A Study of Opinion Mining and Data Mining Techniques to Analyse the Cryptocurrency Market

Akhilesh P Patil
Department of Computer Science and Engineering
Ramaiah Institute of Technology
Bengaluru, India
apatil1997@gmail.com

Akarsh T S
Department of Computer Science and Engineering
Ramaiah Insitute of Technology
Bengaluru, India
ts.akarsh@gmail.com

Dr. Parkavi A
Department of Computer Science and Engineering
Ramaiah Institute of Technology
Bengaluru, India
parkavi.a@msrit.edu

***Abstract*—The value of various Cryptocurrencies such as Bitcoin, Litecoin, Ethereum are always elusive. Hence, it would be a great value addition to investors if a model is able to predict what would be the nature of the crypto market for the next day. Through this paper, a time-series model using Long Short-Term Memory Networks is built to determine the value of cryptocurrency in the future. As a study, three cryptocurrencies – Bitcoin, Litecoin and Ethereum has been taken into consideration. A comparison of the results by using opinion mining to interpret the mood of the market on the current day for different currencies has been done. The sentiment scores got from natural language processing of textual data are used as features to the model used for predictions. The time-series charts are plotted using Plotly – python library for graphing plots. The Mean Absolute Error calculated between the actual and predicted values is used as the uncertainty quantification method. These uncertainty quantification methods are compared to analyze the present-day scenario of the market using opinion mining.**

Keywords—Time Series Analysis, Long Short-Term Memory Networks, Natural Language Processing, Sentiment Score, Plotly, Mean Absolute Error.

I. INTRODUCTION

The past few years have seen a tremendous spike in the stocks of cryptocurrency market. As we witness on a day to day basis, the market value of different commodities and companies are resilient. The cryptocurrency market on the other hand are more volatile than the standard assets. This volatile nature of various currencies like Bitcoin, Ethereum and Litecoin provide a good scope for market analysis through general data mining techniques as well as opinion mining techniques.

When we refer to the data mining techniques, an external API is used to get historic data as well as present day data which is used as input features for the Long-Short Term Memory models. There are many sources such as Quandl, Alpha Vantage, Coin Market etc. which provide us API's to harvest data. For study proposed through this paper, the Coin Market API which provide data with attributes denoting the date, high, opening price, closing price, volume traded, and market cap has been used.

The public sentiment of a particular commodity is compelling enough for the rise or fall of its stocks. A good way for us to get the public opinion is through social media portals such as Twitter or Reddit. This process of obtaining data from social media to analyse the public sentiments is called as opinion mining. The data is obtained from the Twitter Live Streaming API. The tweets are labelled using Natural Language Processing. The series generated by labelling the tweets is used as a feature for building models.

The overall flow of the general architecture is described in brief in the next few sections.

II. LITERATURE SURVEY

Opinion mining is customarily used among researchers to analyze the market using social media. The data obtained from social media platforms such as Twitter give researchers an opportunity to draw interesting insights into the data, which can further be used for predicting the future sentiment of the market. Another customary technique to predict the market cap prices is through time-series analysis. The novelty in this paper is brought about by using important data from Twitter such as sentiment scores of Tweets, popularity through number of mentions and re-tweets as features to the Long Short-Term Memory Networks. A few papers that came close to making similar contributions are referred below.

In the paper cited in [1], various time-series methods such as the Exponential Weighted Averages, Generalized Autoregressive Conditional Heteroskedastic Model, Structural Time Series models were used to ascertain the following day prices of bitcoin. Various uncertainty quantification methods such as Mean Absolute Error and Root Mean Square Error were used to determine the accuracy of the predictions.

Opinion mining of tweets were carried out in a paper referred to in [2]. The price fluctuation of a small cap cryptocurrency named ZClassic was analyzed using the sentiments of Tweets obtained over a period of three weeks. A series of Tweets labelled as Positive, Negative, Neutral were fed as input to a Gradient Boosting Tree Algorithm.

A time series forecasting similar to [1] was done in a paper cited in [3]. Dynamic Linear Models and several Multivariate Vector Autoregressive models were used for time series forecasting for different intervals. This was done for several cryptocurrencies such as Bitcoin, Litecoin, Ripple and Ethereum.

In the paper referred to in [4], various machine learning techniques were compared using their performances on data

ISBN: 978-1-5386-6078-2 © 2018 IEEE

collected for a period of 3 years. Three methods namely, XgBoosting, simple moving averages strategy and Long Short-Term Memory Networks were used. Sharpe Ratio and Geometric Mean Return were used as performance metrics.

A good example of the contributions made by our paper would be the combined efforts of [5] and [4]. As cited in [5], the fluctuations of cryptocurrency prices were determined based on the comments and replies of users. A crawler was used to obtain the comments from websites related to different cryptocurrencies. Vader based tagged values were used to understand the sentiments of the textual data.

The uniqueness in the paper cited in [6] was brought about by the optimized Bayesian implementation of the Recurrent Neural Networks – Long Short-Term Memory Networks. The accuracy was compared with that of ARIMA models and found to be higher than that of the latter.

The papers from [1] through [6] were referred for understanding the work carried out to date with respect to the time series modelling using different techniques. From papers cited in [7] through [12], the textual analysis of data obtained from social media platforms were mined to establish the nature of the cryptocurrencies. The oddity in our paper is brought about by using the features obtained from the Twitter stream for short term predictions by modelling them into the Long Short-Term Memory Networks.

III. Architecture Design

The architecture design is shown in Fig 1.

A. Data Gathering

The entire process generally begins with the gathering of data – both historic as well as streaming. The data we are dealing with is both structured as well as unstructured. The historic data is obtained through a request to an external API such as Coin Market. This API gives us access to cryptocurrency information for key searches such as Bitcoin, Ethereum, Litecoin. The attributes such as open, close, high, low, volume and date constitute the dataset. This forms the historical part of the data as well as the structured component of the process. This data can be easily downloaded from Kaggle. But, if we need to incorporate recent data such as present-day prices, we will have to use API's. This is the purpose of using an external API over existing datasets.

Another source from which we reap data is the Twitter Stream API. This contains textual information about the Cryptocurrency market that can further be used as a feature for training the Long Short-Term Memory networks. Opinions from Twitter form the unstructured part of the data which plays an important role in the rise and fall of Cryptocurrency.

B. Data Pre-Processing

In this step, we only deal with the textual data obtained from Twitter. The tweets obtained from users contain data noisy information such as emoticons, and some other symbols which cause erroneous predictions. These must be stemmed before any natural language processing can be carried out. It is also a good practice to stem the stop words – the common words that do not play any role in determining the sentiment of the text. The further process includes the labelling of the human annotated tweets as 0, 1, 2 which denotes Neutral, Positive, Negative respectively.

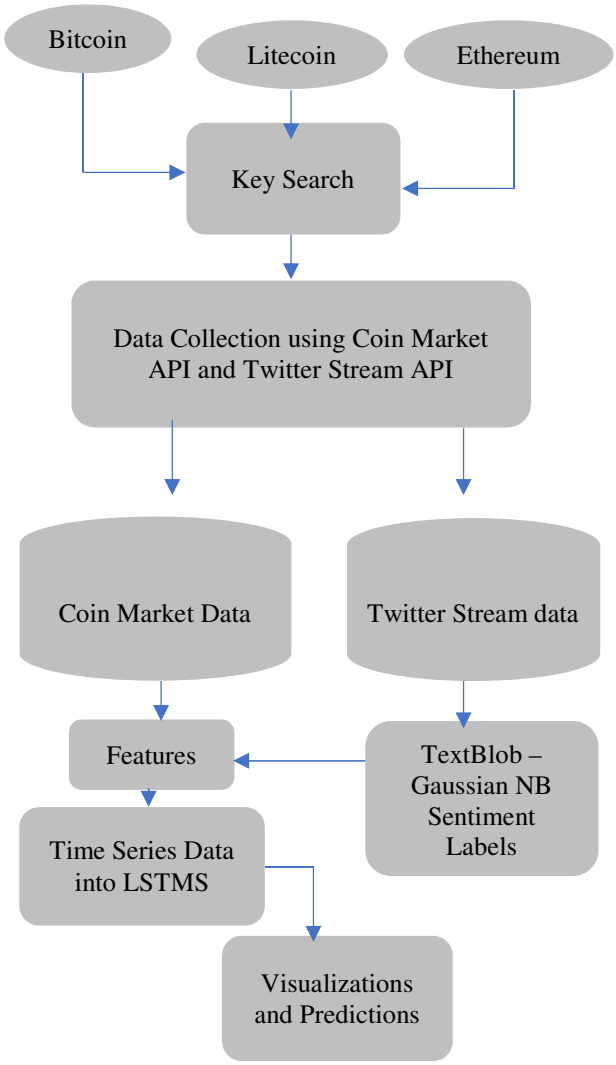
C. Training the Long Short-Term memory Networks

The model built uses past data of coins such as Bitcoin and Litecoin to predict the closing price for the succeeding days. Since we are modelling using long term dependencies, we must assign the prior time frame over which the network has access to. Hence, we split the time frame into windows of ten successive days. This will lead to the addition of large number of data frames.

D. Visualizing the output array

After training the models over layers of Neural Networks, predictions for the closing prices of the successive days are made. An analysis of how this validates with real time data is done. Certain Uncertainty Quantification methods such as Root Mean Square Error and Mean Absolute Error are visualized as heat maps.

A visualization of the data obtained from Twitter for this purpose is carried out. Certain graphs such as line charts plotted for Sentiment versus Time Frame, Bar Charts showing analysis of popularity of Bitcoin or Litecoin can be used as measures of validations. For the purpose of



ISBN: 978-1-5386-6078-2 © 2018 IEEE

visualizations, we have used Plotly - a visualization package.

Fig 1. Architecture Design

IV. Analysing The Coin Market Data

A. *Overview of the Data*

The dataset for each cryptocurrency consists of attributes such as Date, Open, High, Low, Close, Volume and Market cap. The snippet of data for Bitcoin is shown in Fig 2.

	Date	Open	High	Low	Close	Volume	Market Cap
0	2018-11-10	209.97	213.86	209.81	212.53	1377760000	21917195708
1	2018-11-09	211.99	213.32	209.51	210.07	1554750000	21659329261
2	2018-11-08	217.33	218.34	212.20	212.23	1769080000	21877424057
3	2018-11-07	218.90	221.65	216.80	217.18	1927830000	22383497662
4	2018-11-06	209.47	218.45	207.89	218.45	1856940000	22445690692

Fig 2. The snippet of the dataset

This snippet shown represents the head of the dataset for Bitcoin. The same structure exists for other currencies as well. To verify if the data extracted is accurate, we plot a plot of Date versus Volume and Date versus Closing price. The graph is shown in Fig 3.

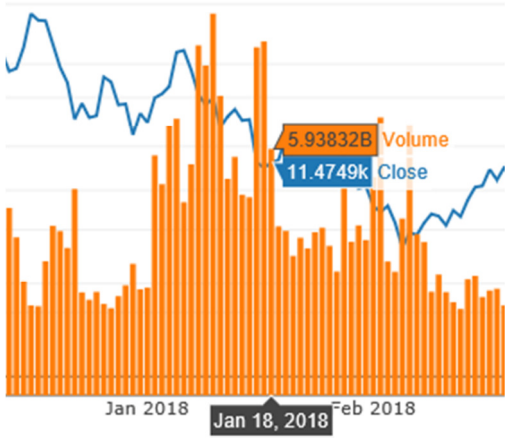


Fig 3. Plot of Volume and Close versus Date for Jan 18th, 2018

The volume is represented as orange bars and close as blue line. The above graph shows the respective values on January 18th, 2018. The next plot – Fig 4 shows a similar comparison for an earlier data say about a year ago – June 15th, 2017. This plot is shown below. We can see that the value of Bitcoin was much lower when compared to the data collected on January 18th, 2018. The similar graph can be shown for other cryptocurrencies such as Litecoin and Ethereum. After doing a test and a train split, we feed the data into a Long Short-Term Memory Network.

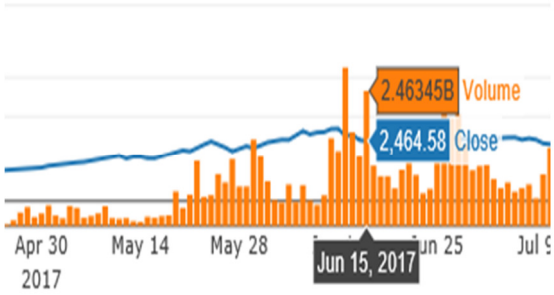


Fig 4. Plot of Volume and Closing Price vs Date on Jun 15th, 2017

B. *Train and Test Split*

While building a deep learning model, it is customary to split the collected data into a train and a test seta arbitrarily. The model was built on the test set and validated on a test data. A split date as December 1st, 2017 was demarcated for this reason. We train our model over data where cryptocurrencies were a bit low priced. The graph of Train and Test Split versus Date is shown below.

Fig 5. Graph of Test and Train Split versus Date.

C. *Preparing Data for the Long Short Term Memory Networks.*

Before beginning to build the models, some modifications to our data needs to be done. By this we mean that unnecessary features from the previous data must be eliminated and important features must be integrated. This is also where the most important feature – the data from Twitter comes in. Features like sentiment score, number of mentions about a particular cryptocurrency and number of retweets were added.

Certain attributes like opening price, daily highs and lows were considered. New attributes such as "close off high" that represents the difference between the high and

ISBN: 978-1-5386-6078-2 © 2018 IEEE

close for that day was added. The difference between high and low divided by the opening price is represented as "volatility".

The model built makes use of historic data represented as time frames of an arbitrary window (say 10 days) to predict the price of the succeeding day. We must also keep in mind that we are dealing with a lot of data which varies on a large scale. For instance, some features vary from -1 to 1 whereas some of them vary in terms of millions. To bring about uniformity we need to normalize the data.

D. Building the Long Short-Term Memory Network

A function named *build_model* was written to construct a simple sequential model to which a layer of Long Short-Term Memory Networks is added. Obviously, this layer must be made to fit according to our input shape. We also specify the number of neurons used for the computation and we train it for around 50 epochs. The graph of Mean Absolute Error versus Number of Epochs is shown below. If the training of the model went as required, then the Mean Absolute error must reduce with the increase in the number of epochs.

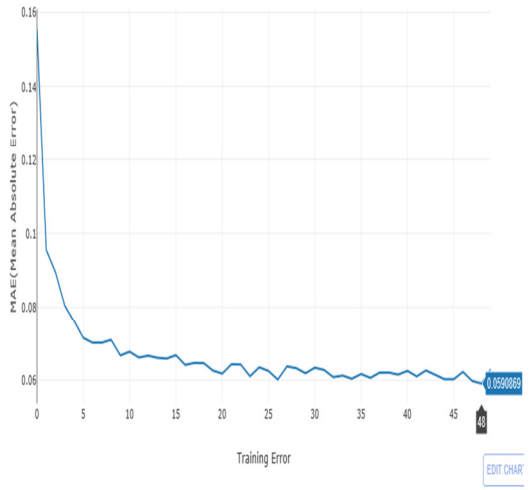


Fig 6. Graph of Epochs vs Training Error.

E. Results of the Predictions

Once done with building the model to predict the following days closing price for Bitcoin, Ethereum and Litecoin, we need to observe how well the model performs on the training set (before the split date December 1st, 2017. Fig 7 shows the predictions for the training data.

From Fig 7, the difference between the expected value and observed values on the test set is observed. To compute the error in the predictions we use Mean Absolute Error. We get an error of 0.054 in the prediction. This is observed in the Fig 7, where the actual value of closing price was supposed to be 218.3 but the model predicted it to be 226.9935. However, the model can be made to perform better by increasing the number of neurons and increasing the number of epochs.

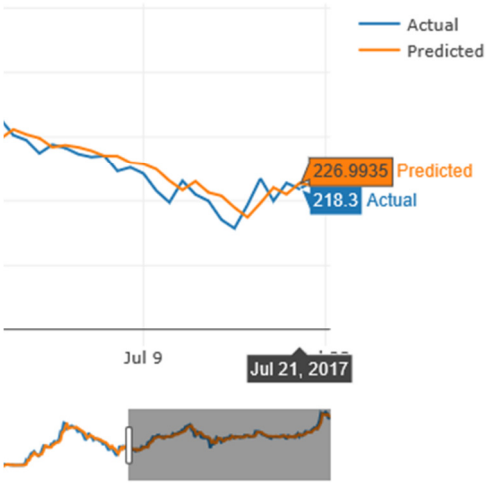
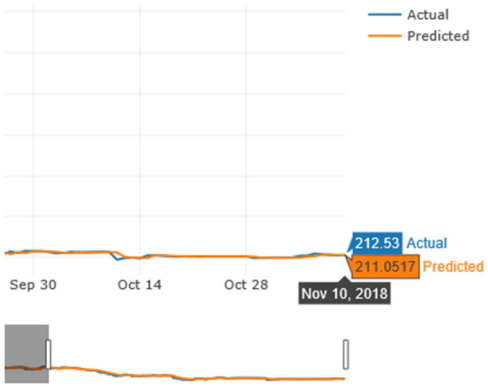


Fig 7. Graph of Actual vs Predicted price for Train Set



Having observed the predictions on the training data, we can now move onto the results for the test set. Fig 8. Shows the predictions for Test set.

Fig 8. Graph of Actual vs Predicted for Test Set

Here, it is observed that our model performs much better on the test set. A mean absolute error of 0.045 on the test set was obtained. However, there is a bit of a flaw in the model. It fails to detect the sudden rise and fall of prices as seen in the below Fig 9.

ISBN: 978-1-5386-6078-2 © 2018 IEEE

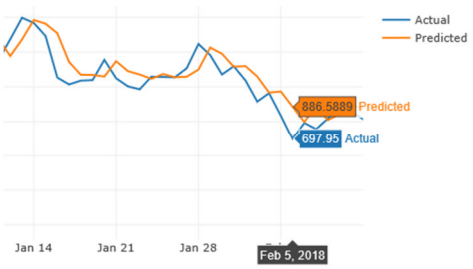


Fig 9. Graph Showing the downside of the predictions.

V. Analyzing Twitter Data

A. Gathering the data and Processing

Twitter provides APIs for collecting data for the past week. But it does not provide free access to historic data. Data was collected according to a search criterion for each of the cryptocurrencies through the API. About 1500 tweets for each of the cryptos was collected. When a user tweets, he does not do so in a text which can be interpreted by a computer. Hence, we need to do some cleaning to make it easier for computers to comprehend the text. Stemming the stop words, eliminating unnecessary characters, removing emoticons and removal of repeated tweets are some common data cleaning strategies to be followed while analyzing textual data. After this we analyze the present-day situation of the cryptocurrency market to validate the predictions we had made in the previous sections.

B. Analyzing Twitter Data

A keen eye is kept on the sentiment of the cryptocurrencies so that we can arrive at a conclusion as to which coin is performing well in the market. The below bar chart gives an idea of the sentiment of each coin. From Fig 10, we can see that Bitcoin is the most popular compared to Litecoin and the other cryptocurrencies. This comparison was made on a total of 2000 collected tweets.

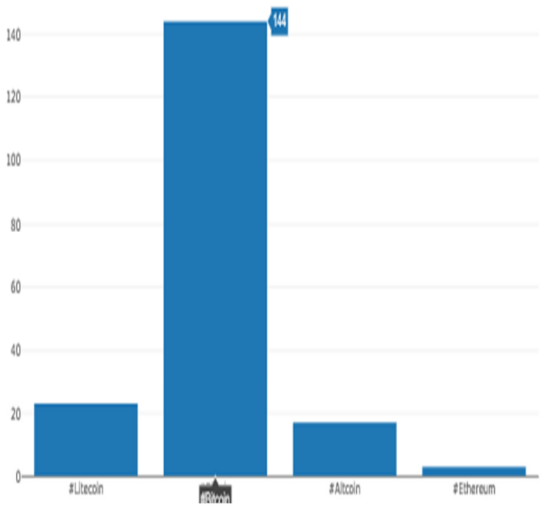


Fig 10. Bar Chart Showing the popularity of Cryptocurrencies.

Now that we know which coin is the most popular, we would want to look at if that popularity is towards a positive cause or a negative cause. The next bar chart in Fig 11, shows the aggregate tweets based on whether the tweets are positive, negative or neutral. Out of the 144 tweets collected on the present-day for Bitcoin 86 were neutral, 15 were negative and 33 were positive. This should imply that, the stocks for bitcoin must have risen. We can compare this with figure 8 for the date November 12th, 2018.

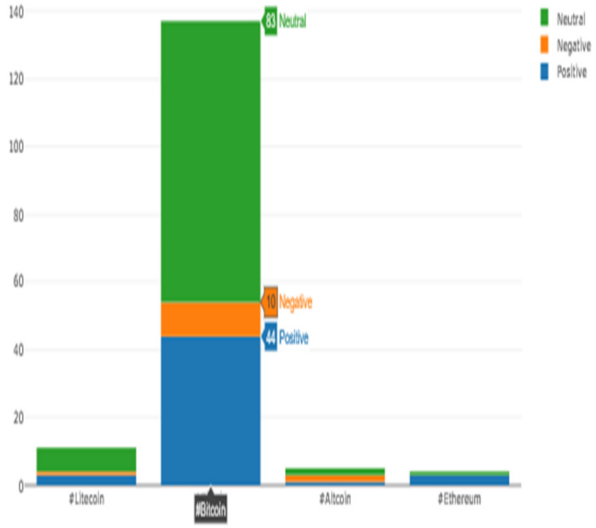


Fig 11. Bar Chart showing the sentiment of the cryptocurrencies.

A time-series chart of live streaming data compared with the time-series predictions of Long Short-Term Memory Networks can also give us a good idea about the surety of the predictions. Fig 12 gives us a good understanding about the same.

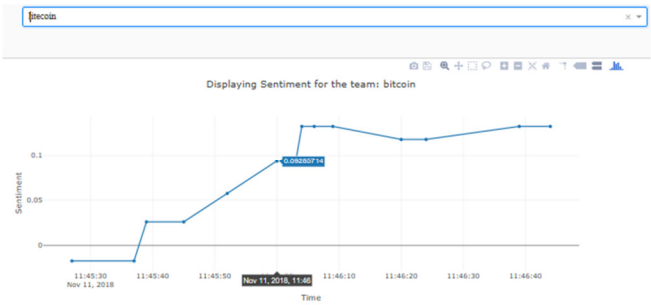


Fig 12. Live Stream Time-Series Chart.

ISBN: 978-1-5386-6078-2 © 2018 IEEE

We can also have a look at the world sentiment by visualizing a heat map of the sentiment of the tweets aggregated country-wise. Fig 13 shows the following heat map.

Fig 13. World Map showing the aggregated sentiments.

Now that we have the analysis of both data obtained from Coin Market Cap as well as Social Media platform such as Twitter, we need to draw conclusions if the analysis for Twitter compliments the analysis for Coin Market Cap. The succeeding section discusses the results which determine if the predictions and the present-day market sentiment go hand-in-hand.

VI. Results And Results

We analyze the results from the predictions made by the time-series models for the different currencies namely Bitcoin, Litecoin and Ethereum. We take into consideration the previous five days predictions and compare them with the present-day prediction. We also take into consideration the Mean Absolute Error for the predictions. The results are summarized in TABLE I and TABLE II.

TABLE I. COMPARISION OF ACTUAL VS PREDICTED VALUES

Date	**Ethereum Price ($)**		**Litecoin Price ($)**		**Bitcoin Price ($)**	
	Predicted	**Actual**	**Predicted**	**Actual**	**Predicted**	**Actual**
Nov 5, 2018	203.9	209.09	412.4	410.89	562.1	563.44
Nov 6,2018	206.14	218.45	414.33	415.45	560.44	562.32
Nov 7, 2018	207.41	208.23	414.78	416.41	560.66	561.31
Nov 8, 2018	208.45	210.43	418.65	421.45	557.98	560.12
Nov 9, 108	212.67	214.14	423.54	425.54	561.54	558.97

From the above table, we can draw conclusions that the most popular cryptocurrency is Bitcoin followed by Litecoin and the least popular one is Ethereum. We can also conclude that the prices are on taking a considerable plunge when compared to the previous day predicted and actual values – if not constant. Now we can compare these values with the sentiment of the Tweets we have obtained on a streaming basis. The TABLE II. shows the popularity and sentiments of each the cryptocurrencies. The below table gives the count of Tweets and the sentiments on November 9th, 2018.

TABLE II. COMPARING THE SOCIAL SENTIMENT

	Total	**Positive**	**Negative**	**Neutral**
Bitcoin	144	64	10	70
Ethereum	3	2	0	1
Litecoin	23	15	4	4

From the above table we can conclude that Bitcoin is the most popular cryptocurrency on the social media compared to others. We can also see that there is a relative neutral to positive response for each of the cryptocurrencies.

We can hence conclude that the results obtained as shown in Table II. validates and also compliments the results obtained as shown in Table I.

REFERENCES

[1] An Experimental Study of Bitcoin Fluctuations using Machine Learning Methods, Tian Guo, Nino Antulov-Fantulin, arXiv:1802.04065 [stat.ML].

[2] Sentiment Based Predictions of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model, Tianyu Ray Li, Anup S. Chamrajnagar, Xander R. Fong, Nicholas R. Rizik, Feng Fu, arXiv:1805.00558.

[3] "Forecasting Cryptocurrencies Financial Time Series," Leopoldo Catania & Stefano Grassi & Francesco Ravazzolo, 2018. Working Papers No 5/2018, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

[4] Anticipating cryptocurrency prices using machine learning Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, Andrea Baronchelli, arXiv:1805.08550 [physics.soc-ph].

[5] Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, Chang Hun Kim

[6] Predicting the Price of Bitcoin Using Machine Learning, Sean McNally, Jason Roche, Simon Caton, 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP).

[7] "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis,"Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) SMU Data Science Review: Vol. 1 : No. 3 , Article 1.

[8] Predicting Bitcoin price fluctuation with Twitter sentiment analysis, EVITA STENQVIST, JACOB LÖNNÖ.

[9] "Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis." Colianni, Stuart, Stephanie M. Rosales and Michael Signorotti. (2015).

[10] Bitcoin Response to Twitter Sentiments, Svitlana Galeshchuk, Oleksandra Vasylchyshyn, Andriy Krysovatyy.

[11] "Using sentiment analysis to predict interday Bitcoin price movements" Vytautas Karalevicius, Niels Degrande, Jochen De Weerdt, (2018), The Journal of Risk Finance, Vol. 19 Issue: 1, pp.56-75, https://doi.org/10.1108/JRF-06-2017-0092.

[12] Bitcoin Spread Prediction Using Social And Web Search Media , Matta, Martina & Lunesu, Maria Ilaria & Marchesi, Michele. (2015).

ISBN: 978-1-5386-6078-2 © 2018 IEEE