2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu (Nepal)

Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis

Dibakar Raj Pant, Prasanga Neupane, Anuj Poudel, Anup Kumar Pokhrel, Bishnu Kumar Lama

Department of Electronics and Computer Engineering, Central Campus Pulchowk
I.O.E, Tribhuvan University
Lalitpur, Nepal
071bct525prasanga@pcampus.edu.np

***Abstract* - The sentiment in Twitter about Bitcoin have direct or indirect influence on overall market value of the Bitcoin. This research is concerned with predicting the volatile price of Bitcoin by analyzing the sentiment in Twitter and to find the relation between them. The tweets of Bitcoin collected from different news account sources are classified to positive or negative sentiments. The obtained percentage of positive and negative tweets are feed to RNN model along with historical price to predict the new price for next time frame. The accuracy for sentiment classification of tweets in two class positive and negative is found to be 81.39 % and the overall price prediction accuracy using RNN is found to be 77.62%.**

Index Terms – Bitcoin, Sentiment, Tweeter, Classified, RNN.

I. INTRODUCTION

The term cryptocurrency [1] is a new and emerging topic in today's world. Cryptocurrency is digital currency governed by cryptographic protocol which uses Blockchain [1]. This concept of cryptocurrency has revolutionized the way we think about money. The continuous increase in adoption and widespread usage has increased its value in real world applications by substantial amount. Various cryptocurrencies have been invented since 2009 but the first one to be launched as a cryptocurrency was Bitcoin [2]. It is a form of electronic cash with no governing financial institution which can be used for online transactions or as exchange between any two parties. Nowadays due to its fluctuating and big-ticket value lion's share of bitcoin transactions occurs in exchange as a stock market rather than in online merchant transactions.

However, it does not have central governing authority and is controlled by the general public. By this reason, Bitcoin is considered a very volatile currency and its price is affected by socially constructed opinions. In the work of Kristoufek [3] it is shown that some of the extreme drops as well as price increases in the Bitcoin exchange rate coincided with dramatic events in China. In another research carried out by American Institute for Economic Research (AIER) [4] shows a major fluctuation in price of bitcoin driven by the impactful news and sentiment over the world during the time period between 2016 and 2017.

The work of J. Bean [5] provides Twitter opinion mining idea in order to visualize the general customer attitude and satisfaction towards the airline company. Further, Nagar and Hahsler [6] suggests a strong correlation exists between the sentiment of news extracted the news corpus and the stock price movement.

Colianni et. al. [7], have used Naive Bayes to find optimal time to trade by correlating prices with Twitter. Pagolu et. al. [8] work on predicting stock through twitter sentiment presents strong correlation between Twitter sentiment and stock price movement.

In this article, new approach of combining the sentiment score with historical price to predict future price is implemented. The system pre-processes the tweets and feeds to sentiment analyser. The sentiment analyser gives sentiment percentage of the day which is feed to RNN predictor along with the historical price of bitcoin. The RNN model then finally predicts the price.

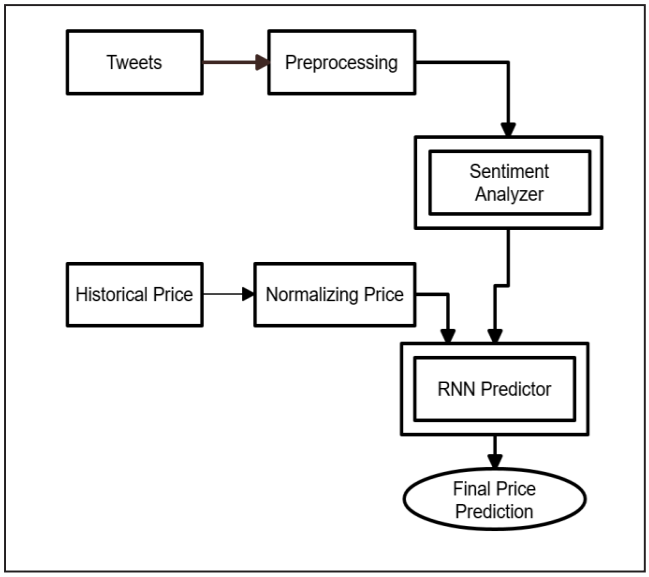

Fig 1. System Flow diagram

978-1-5386-6227-4/18/$31.00 ©2018 IEEE

II. Data Collection And Preprocessing

The Tweeter accounts [9] like BitcoinNews(@BTCTN), CryptoCurrency(@cryptocurrency), CryptoYoda (@CryptoYoda1338), BitcoinMagazine (@BitcoinMagazine), Bitcoin Forum (@BitcoinForums), CoinDesk (@coindesk) and Roger Ver (@rogerkver) are the major sources for the pool of tweets collected from January 1 of 2015 to December 31 of 2017. The price data is collected for the same time period from Coinmarketcap [10].

A. Dataset Creation

The collected tweets are labelled manually as 'p' for positive, 'n' for negative and 'i' for irrelevant or neutral. Total of 2585 positive, 1669 negative and 3200 irreverent tweets are labelled manually (dataset in Appendix-A).

B. Removing Repeated and Irrelevant Tweets

The irrelevant and repeated tweets for example promotional and advertising belonging to different sites and their related other accounts are removed by using FuzzyWuzzy [11] method. They are further processed to word tokenization and stop words filtering.

C. Regex and Weighted Search

Regex [12] search is applied to avoid hyperlinks and different kinds of emojis from tweet. Further Stanford Named Entity Recognizer(NER) [13] along with Regex is used to extract the names of persons, organizations and country present in tweet and it is later used for giving double weight to its sentiment if the extracted names are listed in impactful groups index (NER impactful index in Appendix-B).

III. Feature Extraction

For text classification the two methods of feature extractions Word2Vector [14] and Bag-of-Words [15] are used.

A. Word2Vector

Genism Word2Vector is used for 300 dimensional vector representation of each words present the pool of tweet. The resultant vector from each words of tweet is used as a feature.

B. Bag-of-Words

The frequency distribution of each word in the pool of tweets are used as a feature. The most common words after stop word filtering are regarded as the pivotal words and given its frequency score.

IV. Sentiment Analysis And Correlation

A. Sentiment Analysis

The features extracted from both methods for 4,254 manually labeled tweets are trained with five different algorithms Naïve Bayes [16], Bernouli Naïve Bayes, Multinomial Naïve Bayes, Linear Support Vector Classifier [17] and Random Forest [18]. A voting classifier is created which takes output of each of these algorithms (ie. positive and negative) and then classifies the new tweet to that class for which the vote is maximum. Thus classified tweets are subjected to NER search for presence any impactful keywords and given weights accordingly.

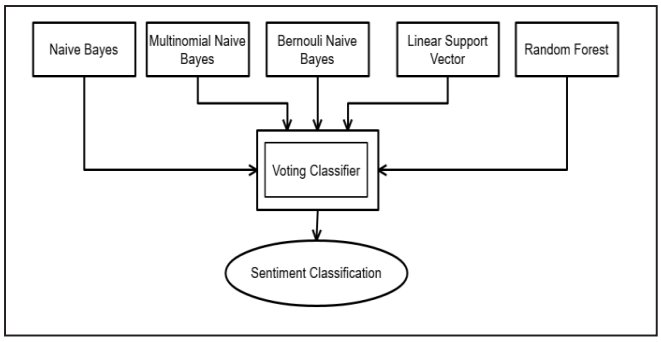

Fig 2. Voting classifier for sentiment classification

The voting classifier consists of following classifiers:

1) Naïve Bayes: It is a conditional probability model which assumes that features are statistically independent of one another. Mathematically,

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (1)$$

Where,
$p(C_k)$ represent prior probability of class C_k
$p(x|C_k)$ represent class conditional feature probability
$p(C_k|x)$ represent probability x belonging to class C_k

Different models of Naïve Bayes like Multinomial and Bernouli are also used.

2) Linear Support Vector: It is a classifier that classifies by constructing hyperplanes which separates the cases that belong to different categories.

$$C(x) = \begin{cases} 1, & w.\emptyset(x) + b \geq k \\ -1, & w.\emptyset(x) + b \leq -k \end{cases} \quad (2)$$

where, $w=\{w1,...,w_n\}$ represent a weighted vector,
$x=\{x1,...,x_n\}$ represent input and $\emptyset(x)$ is a kernel function

3) Random Forest: It is popular ensemble learning algorithm generally used for classification related tasks. They first create a multitude of decision trees during the time of training and then predict the final output class which comes out to be the mode of all the predicted class from the individual trees.

$$R = \{h(x|\emptyset_1), \ldots\ldots\ldots, h(x|\emptyset_k)\} \quad (3)$$

$$Y = MODE\{h(y_1), \ldots\ldots\ldots, h(y_k)\} \quad (4)$$

Where $h(x|\emptyset)$ is a decision tree with parameter $\emptyset$

And $h(y)$ is the predicted class of decision tree,

R is the ensemble of decision tree $h(x|\emptyset)$,

and Y is the class with maximum votes

B. Correlation with Price

The Pearson Correlation coefficient test is a measure of the linear correlation between two variables.Mathematically,

$$r = \frac{cov(X,Y)}{std.(X) * std.(Y)} \quad (5)$$

Where, $cov(X,Y)$ is covariance

std.(X), std.(Y) is the standard deviation

r is the Pearson coefficient between X and Y

The tweets from January 1, 2018 to June 30, 2018 are collected and sentiment score of each day is calculated. For time lag between the impact of sentiment on its price, cross correlation test is performed which showed the lag of one day meaning that sentiment of today has impact in price of tomorrow. Further, Pearson correlation test is performed with the sentiment score and corresponding price of the next day.

V. RNN PREDICTOR

For the time series prediction Recurrent Neural Network and its variation like LSTM and GRU is used. The figure below presents the model of RNN network where first layer is input layer and middle one as GRU layer and the last is dense layer as output layer.

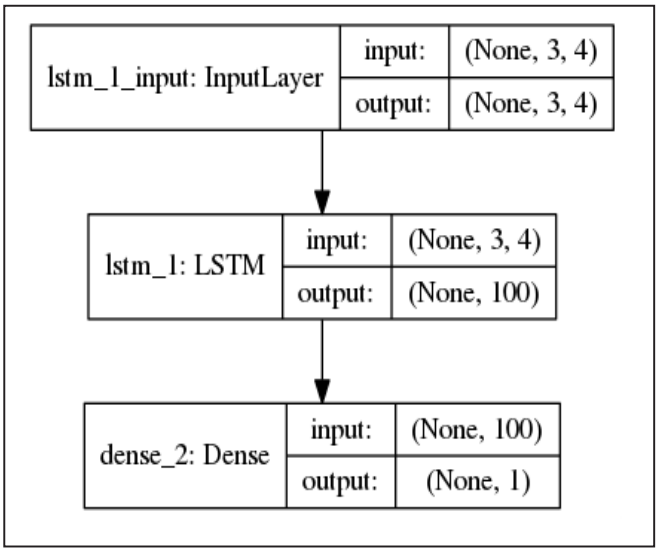

Fig 3. RNN with LSTM model

VI. RESULTS AND ANALYSIS

To choose between two features extraction method same model of Random Forest is trained with both feature set independently.Word2Vector feature yielded the classification accuracy of only 69.82% whereas Bag-of-words features yielded the accuracy of 78.49%.

Table I. Bag-of-Words and Word2Vector Comparison

Feature	Parameter (n_estimator)	Accuracy
Bag-of-Words	N = 100	78.12 %
Word2Vector	N = 100	68.33 %
Bag-of-Words	N = 120	78.49 %
Word2Vector	N = 120	69.82 %
Bag-of-Words	N = 150	77.95 %
Word2Vector	N = 150	68.61 %

This reveals that for the lower number of dataset and at sentence level sentiment classification Word2Vector does not perform well so Bag-of-words is preferred.

The overall accuracy of sentiment classification by voting classifier with validation split of 1:3 is achieved as 81.39% as shown in table below:

Table II. Confusion Matrix

Accuracy	Precision	Recall	F-Measure
81.39%	82.90%	84.86%	83.86%

Confusion Matrix for Voting Classifier of Sentiment Classification

The Pearson Correlation Coefficient calculated between the sentiment score and percentage price change of the next day between the period of January 1, 2018 to June 30, 2018 is given in the table below:

Table III. Sentiment and Price Correlation

Price Fluctuation	***Pearson Correlation Coefficient***
More than 2 %	Negative Sentiment: 0.34
	Positive Sentiment: 0.21
More than 4 %	Negative Sentiment: 0.41
	Positive Sentiment: 0.26

Pearson Correlation Coefficient Comparison for different range of price

As Table-III shows, for the price fluctuations more than 4% (ie. more than $500 during that study period) in a single day, Pearson correlation coefficient is found to be 0.41 for negative sentiment and corresponding fall in price and 0.26 between the positive sentiment and its corresponding increase in price. This shows there is a moderate (according to Evans 1996) [19] correlation between rise of negative sentiment and consequent fall in price of Bitcoin but a weak relation between increase of positive sentiment and consequent increase in price.

The price prediction accuracy for RNN model is found to be 77.62%. The figure below shows a comparative plot between the predicted and actual price of Bitcoin.

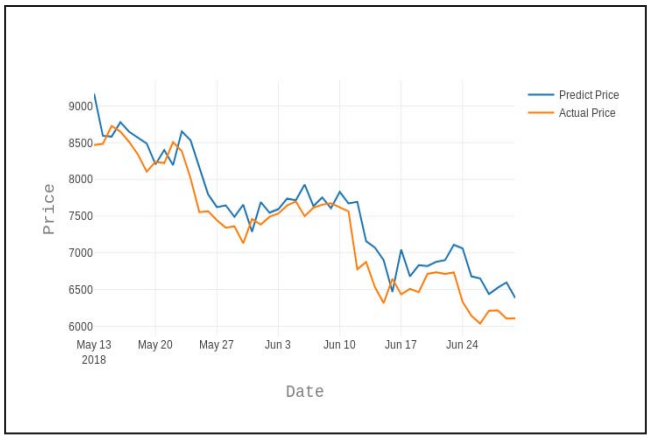

Fig. 4. Predicted price and Actual price comparison.

The two figures below shows the accuracy of the RNN model:

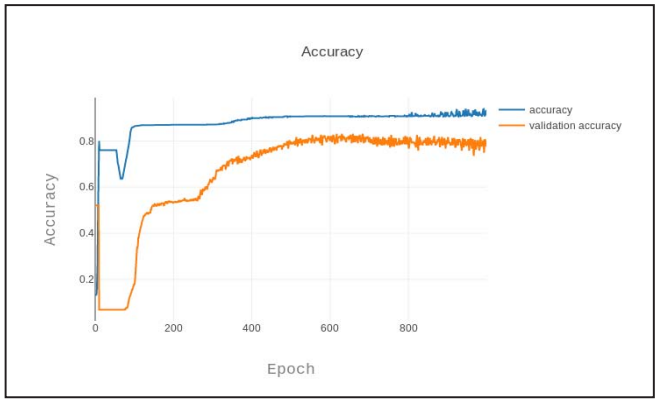

Fig.5. Validation Accuracy Vs Epoch.

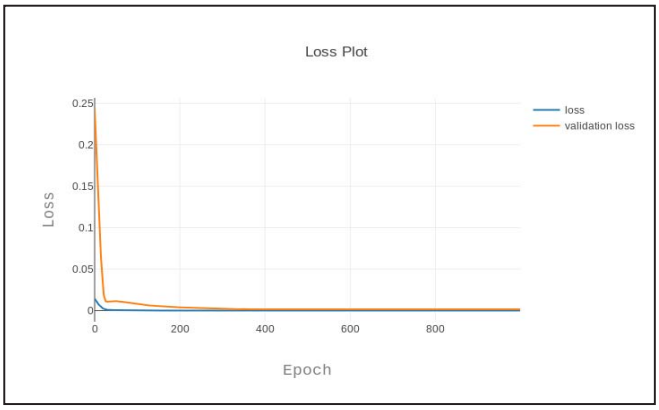

Fig.6. Validation Loss Vs Epoch.

VII. Conclusion

The major contribution of this work is a sentiment analyser which can distinguish between the positive and negative tweets of Bitcoin over the Twitter with the accuracy of 81.39%. Further the RNN model which can predict the price of Bitcoin the following day taking consideration of historical price and the positive and negative sentiment scores with the accuracy of 77.62% is another useful work. It discusses the different methods of feature extraction and provides a comparison between them. It also shows a moderate correlation of 0.41 between rise of negative opinions in the Twitter related to Bitcoin and its consequent fall in price.

Acknowledgment

The authors would like to acknowledge Mr. Sandip Pandey for his laudable contribution during this research work.

References

[1] U. W. Chohan, “Cryptocurrencies: A Brief Thematic Review”, SSRN Electronic Journal, 2017.
[2] S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system., 2008.
[3] L. Kristoufek, "What are the main Drivers of the Bitcoin Price? Evidence from Wavelet Cohernce Analysis", 2015.
[4] “Bitcoin largest price changes coincide major News events about Cryptocurrency” [Online]. Available: www.aier.org [Accessed 13 July 2018]
[5] J. Bean, "R by example: Mining Twitter for consumer attitudes towards airlines", 2011.
[6] A. Nagar and M. Hashler, "Using text and data mining techniques to extract stock," vol. XX, 2012.
[7] S.Colianni, S. Rosales and M. Signorotti, “Algorithmic trading of cryptocurrency based on Twitter sentiment analysis.” , 2015
[8] V. Sasank Pagolu, K.N. Reddy,G. Panda and B. Majhi "Sentiment analysis of Twitter data for predicting stock market movements",SCOPES, 2016.
[9] Tweeter Accounts:
‘BitcoinNews’ Available: https://twitter.com/BTCTN?lang=en
‘CryptoCurrency’ Available: https://twitter.com/cryptocurrency?lang=en
‘CryptoYoda’ Available: https://twitter.com/CryptoYoda1338?lang=en
‘BitcoinMagzine’Available: https://twitter.com/bitcoinmagazin?lang=en
‘BitcoinForum’Available:https://twitter.com/BitcoinForumCom?lang=en
‘CoinDesk’ Available: https://twitter.com/coindesk?lang=en
‘RogerVer’ Available: https://twitter.com/rogerkver?lang=en
[10] Coin Market Cap, Available: https://coinmarketcap.com/ [Accessed 22 July 2018]
[11] FuzzyWuzzy, "Geeksforgeeks.org," [Online]. Available: https://www.geeksforgeeks.org/fuzzywuzzy-python-library/.[Accessed 24 July 2018].
[12] C. Frenz, “Introduction to Searching with Regular Expressions”, Proceedings of the 2008 Trenton Computer Festival, 2008
[13] Stanford Named Entity Recognizer(NER), “Stanford.edu”, [Online] Avilable: https://nlp.stanford.edu/software/CRF-NER.shtml [Accessed 22 July 2018]
[14] “Word to Vector”,[Online]. Available: https://towardsdatascience.com/word-to-vectors-natural-language-processing-b253dd0b0817 [Accessed 25 June 2018]
[15] M. McTear, Z. Callejas and D. Griol, “The Conversational Interface”, Springer, 2016.
[16] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach,1995.
[17] C. Cortes and V. Vapnik, "Support-vector networks". Machine Learning,1995
[18] T. Kam Ho, Random Decision Forests. Proceedings of ICDAR, Montreal, 1995.
[19] J. D. Evans, Straightforward statistics for the behavioral sciences, Brooks/Cole Publishing, 1996

APPENDIX-A

Some example of positive and negative labelled Tweets during creation of Dataset

Positive Tweets	Negative Tweets
1. Church in Z rich Accepts Donations in Bitcoin, BCH, Ether, Ripple and Stellar 2. Overall Capital of Crypto Markets Exceeds $750 Billion. 3. Turkish Minister Proposes National Cryptocurrency. 4. Amazon set to use Bitcoin as payment for people.#Bitcoin 5. Australian Gold Refinery Announces Plan to Develop Cryptocurrency #Bitcoin	1. Trader in Chicago Firm Stole Million BTC and Faces 20 Year Sentence #Bitcoin 2. Cryptocurrency Regulator Found Dead at His Home in South Korea #Bitcoin 3. Lawyers Discuss Challenges Posed by Cryptocurrencies During Divorce #Bitcoin 4. Scammers Are Ruining Crypto Twitter and Twitter Is to Blame #Bitcoin 5. US Navy Bust Bitcoin Drug in Naval Academy #Bitcoin

APPENDIX-B

Implemented NER impactful index list for giving more weight to the sentiment of tweets.

Country (According to average market Capital for first six months of 2018)	**Organizations and Authority** (According to frequency and reputation)	**Person and Authority** (According to frequency and reputation)
Russia	Facebook	Putin
USA	Central Bank	President
China	Financial Authority	CEO
South Korea	State Government	-
Japan	-	-