Short-Term Bitcoin Price Fluctuation Prediction Using Social Media and Web Search Data

Aditi Mittal, Vipasha Dhiman, Ashi Singh, and Chandra Prakash
Department of Information Technology
Indira Gandhi Delhi Technical University for Women Delhi, India

***Abstract*—In recent years, the social network has been widely used among the public to share their views and for communication as well. Further, sentiments of the text used in emails, blogs, and social media posts affect human decision making and behavior. Bitcoin being a decentralized and peer-to-peer cryptocurrency has attracted a large number of users on the web search and social media. The goal of this paper is to correlation among Bitcoin price and Twitter and Google search patterns. Linear regression, polynomial regression, Recurrent Neural Network, and Long Short Term Memory based analysis concludes that there is a relevant degree of correlation of Google Trends and Tweet volume data with the price of Bitcoin, and with no significant relation with the sentiments of tweets.**

***Index Terms*—Bitcoin, Google Trends, Twitter, Sentiment Analysis, RNN, LSTM**

I. Introduction

Cryptocurrency prediction and trading with machine learning is a new and upcoming area in the field of research. Machine learning has already been in use for stock market prediction and trading and has already been researched to a great extent. Similarly, various algorithms can be applied to the cryptocurrency in the market as well.

Bitcoin was the first successful cryptocurrency in the market. It is one of the widely used decentralized cryptocurrency that means it does not have any central authority like a bank. It was introduced through a paper in 2008 under the name of Satoshi Nakamoto [1]. It can be sent from the user to the user through a network known as the peer-to-peer (P2P) network. The number of Bitcoins circulated in the market is controlled by a special algorithm that was specified in the paper. The records of all the transactions of cryptocurrencies are posted in a public ledger known as Blockchain. Also, bitcoin is highly volatile.

Fig. 1 shows how the price has varied in all-time. This indicates that people use cryptocurrency not only as a currency but also for other purposes like a future investment. When it was first released in 2009, its price was less than one cent. After eight years in 2017, it reached an all-time high of around $20,000! Researchers at the University of Cambridge estimated that there were around four million users using a cryptocurrency with most of them using bitcoin in 2017 [2].

A machine learning technique can be used to model and predict whether the price will increase or decrease for the currency, in this case, Bitcoin, that needs to be traded in the near future. Based on the prediction, decisions about buying or selling can be made to earn the maximum profit. Bitcoin prediction will help the investors to do effective investments of their money. The model will indicate at what time to withdraw the investment and at what time to make a new investment.

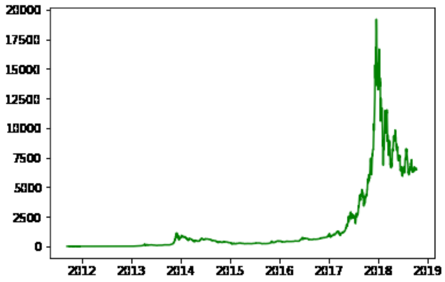

Fig. 1. Variation in Price of Bitcoin from its inception

It was observed that when there is negative news published about Bitcoin, more people sell Bitcoin instead of buying it. This causes a decrease in price. When there is positive news for Bitcoin, people buy Bitcoin more than the people selling it. This cause an increase in price.

In this work machine learning techniques (Linear Regression, Polynomial Regression, Recurrent Neural Network (RNN) and Long Short-term Memory (LSTM)) are explored to find the relation of the price of Bitcoin with polarity of tweets, the volume of tweets and number of searches in Google trends. These models are evaluated for the prediction of Bitcoin price for the near future.

II. LITERATURE SURVEY

This section presents the previous studies in an area of price prediction of Bitcoin. There were several attempts to analyze the Bitcoin transaction network.

Madan, Saluja and Zhao [3] applied machine learning techniques for the prediction of the price of Bitcoin. Their dataset had 25 optimal features related to Bitcoin for each day of five years. The accuracy was found out to be 98.7%. They also performed the prediction of the price of Bitcoin by using only the previous price as the feature at 10 min interval, and the accuracy was found out to be in the range of 50-55%.

Abraham, Higdon, Nelson and Ibarra [4] gave a method for predicting the changes in the price of Bitcoin and Ethereum by using Twitter and Google data. Through the tweets, they founded that Tweet volume and not the sentiment is a better predictor for the price direction of whether it will increase or

978-1-7281-3591-5/19/$31.00 ©2019 IEEE

decrease. Tweets and Google data were used as the input to the linear model for prediction of the direction of change in price. It was found out the Google Trends and tweet volume was correlated with the bitcoin price with correlation value as 0.817 and 0.841 respectively.

Kaminski [5] analyzed the relation between the Bitcoin price and the Tweets sentiments. The dataset was collected from 23rd November 2013 - 7th March 2014 with 160,000 tweets with bitcoin keyword and concluded that the Bitcoin price depends on sentiment of tweets for a smaller period and requires other factors to be considered for the Bitcoin price prediction.

Georgoula, Pournarakis, Bilanakos, Sotiropoulos and Giaglis [6] applied time-series analysis to model the relation between Bitcoin price and economic factors and sentiments derived from Twitter posts. Application of regression on a short window concluded that the Twitter sentiment and Wikipedia has shown a positive correlation with the price of Bitcoin. Also, the Bitcoin price was affected by the exchange rate of the US dollar and the euro.

McNally, Roche and Caton [7] ascertained with what accuracy can the price of Bitcoin be predicted by the use of machine learning techniques. The price data were collected from the Bitcoin Price Index. They implemented Bayesian optimized RNN along with the LSTM network with the highest accuracy of 52% for LSTM. The ARIMA model was also implemented for comparing it with the deep learning models which performed poorly.

Jang and Lee [8] analyzed a time series of the Bitcoin process and also performed Bayesian neural networks(BNN) algorithm and SVM on it. The dataset was extracted from the bitcoin charts and blockchain.info. The work showed that BNN performed well for prediction of the Bitcoin price and was also able to explain the reason for high volatility in its price.

Guo and Fantulin [9] focused on making the short-term prediction of the fluctuations in the price of Bitcoin. The data was collected from September 2015 to April 2017 and was divided into 70% training, 20% testing along with 10% validation datasets. The order book data was collected from OKCoin. Experiments were performed using various statistical and machine learning approaches like exponential weighted moving average approach, generalized auto regressive conditional heteroskedasticity model, structural time series model, etc. and then the results were compared. The MAE and RMSE were used to evaluate the performance of the models. It was found out that ensemble method extreme gradient boosting and regularized elastic-net better performed than the other methods.

Mern and Anderson [10] tried to construct a model to predict the next-day trading price of Bitcoin by using data extracted from market, trends, network and some other economic indicators like Dow Jones Industrial Average, Volatility Index, Brazilian currency index, etc. Logistic regression, SVM and convolution neural network were applied on the dataset. The validation accuracy(96.7%) of convolution neural network(CNN) was found out to be the best among the three techniques.

III. Data Collection

A. Bitcoin Data

The data of Bitcoin was collected from bitcoincharts [11]. The dataset has three attributes, namely- Timestamp, Weighted Price and Volume of Transactions. Every record has data of Weighted Price and Volume of the transaction for every 1-minute from 9 April 2014 till 07 January 2019. The data was aggregated to get the average price of Bitcoin for each day.

B. Tweets from Twitter's API:

Tweepy, a library in Python, was used for accessing the API to extract the data. Tweets were collected using only the #bitcoin hashtag which provided a large number of tweets. Also, retweets were not collected to avoid the problem of redundancy. For each post, the post ID, the time stamp, and the text were collected. Tweets were also filtered based on the language; that is, only English language tweets were used for the analysis. The tweets were collected from 16th December 2017 till 20th February 2019. The final dataset was made up of around 7.5 million tweets.

Tweets have characters and symbols that do not give any relevant information like hashtags, links, and emoticons. Preprocessing includes removal of capital letters so that words with the same meaning doesn't have different impact due to capitalization. For this, preprocessing packages for Python and regular expressions were used. Regular expressions are used to find the text with similar patterns. They were used to remove the hashtags, emoticons, and links. For VADER, if the compound score is larger than 0.05 the polarity of the tweet will be considered as positive, and if the compound score is smaller than -0.05, the polarity of the tweet is considered to be negative. In all other cases, the text is taken as neutral.

C. Tweet Volume

Tweet Volume means the number of tweets in a particular period. In this work, data is analyzed for each day. Online data at www.bitinfocharts.com contains data about the number of tweets done per day about Bitcoin by the users from 9 April 2014 to 07 January 2019, having 1735 rows.

D. Google Trend Data

Google provides search data from 2004 until the date specified. The data is in the form of Search Volume Index (SVI) instead of Search Volume. Google trends do not provide daily data for a period greater than 90 days. For a period greater than three months, Google trends provide month wise data. For querying data more than three months, adjustments according to Erik Johansson method are done. The Google trends data collected has 1735 columns which contain data from 9 April 2014 to 07 January 2019.

The adjustments in the dataset provided by Google trends is made according to the Erik Johansson method [12]. The first step involves the collection of data into parts of 3 months in continuity, and then combine all the data. The second step is to arrange the data for the same period but combined for each month to get the monthly SVI. Now, the determination of the adjustment factor is performed by dividing the monthly SVI with the daily SVI where the dates are actually overlapping. The last step involves the multiplication of adjustment factor and the daily SVI.

IV. Machine Learning Techniques

Machine learning algorithms have been implemented to find the correlation among the bitcoin price and Google trends and Tweet volume data pattern.

A. Linear Regression

Linear regression is a machine learning techniques that establishes a linear relationship of a dependent variable with one or more independent variables [13]. Linear regression with one independent factor is known as simple linear regression as shown in 1, and with more than one independent factor, it is known as multiple linear regression. The goal of linear regression is to determine a linear model which can be used for prediction of the price of Bitcoin using the three factors - Tweet sentiment, Google Trends, and Tweet Volume. The model is trained with 70% of the data and tested on the remaining data.

$$Y = A_0 + A_1 x \tag{1}$$

The values of A_0 and A_1 should be chosen that gives the least error. There are various kinds of error metrics that can be used to evaluate the model. If the sum of squared error is considered as the metrics to evaluate the model, then the following formula is used to calculate the error.

$$Error = \sum (actual\_output - predicted\_output)^2 \tag{2}$$

$$A_0 = \bar{y} - A_1 \bar{x} \tag{3}$$

Once A_0 is calculated, then A_1 is calculated using the following formula

$$A_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{4}$$

B. Polynomial Regression

It is a type of regression analysis which shows the relationship among non-linear dependent variable B and independent variable A derived in the n-th degree of a polynomial in A [14]. Polynomial regression shapes a non-linear model to the data, in the form of statistical estimation problem it can be seen as linear because the regression function E(B|A) is linear. That is the reason, why polynomial regression is taken as a special type of multiple linear regression. We can model the expected value of B in the form of an n-th degree polynomial, yielding the polynomial regression model as follows:

$$B = \beta_0 + \beta_1 A + \beta_2 A^2 + ... + \beta_n A^n + \varepsilon \tag{5}$$

The polynomial regression model

$$B = \beta_0 + \beta_1 A_i + \beta_2 A_i^2 + ... + \beta_m A_i^m + \varepsilon (i = 1, 2, 3...n) \tag{6}$$

can be represented in the form of matrix **X**, $\vec{\beta}$ - a parameter vector , $\vec{A}$ - a response vector and $\vec{\epsilon}$ - a vector of random errors. The i^{th} row of $\vec{A}$ and **X** will have the A and B value of the i^{th} sample data.

C. Recurrent Neural Network

Recurrent Neural Network is a generalization of feed-forward neural network that has an internal memory [15]. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it has learned from the previous one input.

in_t is the input given in the neural network at time instance t and h_t is the hidden state at any time t. h_t is calculated using the following equation:

$$h_t = s(Uin_t + Wh_{t-1}) \tag{7}$$

s(x) is the activation function which can be sigmoid, tanh, stepwise etc.. h_{-1} is initialized with zero. O_t is the output produced at time t which can be calculated as:

$$O_t = s(Vh_t) \tag{8}$$

D. Long Short Term Memory

Long Short-Term Memory (LSTM) networks are a modified version of recurrent neural networks, which makes it easier to remember past data in memory [16]. The vanishing gradient problem of RNN is resolved here. LSTM trains the model by using back-propagation. In an LSTM network, three gates are present:

1) input gate - discover which value from input should be used to modify the memory.
2) forget gate - discover what details to be discarded from the block.
3) output gate - the input and the memory of the block is used to decide the output.

The data which needs to be discarded from the current cell state is identified using the following equation:

$$f_t = \sigma(\, U_f[\, s_{t-1}, y_t] + v_f) \tag{9}$$

The new data that is to be saved into the cell state is determined.

$$i_t = \sigma(\, U_i[\, s_{t-1}, y_t] + v_i) \tag{10}$$

$$C'_t = \tanh(\, U_C[\, s_{t-1}, y_t] + v_C) \tag{11}$$

Then the data about the old subject is dropped, and new data is added.

$$C_t = f_t * C_{t-1} + i_t * C'_t \tag{12}$$

Finally, the output is determined using the following equations:

$$o_t = \sigma(\ U_o[\ s_{t-1}, y_t] + v_o) \tag{13}$$

$$s_t = o_t * \tanh(\ C_t) \tag{14}$$

where, s_{t-1} is previous output, v is bias, U is the weight s_t is final output of current cell, o_t is the output, C_{t-1} is previous candidate cell state, C_t is the new candidate that can be added to the current cell state. and y_t is the current input.

V. Results

The performance of forecasting is evaluated by accuracy of prediction of increase or decrease in price, accuracy with margin as $500, $100, $50 and $25.

A. Linear Regression

1) Google Trends and Tweet Volume : Table I shows the R2 Score and Pearson R correlation coefficient obtained for the three parameters using the linear regression model. Figure 2 shows that there is a significant degree of correlation of Google trends SVI data with the Bitcoin price, with Pearson R-value as 0.79 and p-value as 0. The R2 score was found out to be 0.755. Figure 3 shows the actual and predicted price using linear Regression and Google trends.

TABLE I
Linear Regression results

Parameter	R2 Score	Correlation Coefficient
Tweet Volume	0.690	0.740
Google Trends	0.755	0.790
Tweet Sentiments	0.049	-0.300

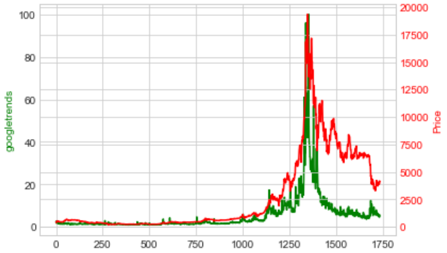

Fig. 2. Bitcoin Google Trends SVI and Prices

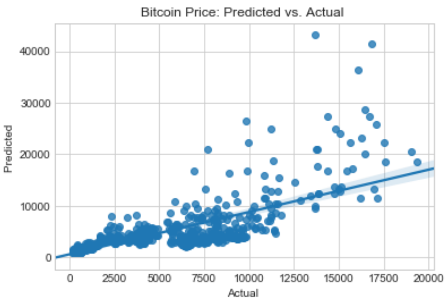

Fig. 3. Linear Regression Result for Bitcoin Google Trends SVI

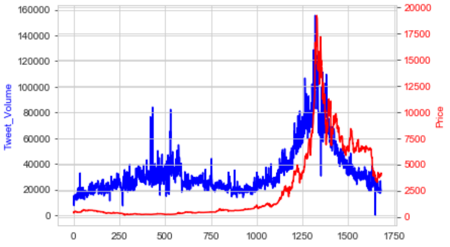

Fig. 4. Bitcoin tweet Volume and Prices

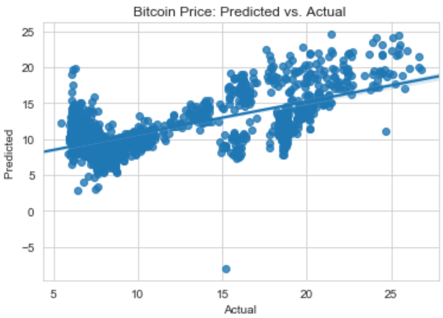

Fig. 5. Linear Regression Results for Bitcoin Tweet Volume

Figure 4 shows a significant correlation of Tweet Volume with the Bitcoin price with Pearson R correlation coefficient as 0.74 and p-value of 0. The R2 score value was found out to be 0.69.

Figure 5 shows the result of applying linear regression to the Tweet Volume data of Bitcoin.

2) Sentiment Analysis on Collected Tweets: Sentiment analysis is the process of understanding an opinion in a particular statement whether in written or in spoken form [17]. A large number of tweets are made using bots with different user names. Still, those tweets will influence peoples decision to buy or sell the Bitcoin as they may have positive or negative emotions, and therefore impact and fluctuate the price of the Bitcoin. Also, a large number of tweets are neutral that is they don't have any sentiment and instead are only facts which do not influence the people's decision about Bitcoin.

In 6 which shows the graph between the price of Bitcoin and the Average Polarity of the tweets in a day, there is no clear relationship. This shows that the correlation between price and the sentiment of tweets is not satisfactory and won't affect the price of Bitcoin to a significant extent.

With the sentiments polarity results are given by VADER, the average of compound scores of all the tweets for each day was considered. The dataset was cross-validated with 70% training and 30% testing sets. The figure 7 shows the result of applying linear regression to the sentiment of tweets and the price of Bitcoin with Pearson-R correlation coefficient as -0.3 and R2 Score value as 4.9.

This shows that Tweet Sentiment has poor correlation with bitcoin price as the average polarity of all tweets are summing up to positive irrespective of the price of bitcoin. The possible

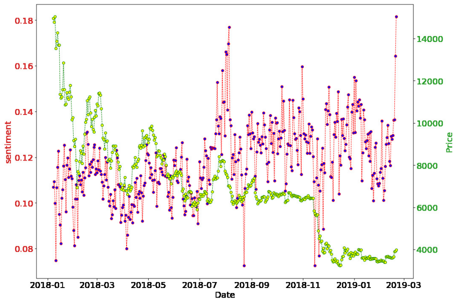

Fig. 6. Bitcoin Price and Daily Average Polarity of Tweets

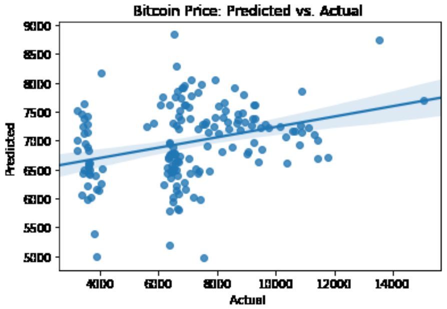

Fig. 7. Linear Regression Results for Tweets Average Polarity

reason for the same can be that people might be interested in other technologies of bitcoin like privacy, blockchain, etc.

B. Polynomial Regression

Here, the best-fit curve of different degrees was modeled and compared. It was observed that the R^2 score at higher degree is better than lower, but after a saturation point it started decreasing again. Fig 9 and 8shows the graph of bitcoin price predicted with twitter volume and Google trends at different degrees of the polynomial. Table II shows the results of predicting the bitcoin price based on Google Trends and Tweet Volume using Polynomial Regression.

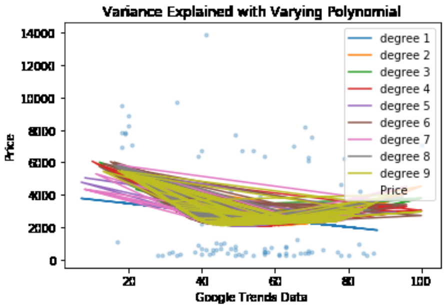

Fig. 8. Polynomial regression with price vs Google trends

C. Long Short Term Memory & Recurrent Neural Network

RNN and LSTM require various parameters such as batch size, number of units and epochs. With different parameters, they give different results. The models were trained using 70% of data and tested using remaining data. Table III shows the values of the various parameters for RNN and LSTM that gave the best accuracy. Table IV shows the results of predicting the price based on Google Trends and Tweet Volume using RNN and LSTM.

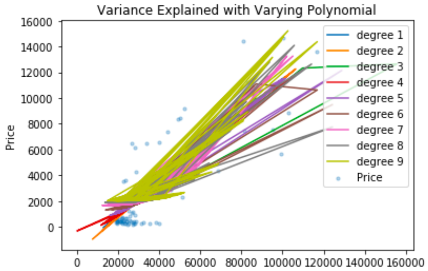

Fig. 9. Polynomial regression with Price vs Tweet volume

TABLE II
POLYNOMIAL REGRESSION RESULTS

Measure	Google Trends	Tweet Volume
Accuracy	66.66%	77.01%
500$ margin acc.	6.89%	10.34%
100$ margin acc.	3.44%	1.14%
50$ margin acc.	2.29%	1.14%
25$ margin acc.	2.29%	1.14%

TABLE III
RNN AND LSTM PARAMETERS USED

	RNN		LSTM	
	Trends	Volume	Trends	Volume
Activation Function	sigmoid	sigmoid	sigmoid	sigmoid
Units	5	4	6	4
Learning Rate	0.001	0.01	0.01	0.01
Epochs	70	60	80	70
Batch size	5	5	4	6

Fig 10 and 11 shows the graph between the actual price and the price of Bitcoin predicted based on the Google Trends and Tweet Volume respectively using RNN. Fig 12 and 13 shows the graph between the actual price and the price of Bitcoin predicted based on the Google Trends and Tweet Volume respectively using LSTM.

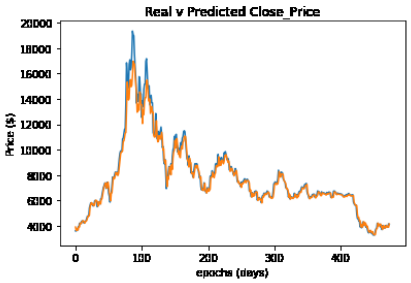

Fig. 10. Actual and Price of Bitcoin predicted using Google Trends & RNN

TABLE IV
RNN AND LSTM RESULTS FOR GOOGLE TRENDS AND TWEET VOLUME

	RNN		LSTM	
Measure	Trends	Volume	Trends	Volume
Accuracy	62.45%	53.46%	50.00%	49.89%
500$ margin acc.	83.33%	80.80%	82.62%	80.93%
100$ margin acc.	33.75%	32.7%	37.71%	30.50%
50$ margin acc.	18.14%	17.51%	22.24%	15.46%
25$ margin acc.	8.86%	8.5%	11.86%	6.35%

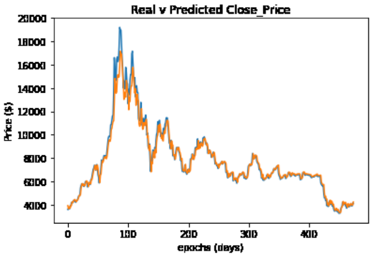

Fig. 11. Actual and Price of Bitcoin predicted using Tweet Volume & RNN

D. Effect on the performance without Google Trends and Tweet Volume

For the prediction of bitcoin price using only past price, RNN, LSTM, ARIMA models were evaluated on daily, weekly, monthly and yearly data. For daily data, accuracy for predicting direction of change in price for RNN and LSTM was found out to be 43.78% and 42.98% respectively. ARIMA gave the worst accuracy of 38.02%. The results got poor with aggregation due to smaller dataset. Table V shows the comparison of the three techniques using day wise data.

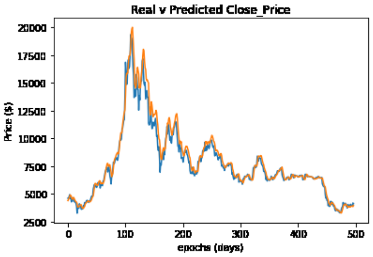

Fig. 12. Actual and Price of Bitcoin predicted using Google Trends & LSTM

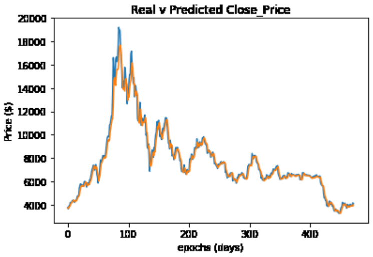

Fig. 13. Actual and Price of Bitcoin predicted using Tweet Volume & LSTM

TABLE V
COMPARISON OF RNN,LSTM AND ARIMA ON DAILY DATA

Measures	**RNN**	**LSTM**	**ARIMA**
Accuracy	43.78%	42.98%	38.02%
500$ margin acc.	35.23%	30.60%	28.020%
100$ margin acc.	25.56%	23.98%	21.10%
50$ margin acc.	23.21%	21.43%	12.32%
25$ margin acc.	32.45%	31.75%	20.47%

VI. CONCLUSION AND FUTURE WORK

Among tweet volume, Google trends and tweet sentiments, tweet sentiment analysis has shown the worst results. After applying the algorithms - LSTM, RNN, Polynomial regression - on tweet volume and Google trends, the accuracy of direction of bitcoin price is predicted with accuracy 77.01% and 66.66% of polynomial regression with Tweet Volume and Google trends respectively. But, for better accuracy, other factors such as Wikipedia search and Facebook posts can be taken into account. Also, the factors were implemented separately and can be combined together to check the overall accuracy.

REFERENCES

[1] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” 2008.

[2] G. Hileman and M. Rauchs, “Global cryptocurrency benchmarking study,” *Cambridge Centre for Alternative Finance*, vol. 33, 2017.

[3] I. Madan, S. Saluja, and A. Zhao, “Automated bitcoin trading via machine learning algorithms,” *URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan*, vol. 20, 2015.

[4] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, “Cryptocurrency price prediction using tweet volumes and sentiment analysis,” *SMU Data Science Review*, vol. 1, no. 3, p. 1, 2018.

[5] J. Kaminski, “Nowcasting the bitcoin market with twitter signals,” *arXiv preprint arXiv:1406.7577*, 2014.

[6] I. Georgoula, D. Pournarakis, C. Bilanakos, D. Sotiropoulos, and G. M. Giaglis, “Using time-series and sentiment analysis to detect the determinants of bitcoin prices,” *Available at SSRN 2607167*, 2015.

[7] S. McNally, J. Roche, and S. Caton, “Predicting the price of bitcoin using machine learning,” in *Parallel, Distributed and Network-based Processing (PDP), 2018 26th Euromicro International Conference on*. IEEE, 2018, pp. 339–343.

[8] H. Jang and J. Lee, “An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information,” *IEEE Access*, vol. 6, pp. 5427–5437, 2018.

[9] T. Guo, A. Bifet, and N. Antulov-Fantulin, “Bitcoin volatility forecasting with a glimpse into buy and sell orders,” in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 989–994.

[10] J. Mern, S. Anderson, and J. Poothokaran, “Using bitcoin ledger network data to predict the price of bitcoin.”

[11] “Bitcoin (usd) price,” https://www.coindesk.com/price, accessed: Jan 7, 2019.

[12] “Erik johansson method,” http://erikjohansson.blogspot.com/2014/12/creating-daily-search-volume-data-from.html, accessed: Jan 7, 2019.

[13] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2012, vol. 821.

[14] J. D. Opsomer and D. Ruppert, “Fitting a bivariate additive model by local polynomial regression,” 1997.

[15] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, “Recurrent neural network based language model,” in *Eleventh annual conference of the international speech communication association*, 2010.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] B. Pang, L. Lee *et al.*, “Opinion mining and sentiment analysis,” *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.