotes how our model predicts the price of future and as this gives the idea to the user about the market so using this they can trade carefully and more wisely. The huge dip is seen between the Nov 2017 to Dec 2018 and the reason behind this is the rumor going on that China government can ban bitcoin mining and this news has dropped the price of bitcoin from $7k to $4k and our model has very effectively predicted this.

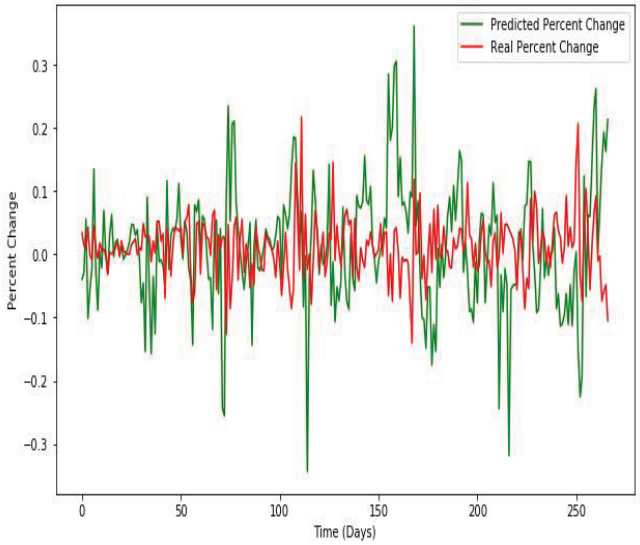

Fig. 4. Percentage change in price over days

Above graph represents a change of bitcoin price in percentage value per day. This is more useful when we want to do

intraday trading as the percentage change per day is shown. The above graph depicts the price in percentage versus time in days. Actual percentage change value is represented using red line and predicted percentage change value is represented using a green line.

IV. CONCLUSION

We have studied about the price fluctuations in cryptocurrency like bitcoin, etc. using the twitter sentiments of the users and other information taken from the bitcoin blockchain which are selected as features using feature selection methods. We have done the time series analysis using Bi-directional LSTM which is a type of RNN and the accuracy we have achieved is about 50% which is low though precision is 60% which means what our model has predicted is almost 60% true. As the cryptocurrency prices fluctuate more i.e. they are volatile so we have not focused much on the accuracy part or precise prediction of the prices as a loss can occur to the user. These results have shown that the model can be improved. Moreover, accuracy can be increased more by feeding more data to the model.

REFERENCES

[1] Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System", 2009.

[2] Stephen Chan, Jeffrey Chu, Saralees Nadarajah and Joerg Osterrieder, "A Statistical Analysis of Cryptocurrencies", Journal of Risk and Financial Management, 2017.

[3] O. Momoh, "Initial Coin Offering (ICO)", Investopedia, 2018. [Online]. Available: https://www.investopedia.com/terms/i/initial-coin-offering-ico.asp.

[4] "7 Reasons Why You Should Invest in Bitcoins, Cryptocurrencies and ICO's", Medium, 2018. [Online]. Available: https://medium.com/@fastinvest/7-reasons-why-you-should-invest-in-bitcoins-cryptocurrencies-and-icos-af032a03bc39.

[5] "All Cryptocurrencies | CoinMarketCap", Coinmarketcap.com, 2018. [Online]. Available: https://coinmarketcap.com/all/views/all/.

[6] Kristoufek L.,"BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era," Scientific Reports, 3: 3415, doi: 10.1038/srep03415, PMID: 24301322,2013

[7] "Bitcoin Block Explorer - Blockchain", Blockchain.info, 2018. [Online]. Available: https://blockchain.info/.

[8] "Docs", Dev.twitter.com, 2018. [Online]. Available: https://dev.twitter.com/rest/tools/console.

[9] "API Access", Docs.bitfinex.com, 2018. [Online]. Available: https://docs.bitfinex.com/docs/api-access.

[10] "Word2vec", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Word2vec.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Computation., 1997.

[12] Y. Gal, "A theoretically grounded application of dropout in recurrent neural networks", arXiv:1512.05287, 2015.