International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Price Prediction and Analysis of Financial Markets based on News, Social Feed, and Sentiment Index using Machine Learning and Market Data

Tapan Mehta[1], Ganesh Kolase[2], Vivek Tekade[3], Rahul Sathe[4], Anand Dhawale[5]

[1,2,3,4] *Student, Dept. Of Computer Engineering, Modern Education Society's College of Engineering, Pune, India*
[5]*Professor, Dept. Of Computer Engineering, Modern Education Society's College of Engineering, Pune, India*

Abstract - *Cryptocurrency is the whole new market for trading, earning money and gaining profits using a complete digital mode of transaction. In this paper we focus on Cryptocurrency named Bitcoin. The predictions of prices are done in real-time based on news and tweets using a LSTM model. Our dataset consists of various features related to bitcoin over 7 years recorded daily.*

Keywords: **Bitcoin, Sentiment, Cryptocurrency, LSTM model, Twitter.**

1. INTRODUCTION

Accurate Price prediction of any currency is always a tedious task. Various Machine Learning algorithms have been used in Price prediction of the stock market. Hence, it is now possible to predict the price of highly volatile Cryptocurrencies. Bitcoin was invented in 2008 by an unknown person or group of people using the name Satoshi Nakamoto and initiated in the year 2009 when the source code was released as open-source. Bitcoins were created as a reward for process mining. Unlike fiat currency, Bitcoin is created, distributed, traded, and stored with the procedure of a decentralized ledger system recognized as the block chain. Being highly volatile the price of Bitcoin depends on the very large number of variables including people's opinions, buzz, and news around the world. Due to encroachment in technology, it is possible to process text or spoken languages into an analyzable form. Sentiment analysis is the machine learning methodology for NLP. Sentiment Analysis is the procedure of 'computationally' defining whether a section of text is positive, negative, or neutral. It is also termed as opinion mining, evaluating the opinion and attitude of the speaker.

1.1 Data Creation

The data is collected from Twitter and well-known news sources. This data consists of news and tweets from Twitter. Data is directly or indirectly related to Bitcoin. The news is scraped using the Scrapy framework. Scrapy provides complete packages for Scrapping requirements. The tweets are collected using the Twint Python library. The scraped data is stored in CSV file format in local storage. The data is preprocessed for sentiment analysis. The sentiment analysis will label the data into three types, p for positive news, n for negative news, and neutral.

1.2 Data Preprocessing

Preprocessing is an important step after data gathering. It is very hard and unwise to directly use raw data for machine learning. Preprocessing includes cleaning, integration, transformation, and reduction techniques. Data cleaning includes handling of missing data and noisy data. Data transformation includes normalization, attribute selection, discretization. Data reduction comprises Data cube aggregation, attribute subset selection, dimensionality reduction. After applying all the above techniques, the data becomes usable for machine learning. The extracted data contained many features out of which few were selected. Stop words were removed for better sentiment analysis from tweets.

1.3 Sentiment Analysis

Sentiment Analysis is the method of 'computationally' defining whether a section of text is positive, negative, or neutral. It's also termed as opinion mining, which consists of evaluating the behavior of the individual. The preprocessed data is fed to our model which labels the data. Initially, the model is trained for labeling.

1.4 Machine Learning Training Model

Long Short-Term Memory (LSTM) networks are a kind of RNN (recurrent neural network) mostly used in sequence prediction problems. This is a conduct needed in intricate problem domains like machine paraphrase, speech identification, etc. LSTM lies in the complex structure of DL. Studying and implementing LSTM is tedious work.

2. LITERATURE SURVEY

The initial part of the paper [1] is database collection. Quandl and CoinmarketCap databases are used for retrieving bitcoin values. After acquiring this time-series data recorded daily for five years at different time instances. They have normalized and smoothened it. For this, they have implemented different normalization techniques. The techniques are log transformation, z-score normalization, box cox normalization, etc. After this, data is smoothed over the complete period. After feature selection, the sample inputs are fed to the model. The variation in the bitcoin values is denoted a pattern. The pattern consists of variations in a positive or negative type compared to the previous day's data. After establishing the learning framework and completing the normalization, they intend to use the two methods. Bayesian Regression and GLM/Random Forest, then choose the best method to solve the Bitcoin prediction problem. The accuracy is compared with different models after the final Prediction

The aim of their work [2] was to derive the accuracy of Bitcoin Prediction using different machine learning algorithms and compare their accuracy. They have collected the dataset for the document with the following details from quandl.com and applied machine learning algorithms viz. decision tree and regression for prediction and price forecast. Test outcomes are matched for decision trees as well as regression models. The proposed learning method suggests the best algorithm to choose and adopt for the cryptocurrency prediction problem. The experimental study results show that linear regression outperforms the other by high accuracy on the price prediction.

The goal for their [3] innovative project is to show how a trained machine model forecasts the value of a cryptocurrency if we provide a sufficient quantity of data and computational power. They have collected the historical data from poloniex.com using a REST API call. API gives data from 2015 to the in time intervals of 5 mins and 2 hours. The collected data is then placed into a Data Frame. Convolutional Neural Networks (CNN) is a deep learning methodology used for classification. However, here we tweak it to be used for prediction. By setting up a one-dimensional network instead of 2D or 3D, they predict the output by feeding in a list of the close prices from our dataset The neural networks built on in this project were completed using the Keras libraries. Keras offers neural network API which can run on Tensorflow or Theano. Keras facilitates seamless prototyping. Like all python libraries Keras also takes advantage of the modularity concept providing users with independent configurable modules. Since all the code is purely written in python, python developers do not find it hard to debug or run complex modified code. Predicting the future will always be on the top of the list of uses for machine learning algorithms. Here in this project they have attempted to predict the prices of Bitcoins using two deep learning methodologies. The Web application is designed on the Django web framework and has two pages for one for the CNN network and other for the LSTM (Long Short Term Memory) network

Akhilesh P. Patil has proposed in this paper [4] usage of Short-Term Memory Networks for predicting the future price of cryptocurrency through a time series model. Major considerations of cryptocurrencies in the market are Bitcoin, Ethereum, and Litecoin.If you have a Table, simply paste it in the box provided below and adjust the table or the box. If you adjust the box, you can keep the table in single column, if you have long table. In this paper they have compared various opinions on the cryptocurrencies

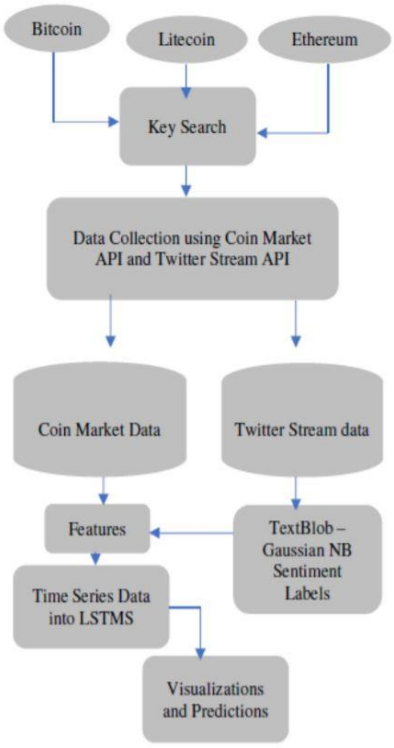


Image source [4]

Based on which they have declared the sentiment scores from natural language processing of the textual data. Features given to the model are the sentiment scores derived as explained above which are used for future predictions. The output is represented in the form of a time series graph using Python library Plotly. Here they have used the uncertainty quantification method which consists of calculating The Mean absolute error is the calculation of actual and predicted. This Comparison between Uncertainty quantification methods is done in this paper to get current cryptocurrency trends using the opening mining technique.

Pavitra Mohanty, Darshan Patel, Parth Patel, Sudipta Roy have presented in this paper [5] a way of predicting future fluctuations of cryptocurrencies. Here they have used Apache Flume for the gathering of Users' comments from Social media and data of price is collected through various exchanges. In this paper, they have used LSTM (Long Short-Term Memory) for forecasting Bitcoin trends through Twitter Data. Sentiment Index from data is derived leading to positive, negative, or neutral sentiments. Here, they also have used information from the Block chain as one of the major considerations affecting bitcoin market trends. More weightage on LSTM is given for prediction of future prices of Bitcoin. Due to high volatility in the market the model does not meet the accuracy requirements. The precision given by your model is 60% and accuracy is 50%.

Dibakar Raj Pant, PrasangaNeupane, Anuj Poudel, Anup Kumar Pokhrel, Bishnu Kumar Lama have proposed in this paper [6] the approach for predicting Bitcoin prices based on analysis of tweets gathered from news and twitter. The data collected is classified into 3 categories- positive, negative, and neutral. Positive and negative tweets with historical data are given as input to the RNN model for prediction of price. In this paper sentiment analysis has a major execution weight in the workflow. RNN model is used for future price prediction using the historical data. It also shows a moderate correlation of 0.41 between the rise of negative opinions on Twitter related to Bitcoin and its consequent fall in price the positive and negative sentiment scores with the accuracy of 77.62% is another useful work.

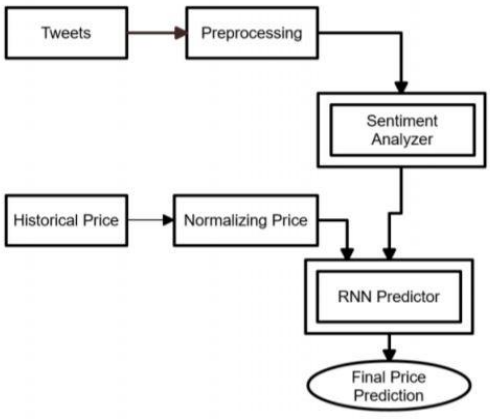


Image source [6]

This paper [7] prediction the price of the two cryptocurrencies like Bitcoin and Litecoin is done based on sentiment analysis of tweets. Multiple Linear Regression is used which forecasts the price with R2_score of 44% and 59% respectively. From these scores, they can infer that Bitcoin's price does not get much affected by the sentiments of tweets in comparison to the Litecoin's price.

The fluctuation of the prices of Bitcoin has a reliance on various factors like mining cost, economic factor. The price of the cryptocurrencies for 2 hours is predicted and the dependency of cryptocurrency price on the number of positive tweets in this duration is returned. It is noted that social factors play a major role in deciding the price of a cryptocurrency. Their proposed framework works in two phases Training phase and a Detection phase. The training phase is a one-time activity. For carrying out the training phase, they have collected Twitter data and the concurrent Bitcoin and Litecoin prices. The amount of positive, neutral, and negative.tweets present in one chunk is calculated. The count of positive tweets, neutral and negative tweets are the features of the dataset, and the mapped average price is the label of the dataset. The model is validated with the original labels of the given dataset. If the result of validation is acceptable, then the model has used prediction of future price, if not then a new model is to be designed. In the detection phase, real-time tweets are inputted to the model, and the model predicts the average price for two hours.

3. SYSTEM ARCHITECTURE

The above figure describes the system architecture. The Training data for the model consists of past seven year data out of which news is scraped from trusted news sites using scrapy framework and twitter data using twint library of only those users who have followers more than fifty thousand,all this data is stored csv in the form of rows and columns .Next we work on cleaning and pre-processing the data .We use nltk library to remove the stopwords from the data which helps us to get more accurate sentiment score. In the next step Sentiment Analysis is done using a Vader sentiment library on the processed data which helps in determining the trends by giving us the positive, negative and neutral score. Alongside we have taken the Bitcoin Closing, Opening, High and Low prices of each day from 2013 to 2020. So now we map the sentiment index of a particular day with above prices of the next day by doing the respective date manipulation. In this way the training data is prepared.

Now the detection phase begins,in the detection phase real-time tweets are inputted to the model and the model predicts the price for the duration of hours after which we are running the scheduler .The scheduler runs the twitter scrapper in the crontab on a remote server to get the real time tweets, the twitter scraper is built using the twint library . After getting the tweets, it is given to the sentiment analysis model and then we take the mean of all the sentiment index of that time period which is fed to the machine model then finally we get the predicted price for that time period.

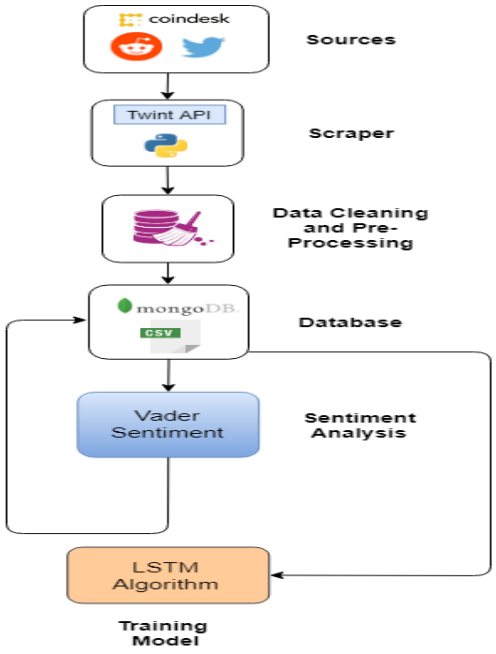


Fig 3. System Architecture

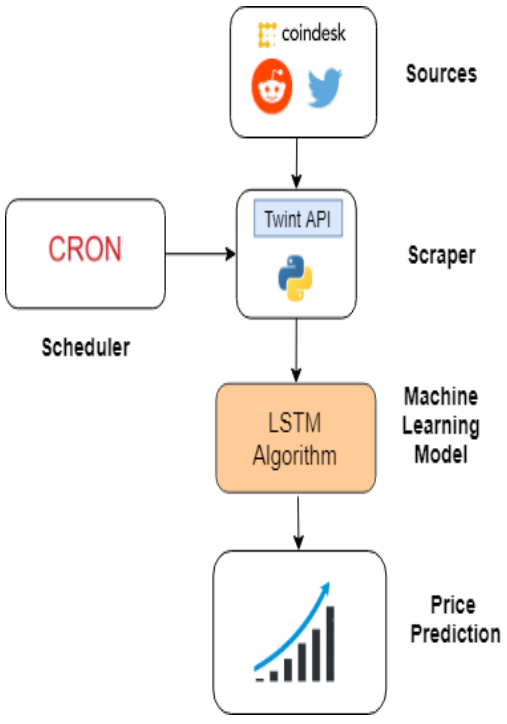
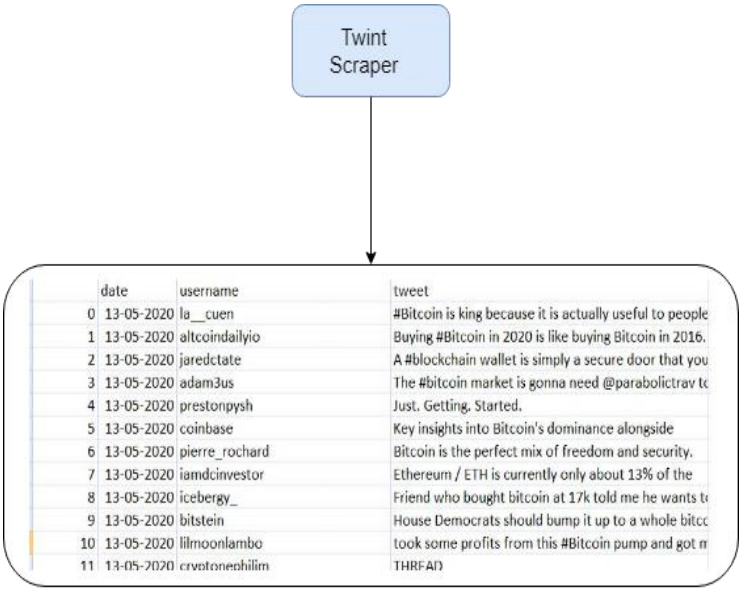


Fig 4. Real Time Scheduler

4. DATA EXTRACTION FOR TRAINING MODEL

4.1 Data Scraping



	date	username	tweet
0	13-05-2020	la__cuen	#Bitcoin is king because it is actually useful to people
1	13-05-2020	altcoindailyio	Buying #Bitcoin in 2020 is like buying Bitcoin in 2016.
2	13-05-2020	jaredctate	A #blockchain wallet is simply a secure door that you
3	13-05-2020	adam3us	The #bitcoin market is gonna need @parabolictrav t
4	13-05-2020	prestonpysh	Just. Getting. Started.
5	13-05-2020	coinbase	Key insights into Bitcoin's dominance alongside
6	13-05-2020	pierre_rochard	Bitcoin is the perfect mix of freedom and security.
7	13-05-2020	iamdcinvestor	Ethereum / ETH is currently only about 13% of the
8	13-05-2020	icebergy_	Friend who bought bitcoin at 17k told me he wants t
9	13-05-2020	bitstein	House Democrats should bump it up to a whole bitcc
10	13-05-2020	lilmoonlambo	took some profits from this #Bitcoin pump and got m
11	13-05-2020	cryptonephilim	THREAD

Fig 5. Data from Twint

Our project is based on sentiment analysis of tweets and news. Tweets are extracted using Twint Python library which helps to extract tweets based on required conditions. We have searched and extracted tweets using the keyword 'Bitcoin' of people with more than 50000 followers. Hence making the tweets more reliable for training and prediction purposes. We have collected tweets of the past 7 years from the influencers of Bitcoin.

Currency	Date	Closing Price	24h Open	24h High	24h Low (U
BTC	01-10-2013	123.65499	124.3047	124.752	122.5635
BTC	02-10-2013	125.455	123.655	125.759	123.6338
BTC	03-10-2013	108.58483	125.455	125.666	83.32833
BTC	04-10-2013	118.67466	108.5848	118.675	107.0582
BTC	05-10-2013	121.33866	118.6747	121.936	118.0057
BTC	06-10-2013	120.65533	121.3387	121.852	120.5545
BTC	07-10-2013	121.795	120.6553	121.992	120.432
BTC	08-10-2013	123.033	121.795	123.64	121.3507
BTC	09-10-2013	124.049	123.033	124.784	122.5927
BTC	10-10-2013	125.96116	124.049	128.017	123.8197
BTC	11-10-2013	125.27966	125.9612	126.437	124.1138
BTC	12-10-2013	125.9275	125.2797	126.037	123.1297

Fig 6. Bitcoin Prices

The bitcoin prices are retrieved from the past 7 years from coindesk.com. The prices dataset consists of 4 types of prices of each day viz. Closing Price, Opening Price, Highest Price and Lowest price of that particular day. This gives clear insight of price variations occurring in a day.

4.2 Sentiment Analysis

Date	Positive	Negative	Neutral	Compound
18-11-2013	0	0	1	0
19-11-2013	0.152	0	0.848	0.38155
25-11-2013	0	0	1	0
26-11-2013	0	0.054	0.946	-0.1462
27-11-2013	0	0.169	0.831	-0.6249
28-11-2013	0	0.054	0.946	-0.1462
29-11-2013	0.276	0.061	0.663	0.6369
02-12-2013	0	0	1	0
03-12-2013	0.163	0.105	0.732	0.3094
04-12-2013	0.079	0	0.921	0.3612
06-12-2013	0.239	0	0.761	0.5719

Fig 7. Sentiment Analysis

Sentiment analysis in our module is done in 2 phases. Initially, as the tweets are scrapped each tweet(text) is passed through a process of stopword removal.The stopword removal is done using NLTK library using 'nlppreprocess'. After the stop word removal is done on the text, it is passed to VaderSentiment for sentiment analysis.

The output on each tweet has four parameters Positive, Negative, Neutral and Compound. The mentioned parameters are the individual weights depicting the behaviour of the user in the tweet.

Each day has multiple tweets as a result we get multiple sentiment values for each day so we calculate the mean sentiment values for each day as shown in fig [7].

4.3 Mapping of Tweets and Prices

Date	Positive	Negative	Neutral	Compound	Closing_price	Opening_price
18-11-2013	0	0	1	0	693.65	510.6025
19-11-2013	0.152	0	0.848	0.38155	531.54249	693.65
25-11-2013	0	0	1	0	789.36475	768.8475
26-11-2013	0	0.054	0.946	-0.1462	893.1815	789.36475
27-11-2013	0	0.169	0.831	-0.6249	934.355	893.1815
28-11-2013	0	0.054	0.946	-0.1462	1068.363	934.355
29-11-2013	0.276	0.061	0.663	0.6369	1154.92593	1068.363
02-12-2013	0	0	1	0	1028.845	1019.78966
03-12-2013	0.163	0.105	0.732	0.3094	1071.2848	1028.845
04-12-2013	0.079	0	0.921	0.3612	1139.33083	1071.2848
06-12-2013	0.239	0	0.761	0.5719	759.43041	1004.61633
09-12-2013	0.1095	0	0.8905	0.31845	916.77599	841.83966
15-12-2013	0.154	0	0.846	0.2912	868.95316	848.9975
16-12-2013	0.115	0.08	0.805	0.1531	653.80483	868.95316

Fig 8. Mapping

After the sentiment analysis is done, for creating training dataset bitcoin prices have to be mapped with the tweets. As the variations in Bitcoin's prices are due to these tweets hence price mapping is one of the very essential aspects. There are variations in price trends due to previous day tweets.

Hence mapping is done as follows:

Today's_tweets => price (tomorrow).
Using this strategy for mapping, it will help in predicting tomorrow's prices.

5. REAL TIME SCHEDULER

In this project we have used a real time scheduler for scraping data from considered sources in real time which will help in predicting the immediate trends of prices.

For a real time scheduler we will be using CronTab. It is a Linux based real time scheduler and executes the code after a specified interval mentioned in the execution engine of the scheduler.

Fig 9. CronTab Scheduler Command

As the scheduler executes periodically the following steps are executed -

I. Data extraction using Twint Library from Current System time till last 'n' numbers of hours.
II. Pre -processing of extracted data.
III. Perform Sentiment Analysis then find its mean.
IV. Using CoinDesk Api to retrieve current Bitcoin price and then mapping it to the sentiment values.
V. Mean given as input to Machine Learning which gives the predicted price of the next day.

6. EXPERIMENTAL ASSESSMENT

In this project we have tried to predict the bitcoin prices,using the sentiment value and it's analogous actual bitcoin prices of each day in the past. Here the training data consists of past data which has sentiments and its price trends, as it is a time-series data. For dealing with time-series data LSTM (Long short term memory) machine learning algorithm is the most efficient one. LSTM can also deal with missing time/date frames and maintain the accuracy of the model. Hence LSTM is preferred over

other ML algorithms when it comes for training of data consisting of time-series data.

6.1 Training Model:

The data on which the model is trained consists of 1559 tuples and each tuple consists of Date, Positive, Negative, Neutral, Compound sentiments with Bitcoin price of the next day. Out of 1559 tuples we have used 70% of it for training purposes and remaining for testing. The data is normalized in the range of 0 to 1.

For training purposes we have used tensorflow backend .Our model is a multi-feature sentiment analysis as it contains multiple sentiment values for each tweet and news which is input to the training model. Also the Multi-feature model provides better accuracy than the Single feature mode.

Then the input sentiment values are given to series_to_supervised function which converts the original values to a set of lag shifted values which are further reframed then passed on to the LSTM model.

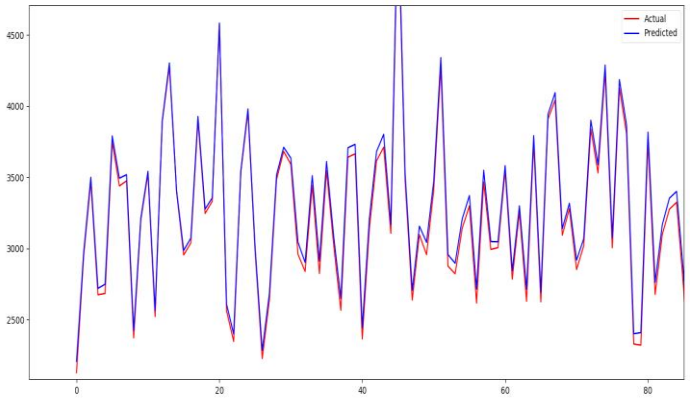


Fig 9. Validation Graph

6.2 Prediction Output

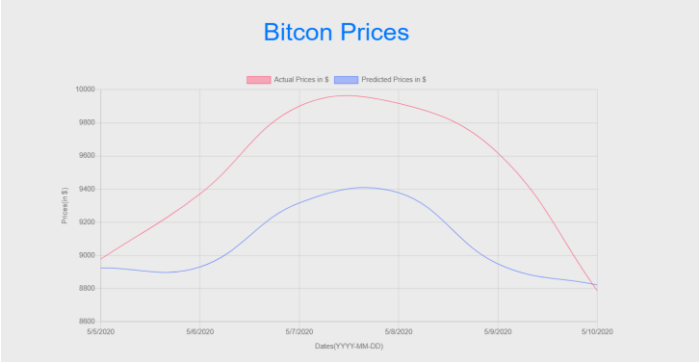


Fig 10. Predicted Prices

Date	Actual	Predicted
5/5/2020	8978.284	8924.906
5/6/2020	9371.684	8930.627
5/7/2020	9900.679	9316.75
5/8/2020	9917.248	9378.255
5/9/2020	9617.518	8948.799
5/10/2020	8786.655	8821.713

Fig 11. Predicted and Actual Prices

7. FUTURE SCOPE

As this project consists of a real-time scheduler which extracts data in real time and makes real time Bitcoin's price prediction, this can be also used in future for trading in real-time as a bot. Where the user has to give some amount to the for trading and the model will give the best profits by investing and selling accordingly at the right time.

As this model only focuses on Bitcoin prediction, it can also be used for prediction of other cryptocurrencies which are also famous in the market viz. Etherum, Litecoin, etc. So profits won't only be earned by investing in Bitcoin but also other cryptocurrencies.

8. CONCLUSION

We have successfully implemented the LSTM model for prediction of Bitcoin's prices in real-time based on sentiment analysis from Twitter and News. The real-time scheduler is live on server continuously extracting data and making predictions using the machine learning LSTM model.
The predictions are shown in graph form on a webpage

ACKNOWLEDGEMENT

It gives us pronounced pleasure in presenting the Survey Paper on "Price Prediction and Analysis of Financial Markets based on News, Social Feed, and Sentiment Index using Machine Learning and Market Data.". We feel very grateful to our guide Prof. A. D. Dhawale for giving us all the help and guidance we needed. We are really glad to have him for his kind support. His valuable suggestions were very helpful.

REFERENCES

[1] Siddhi Velankar, SakshiValecha, Shreya Maji, "Bitcoin Price Prediction using Machine Learning," International Conference on Advanced Communications Technology (ICACT) February 11 - 14, 2018.

[2] Karunya Rathan, SomarouthuVenkat Sai, Tubati Sai Manikanta, "CryptoCurrency price prediction using Decision Tree and Regression techniques". Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)

[3] S. Yogeshwaran, ManinderJeet Kaur, Piyush Maheshwari "Project-Based Learning: Predicting Bitcoin Prices using Deep Learning" 2019 IEEE Global Engineering Education Conference (EDUCON).

[4] Akhilesh P. Patil, Akarsh T. S, Parkavi A, "A Study of Opinion Mining and Data Mining Techniques to analyze the Cryptocurrency Market" 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions, 2018.

[5] Pavitra Mohanty, Darshan Patel, Parth Patel, Sudipta Roy, "Predicting Fluctuations in Cryptocurrencies' Price using users' Comments and Real time Prices," 2018 7th International Conference on Reliability, InfocomTechnologies and Optimization (ICRITO).

[6] Dibakar Raj Pant, PrasangaNeupane, Anuj Poudel, Anup Kumar Pokhrel, Bishnu Kumar Lama, "Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis".

[7] Arti Jain, Shashank Tripathi, Harsh Dhar Dwivedi, Pranav Saxena, "Forecasting Price of Cryptocurrencies using Tweets Sentiment Analysis."