BPTE: Bitcoin Price Prediction and Trend Examination using Twitter Sentiment Analysis

Muhammad K. Shahzad
Dept. of Computing, National University of Sciences and Technology, Pakistan.
mkhuram.shahzad@seecs.edu.pk

Laiba Bukhari
Dept. of Computing, National University of Sciences and Technology, Pakistan.
lbukhari.bese15seecs@seecs.edu.pk

Tayyeba Muhammad Khan
Dept. of Computing, National University of Sciences and Technology, Pakistan.
tkhan.bese15seecs@seecs.edu.pk

S. M. Riazul Islam
Dept. of Computer Engineering
Sejong University, South Korea
riaz@sejong.ac.kr

Mahmud Hossain
Dept. of Computer Science
UAB, USA
mahmud@uab.edu

Kyung-Sup Kwak
Dept. of ICE
Inha University, South Korea
kskwak@inha.ac.kr

2021 International Conference on Information and Communication Technology Convergence (ICTC) | 978-1-6654-2383-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICTC52510.2021.9620216

***Abstract*—Natural Language Processing (NLP) is a challenging and evolving field with the potential of mushroom growth. This technology is expected to assume a pivotal role in bridging the gap between human communication and digital data. In line with that, NLP-driven sentiment analysis has become an attractive research area. On the other hand, bitcoin, arguably the most valuable cryptocurrency, has gained popularity as a major source of investment. In this paper, we propose a framework to perform sentiment analysis on Twitter data. We outline the method and results of predicting bitcoin price for a few days in the future. The framework is expected to be helpful in making informed decisions about our investments and policy based on anticipated future trends.**

I. Introduction

Sentiment analysis is technique to to retrieve sentiments from data by analysing it. Sentiment analysis is also known as opinion mining is a process of analyzing collected data based upon the person's thoughts, feelings, and reviews. In sentiment analysis, we extract features from large data using different machine learning techniques such as statistical analysis and natural language processing. Sentiment can be positive, negative, or neutral. It is domain-centered.

Twitter is an online social networking and micro-blogging site and app in which users write short messages that exceed not more than 140 characters called "tweets". It is a global forum with the presence of eminent personalities from the field of entertainment, industry, and politics. Twitter data is unstructured and in natural language. We collect a dataset of tweets for marking them as positive, negative, or neutral based on a specific topic. The paper suggests using Twitter sentiment analysis to predict bitcoin prices and analyzing its trend. Text is processed, classified, tagged, and tokenized. NLTK pre-processes tweets and converts them to format easy for extracting sentiment.

Twitter is a source of news and information for millions of people daily. Companies and customers post their thinking and feedback on this platform as well. That is why tweets can be used to predict the prices of bitcoins, their trends, and whether they will rise or fall in the future. This article aims to extract the features of tweets and analyze the opinion of tweets as a positive, negative, or neutral influence on the price of bitcoins meaning whether they will increase or decrease.

- Input: Textual content of a tweet
- Output: Label signifying if the sentiment of the tweet is causing an increasing or decreasing trend in the bitcoin market.

We are using python to analyze the sentiments of tweets as rising or falling trends. Sentiments of tweets are graphed to visualize the trends. The scope of this project is limited to Twitter data and predictions related to bitcoin trends from that data.

II. Related Works

Mostly, research work in sentiment analysis is focused on finding the polarity of text that whether the tweet or comment is in favor or against on something of particular interest. The extensive usage of information and communication technology, provide a platform to determine public sentiments. The frequency of protests, agitations, and strikes have risen to an alarming rate. Authors [1], employed modified Naïve Bayes algorithm to help detect of conflicts using tweets from the stakeholders.

Today, Twitter and Facebook are popular examples of Social Networking websites which plays significant role in online social fabric. The compressed format of tweets in particular provides rich information that can be use for decision making after sentiment analysis. Author [2], employed machine learning for sentiment analysis on twitter data collected from different political parties and to predict a party performance in view of public perception.

Twitter through tweets is hugely popular to express sentiments on different occasions. By the last decade research in this field has witnessed mushroom growth. Compressed and challenging format of tweets make it difficult to process tweets sentiments. The reason is use of slangs and abbreviations which is not common. In [3], authors presented review of sentiment analysis on twitter highlighting methodologies

and models along with describing python based generalized approach.

In Presidential election forecast, traditional forecasting models consider poll survey and domestic growth as predictive factors. However, temporally or spatially dense poling has always been costly and uphill task. Due to immense popularity gain of social media has gain tremendous research. Research suggest that social media has shown the potential to reflect on political landscape. The study [4], shows proposed model can with an accuracy of 81% for the 2016 Presidential elections.

Advent of Web 2.0 has enabled social media to rapidly increase users generated contents. In this view, identifying and summarizing sentiments from raw data is ever challenging. Review work [5], presents techniques and trends for opinion minding and sentiments analysis. Another study [6], text minding with sentiment analysis to seafarers - which are more vulnerable to accidents and health hazards due to their work culture at sea.

Artificial intelligence and Machine learning methods have been used with data minding so solve big data problems. Using this effective and time saving techniques are estimated to double the annual revenue from returns from stock market. In work [7], the sentiment analysis analysis is employed on tweets using Twitter API. Authors [8], address the issue of sentiment analysis on twitter by categorizing feelings: positive, negative, or normal.

In view of ever increasing digitization, feelings expressed in twitter is playing important role in decision making. Instead of reading a lot of text, classification in positive and negative is more practical and effective. The survey [9], provides an overview of the current methods that solve sentiment analysis. In [10], address sentiment analysis problem during natural disaster or social movements. They have employed Bayesian network classifiers on two datasets in Spanish.

Sentiment analysis has been widely used to solve key tasks of natural language understanding. It has been found beneficial for deep understanding of opinion and sentiments on social network analysis. The authors [11], used Dynamic Bayesian Networks in modeling time series of the sentiment and their relationship. Further this model employees Gaussian Process Regression for analysing sentiments with time series with topics related at previous time. Authors [12], has discussed sentiment analysis using natural language processing. A lexicon-based techniques [13], has also been used extracting sentiment from user text.

The user generated contents (UGC) from discussion forum, blogs, forums, and social networks. Public open opinions provides crucial input for decision making. The computational treatment of opinions has become a challenging field to solve diverse range of problems. Work [14], evaluation of opinions to get a quick perspective of different views expressed by different users.

Text classification of identifying the class to which a document belongs. This work [15], uses a simple non-weighted features for text categorization. A feature selection approach to find best features for the learning task at hand. This not only help find suitable features but also reduce the number of selected features to learn from. Porter [16], in 1980 presented a simple algorithm for stemming English language words. That paper has presented the main features of the algorithm and subject range it covers.

Abstract sentiment analysis on Twitter Data is indeed a challenging problem due to the research challenging it posses, such as: nature, diversity and density of user data. The enormous amount of data offers great potential to harness sentiment tendency [17]. Digital currencies such as BitCoin has become a trending phenomenon for profiting in the financial markets. Authors [18], connects BirCoin with two phenomena namely; Wikipedia and Google Trends to study their relationships.

Article [19], is the first study on BitCoin price formation by considering both the traditional factors and BitCoin attractiveness for users. Authors [20], tremendous rise of crypto currencies and underlying block-chain have found widespread applications. The underlying clock are synchronized using energy-efficient time synchronization protocol for wireless sensor networks (ETSP) [21] protocol. BitCoin [22], price is highly volatile because of supply, demand, and cost of minding BitCoins.

III. Methodology

We have implemented the price prediction of bitcoin using Twitter data. Data is scraped from Twitter using python script and then filtered for the keyword of bitcoin. Using this we got the username of the person tweeting, id, date, time, text, and URL. This data is then preprocessed to remove all NULL values, arrange them in order and generate visualizations of the data. Data processing is carried using following method.

First data is processed by loading, distribution, preprocessing and cleaning fours stages. Clearning process takes place by removing: URLs, usernames, tweets with not available text, special characters, and numbers.

In fifth stage, text processing is performed by tokenizing, transforming to lowercase, and stemming. In the sith and final stage a world list is build for Bag-of-Words.

A cumulative frequency data plot is shown in Figure 1.

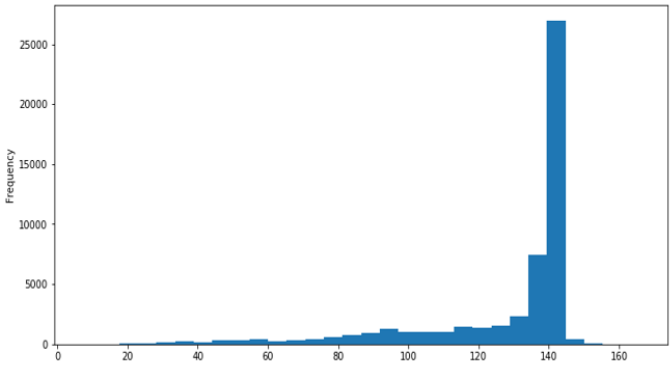

Fig. 1. Cumulative Frequency Data Plot

This preprocessed data will be used to implement machine learning approaches on them. We have used three Artificial

intelligence tools in this paper i.e. Linear regression, LSTM, and DNN Regressor, and provide a comparison of their performance in price prediction of bitcoins.

For the implementation of these tools, we have used several libraries most important of them are; pandas library that is used for efficient dealing of data using data frames and columns. It is best to use for data manipulation and analysis which is needed in our case. We have also used NumPy which makes doing numerical calculations and manipulation very easy which is important for machine learning. For machine learning algorithms we have used TensorFlow and Keras library which has inbuilt functions for many tasks and creating your own model is also very easy in them. In the end, we created visualizations of our results using libraries like Matplotlib and Seaborn that can help in creating different kinds of graphs and heatmaps.

IV. Evaluation Models

A. Model 1: Linear Regression

Linear regression tries to fit the best line on data points and predicts future values based on given values of data. Linear regression is one of the simple classifiers and even then, it gives extraordinary results sometimes. For the implementation of linear regression, we have built-in functions of TensorFlow and then customized some according to our own needs. We have divided data into 80-20 split where 80% of data is used for training purposes and 20 percent for testing purposes. Learning rate, which is the rate with which data converges, was kept at 0.02.

B. Model 2: DNN Regressor

DNN Regressor means the deep neural network that is used for regression problems. It is very effective as it develops connections on its own and learns the most useful patterns that can generate efficient results. We have used 10 hidden layers to formulate our network and each layer is fully connected to the other layer. Data split is kept with 80 to 20 ratios meaning 80% training data and 20% test data are used. Data is then converted to features and it is passed to the TensorFlow model of nn-regression-model for training then model is tested on the rest of the data.

C. Model 3: LSTM

Long Short Term Memory networks are unlike traditional neural networks use feedback mechanisms and they remember the previous state because of these connections. That is why we say that they can retain previous information unlike traditional neural networks and use them to predict the future outcome, in our case future price of bitcoins. These are relatively new, and their performance is increasing with time. We have used them in our comparison. We have used TensorFlow for model creation and prediction of data. Again, we have used an 80-20 split for the distribution of data among training data and test data.

V. Performance Results

To find the best model for bitcoin price prediction, we find the least root mean square error values of all the three tools and compare them. The three tools initially gave us the following test results using TensorFlow.

A. Linear Regression

Results of trained and tested data obtained from the TensorFlow model at the learning rate of 0.02 provided us with outputs shown in Figure 2 and Figure 3.

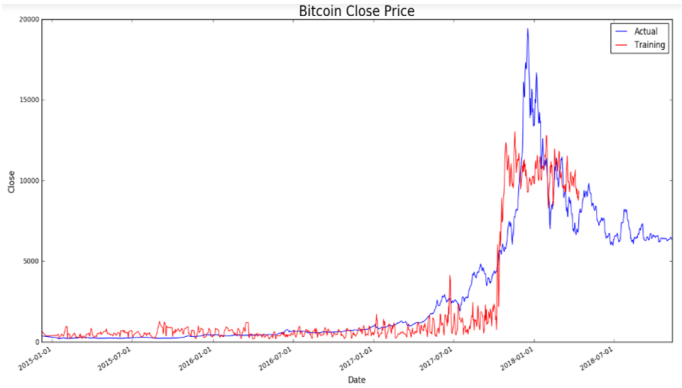

Fig. 2. Linear regression test training results

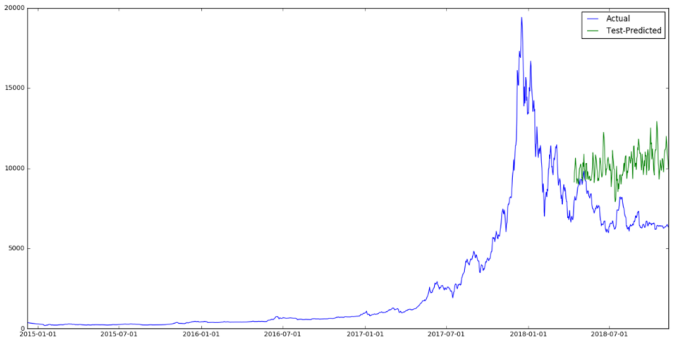

Fig. 3. Linear regression test testing results

B. DNN Regressor

The results of trained and tested data from the TensorFlow model of nn-regression-model gave us output shown in Figure 4.

C. LSTM

Using TensorFlow for model creation and prediction of data, through 80-20 split for the distribution of data among training data and test data, the same distribution used in the above two tools, we obtained some results. Now, to compare the least root mean square error values from the above three tools, we state them in the Table I shown below:

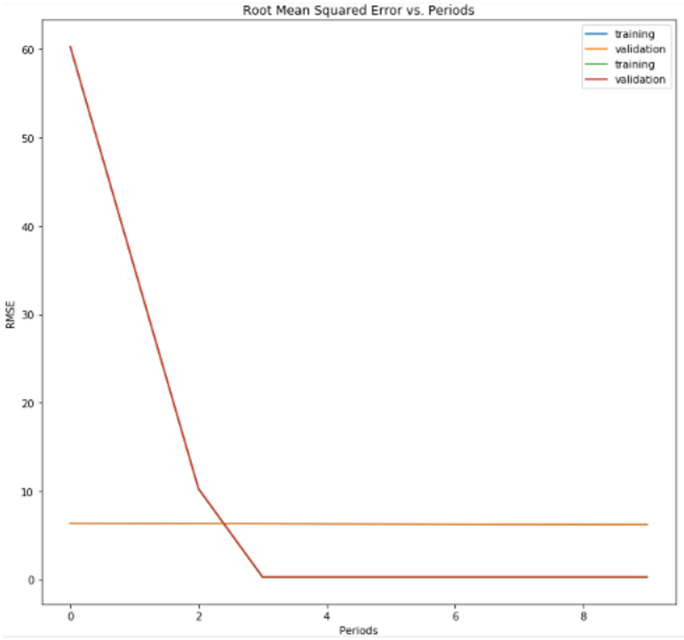

Fig. 4. DNN regressor training and validation (or testing) results

TABLE I
MODEL RESULTS COMPARISON

Model	Training Data	Test Data
[HTML]F2F2F2 **Linear Regression**	7998.24	7994.50
LSTM	53.370	53.372
[HTML]F2F2F2 **DNN Regressor**	6.20	6.21

VI. CONCLUSION AND FUTURE WORK

Various text analysis and machine learning techniques are used to mine opinions from a document. Financial Markets Public opinion regarding companies can be used to predict the performance of their stocks in financial markets. If people have a positive opinion about a product that a company A has launched, then the share prices of A are likely to go higher and vice versa. Public opinion can be used as an additional feature in existing models that try to predict market performances based on historical data. Other features of Bitcoin can also be considered for future work for better results. Microeconomic factors might be included in the model for a better predictive result. Furthermore, a breakthrough evolution in peer-to-peer transactions is ongoing and transforming the landscape of payment services.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea-Grant funded by the Korean Government (Ministry of Science and ICT-NRF-2020R1A2B5B02002478).

REFERENCES

[1] Sengupta, Sarthak, and Anurika Vaish. "Social networking mood recognition algorithm for conflict detection and management of Indian educational institutions." Social Network Analysis and Mining 10, no. 1 (2020): 1-13.

[2] Garg, Prateek, and Vineeta Guide Bassi. "Sentiment analysis of twitter data using NLTK in python." PhD diss., 2016.

[3] Gupta, Bhumika, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, and B. Tech. "Study of Twitter sentiment analysis using machine learning algorithms on Python." International Journal of Computer Applications 165, no. 9 (2017): 29-34.

[4] Liu, Ruowei, Xiaobai Yao, Chenxiao Guo, and Xuebin Wei. "Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach." Annals of GIS 27, no. 1 (2021): 43-56.

[5] Bouras, Dalila, Mohamed Amroune, Hakim Bendjenna, and Nabiha Azizi. "Techniques and Trends for Fine-Grained Opinion Mining and Sentiment Analysis: Recent Survey." Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) 13, no. 2 (2020): 215-227.

[6] Chintalapudi, Nalini, Gopi Battineni, Marzio Di Canio, Getu Gamo Sagaro, and Francesco Amenta. "Text mining with sentiment analysis on seafarers' medical documents." International Journal of Information Management Data Insights 1, no. 1 (2021): 100005.

[7] Reddy, Niveditha N., and E. Naresh. "Predicting Stock Price Using Sentimental Analysis Through Twitter Data." In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), pp. 1-5. IEEE, 2020.

[8] Saglani, Khushboo H., and Nitin J. Janwe. "Machine Learning Based Sentiment Analysis on Twitter Data." International Journal of Emerging Trends in Engineering Research (IJETER) 8, no. 8 (2020).

[9] Sharma, Dipti, Munish Sabharwal, Vinay Goyal, and Mohit Vij. "Sentiment analysis techniques for social media data: A review." In First International Conference on Sustainable Technologies for Computational Intelligence, pp. 75-90. Springer, Singapore, 2020.

[10] Ruz, Gonzalo A., Pablo A. Henríquez, and Aldo Mascareño. "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers." Future Generation Computer Systems 106 (2020): 92-104.

[11] Liang, Huizhi, Umarani Ganeshbabu, and Thomas Thorne. "A dynamic bayesian network approach for analysing topic-sentiment evolution." IEEE Access 8 (2020): 54164-54174.Pavel Ciaian, Miroslava Rajcaniova, and dArtis Kancs. "The Economics of Bitcoin Price Formation", Applied Economics. pp. 1799-1815, 13 Nov (2015).

[12] B. Liu, "Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing," Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA, (2009).

[13] Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. Comput. Linguist, 37, 267–307, (2011).

[14] Jagdale, O.; Harmalkar, V.; Chavan, S.; Sharma, N. Twitter mining using R. Int. J. Eng. Res. Adv. Tech, 3, 252–256, (2017).

[15] AceSoucy, P.; Mineau, G.W. A simple knn algorithm for text categorization. In Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, pp. 647–648, 29 November–2 December (2001).

[16] Willett, P. The porter stemming algorithm: Then and now. The program, 40, 219–223, (2006).

[17] Kanavos, A.; Nodarakis, N.; Sioutas, S.; Tsakalidis, A.; Tsolis, D.; Tzimas, G. Large scale implementations for twitter sentiment classification. Algorithms, 10, 33, (2017).

[18] Kristoufek et al, "BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era", Scientific Reports volume 3, Article number: 3415 (2013).

[19] Pavel Ciaian, Miroslava Rajcaniova, and dArtis Kancs. "The Economics of Bitcoin Price Formation", Applied Economics. pp. 1799-1815, 13 Nov (2015).

[20] Park, Cyn-Young, Grace Tian, and Bo Zhao. "Global bitcoin markets and local regulations." Asian Development Bank Economics Working Paper Series 605 (2020).

[21] Shahzad, Khurram, Arshad Ali, and Nasir D. Gohar. "ETSP: An energy-efficient time synchronization protocol for wireless sensor networks." In 22nd International Conference on Advanced Information Networking and Applications-Workshops (Aina Workshops 2008), pp. 971-976. IEEE, 2008.

[22] Cavalli, Stefano, and Michele Amoretti. "CNN-based multivariate data analysis for bitcoin trend prediction." Applied Soft Computing 101 (2021): 107065.