Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis

4 authors, including:

Otabek Sattarov
Konkuk University
4 PUBLICATIONS   **5** CITATIONS

Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis

Otabek Sattarov
Department of Computer Engineering
Konkuk University
Chungju, Rep. of Korea
otabeksattarov95@gmail.com

Heung Seok Jeon
Department of Computer Engineering
Konkuk University
Chungju, Rep. of Korea
hsjeon@kku.ac.kr

Ryumduck Oh
Department of Software and IT convergence
Korea National University of Transportation
Chungju, Rep. of Korea
rdoh@ut.ac.kr

Jun Dong Lee
Department of Multimedia Engineering
Gangneung-Wonju National University
Gangneung, Rep. of Korea
jlee@gwnu.ac.kr

Abstract - **Bitcoin is one of the main phenomena in recent times together with other cryptocurrencies due to the redefinition of the money term and its price fluctuations. Moreover, scientists are increasingly recognizing Twitter's predictive power for a wide range of events, and particularly for financial markets. This article examines to what degree Bitcoin returns can be estimated using public opinion on Twitter. Using a sentiment analyzer on Bitcoin-related tweets and financial data, the Twitter sentiment was found to have predictive power for Bitcoin's results. Once again, our findings confirm the presence of a correlation between them. We observed 62.48% accuracy when making predictions based on bitcoin-related tweet sentiment and historical bitcoin price.**

Keywords – Cryptocurrency, Bitcoin, Social media, Twitter, Sentiment analyzing, Text mining, Natural language processing.

I. INTRODUCTION

Bitcoin is a decentralized electronic currency system, which after its development by Satoshi Nakamoto brought about a huge shift in the financial system[1]. It reflects an advancement in IT that is focused on peer to peer networks and cryptographic protocols. Bitcoin, because of its decentralized existence and electronic process, is not regulated or controlled by any government or bank. Bitcoin's main aim, or goal, is to promote goods and services transactions. Bitcoin has evolved immensely and managed to attract vast numbers of users, gaining immense popularity because of its regular media mention and dissemination. Because of its popularity, the Bitcoin price, which is constantly fluctuating in real-time like a stock exchange, it's very curious to create a model that can predict Bitcoin's price on real-time using internet social media data. If it can be predicted at a reasonable degree of precision, it can be useful for investors, business people, banks, organizations, etc. that use Bitcoin for transactions. Figure 1 indicates the market share of the cryptocurrencies[2].
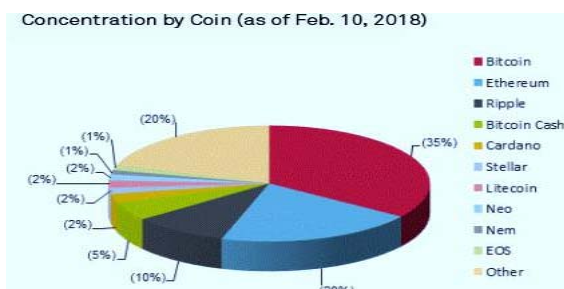

Figure 1. Share percentage cryptocurrencies in market.

*Corresponding author. Tel.: + 82 43 840 362l.
E-mail address: hsjeon@kku.ac.kr (Heung Seok Jeon).

The Internet has evolved massively in the past decade, exchanging information and experiences has become simple with the advent of communication networks including, social media such as Twitter, Facebook, Instagram, blogs, etc. Hundreds of thousands of Twitter users, for example, are creating huge amounts of Bitcoin-related tweet data every day. This enormous data can be useful to research Bitcoin patterns using technology such as Machine Learning, Natural Language Processing, Time Series Analysis, etc. using the continuously generated real-time data from social networking sites. Since Bitcoin is a new phenomenon, it's still new to apply Machine learning and Deep learning with greater accuracy and speed using social media data. Although there is much research surrounding the use of various machine learning techniques for predicting time series, there is a lack of research in this field directly related to Bitcoin. This study could fill the gap by developing a model with the Bitcoin history price and Twitter Sentiment for successful prediction of Bitcoin's future price

We address this task by downloading history prices of Bitcoin and collecting Bitcoin-related Tweets, followed by performing tweet sentiment analysis. Before using the average scores of sentiment analysis to predict the direction of price for Bitcoin, we calculated the relationship between Tweet sentiment and price direction. We applied the Random Forest Regression binary classification model with diverse inputs and evaluated outputs. We are able to boost the accuracy of our model based on the outputs.

The remainder of this paper is organized as follows. Section II involves related work that has been done by other researchers. Methodology and analyzing part of the proposed module will be discussed in Section III and Section IV, respectively. Finally, Section IV concludes our job.

II. RELATED WORK

Social media has been the main source of information about cryptocurrencies, which can be categorized into Twitter, forums related to cryptocurrency, and news sources related to cryptocurrency. In addition to numerous news sources[7], researchers have used messages from forums such as Reddit and Bitcointalk.org[3-6] to conduct sentiment analysis and forecast price volatility in Bitcoin. Most researchers recognize the predictive power of social media and news sentiment for Bitcoin prices and/or volume of trading over the short (1–7 days) and long (30–90 days) span. Even the number of posts or messages correlates with the trading volume of Bitcoin[3]. Additionally, Karalevicius et al.[7] confirm what was mentioned earlier; cryptocurrency investors tend to overreact to news leading to a market trend where the market shifts initially with the sentiment and is then slowly fixed. In addition, Phillip et al.[8] use the epidemic models and the Reddit subject pages to forecast market bubbles and movements for Bitcoin, Ethereum, Litecoin, and

Monero accurately. Analysis of the Twitter sentiment was used in various experiments to forecast price fluctuations for Bitcoin.

In a study conducted by Georgoula et al.[9], using Twitter sentiment analysis, a Support Vector Machine (SVM), and various regression models were used to forecast price fluctuations for Bitcoin. The authors got an 89.6 percent accuracy and found only a short-term correlation between positive Twitter sentiment and the price of Bitcoin. Garcia and Schweitzer[10] use a Vector Autoregressive (VAR) model and Granger-causality testing lexicon-based approach to finding that changes in Twitter sentiment polarity precede Bitcoin price fluctuations. Prices in the market for cryptocurrencies will differ considerably across exchanges, making such a result unclear.

In another research, Pagolu et al.[11] used social media microblogging to forecast stock price, as it accurately reflects the public's feelings and opinions about current events. They applied sentiment analysis and supervised principles of machine learning to Twitter extracted tweets and studied the association between a company's stock price movements and tweet sentiments. The researchers concluded that the paper indicates that there is a clear connection between the rising and falling stock prices with the tweeting public sentiments.

Mata et al.[12] discussed in case the increase in the cost of the Bitcoin is linked to the number of tweets or the results of Web Search media. They compared cost patterns with Google Trend Data, tweet volume, and in particular with positive tweets. They gathered more than 1,900,000 tweets, analyzed them then marked the positive and negative tweets. They found a correlation between Bitcoin price and the number of tweets at the same time, between Bitcoin Price and Positive Tweets, and Bitcoin Price with Google Trend.

Stenqvist and Lönnö[13] researched whether an examination of sentiments on Twitter data would help to suggest price fluctuation for Bitcoin. By analyzing more than two million Bitcoin-related tweets within one month (daily analysis), they reported that applying machine learning to the study might have a connection beyond the number of tweets.

III. METHODOLOGY

We need a time stamp on the price of Bitcoin and on the tweets to do time series analysis. We have used APIs and a little bit of web scraping for the collection of tweets. We have collected tweets for 60 days starting from the 12th of March 2018 to the 12th of May. The main reason for choosing the above time period is within this time price of Bitcoin observed skyrocketing and falling down as well in which gives us to evaluate our model effectively. Figure 2 illustrates the price movements of Bitcoin at the above time period.
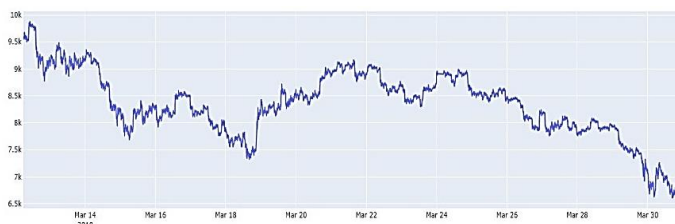

Figure 2. Price movement of Bitcoin from the 12th of March 2018 year to 12th of May 2018 year.

The dataset given has 92550 tweets that were posted almost every minute within the time period mentioned above.

The dataset is a combination of expressions, emoticons, symbols, URLs, and user's mentions. Raw tweets scraped from twitter usually give rise to a noisy dataset. This is because of the casual nature of social media use by people. Tweets have certain unique features, such as retweets, emoticons, user references, etc. that have to be extracted correctly. For this purpose, raw twitter data must be structured to create a dataset that can be easily learned by different classifiers. To standardize the dataset and reduce its scale, we have implemented an extensive number of pre-processing steps. We first do some general pre-processing on tweets which are as follows.

- Convert the tweet to the lower case.
- Replace 2 or more dots (.) with space.
- Strip spaces and quotes (" and ') from the ends of tweet.
- Replace 2 or more spaces with a single space.
- Reducing user mentioned places.
- Turning hashtags to the typical words.
- Classifying emoticons and removing retweets.

Example of the preprocessing techniques is shown in Table 1.

TABLE I.   PREPROCESSING STEPS WITH RESULT

	Processing technique	Result
0.	Original Tweet	RT@bitcoin https://twitter.com/FT/status/10226050 Bitcoin ETF rejected but buuuuuy!!! Ask yourself why you aren't buying lol, tomorrow it will reach 8000 #BUY #NOW #BITCOIN
1.	Remove ''RT'' if present	@bitcoin https://twitter.com/FT/status/10226050 Bitcoin ETF rejected but buuuuuy!!! Ask yourself why you aren't buying lol, tomorrow it will reach 8000 #BUY #NOW #BITCOIN
2.	Remove URLs, excess (white) space and mentions	Bitcoin ETF rejected but buuuuuy!!! Ask yourself why you aren't buying lol, tomorrow it will reach 8000 #BUY #NOW #BITCOIN
3.	Reduce character sequences >3 to 3	Bitcoin ETF rejected but buuuy!!! Ask yourself why you aren't buying lol, tomorrow it will reach 8000 #BUY #NOW #BITCOIN
4.	Apply case-folding	bitcoin etf rejected but buuuy!!! ask yourself why you aren't buying lol, tomorrow it will reach 8000 #buy #now #bitcoin
5.	Remove hashtags if not in Reuters corpus	bitcoin etf rejected but buuuy!!! ask yourself why you aren't buying lol, tomorrow it will reach 8000 buy now

We have used Python's Quandl module API on that particular day to collect the Bitcoin price. Instead of gathering data from a single source at Bitcoin Price, we are collecting the Bitcoin price from 4 big websites- "BITSTAMP," "COINBASE," "ITBIT," "KRAKEN" that tracks the Bitcoin price. Some of those contain data that are missing. So, to get a smoother curve we can take the sum of those and divide it to four. We picked only the "Close Price" in every minute of Bitcoin from the "Opening Price," "Closing Price," "High Price," "Low Price" of Bitcoin prices.

IV. ANALYZING

Next, our goal is to apply sentiment analysis to our collected tweets to calculate these tweets' subjective emotions or opinions about Bitcoin. In particular, we used for analysis the VADER (Valence Aware Dictionary and Sentiment Reasoner)[14]. It should be noted that the sentiment analysis method of VADER is explicitly tailored to the sentiment of social media, which may show more positive results as our data is scrapped from a social media site. The sentiment feature of VADER returns a polarity score for a compound. The polarity score for VADER is between -1 and 1, wherefrom -1 to 0 is negative, 0 is neutral and 0 to 1 is positive.

To evaluate our proposed model, we chose one of the machine learning algorithms, called Random Forest Regression. By comparing other machine learning algorithms, the Random Forest Regression is quite effective with working a different kind of inputs that has not relationships with each other. The algorithm has advantages in predicting future outputs as well. We use sentiment analyzing score and history price of Bitcoin as an input data and implemented a random forest algorithm by using RandomForestClassifier from sklearn.ensemble provided by scikit-learn. We experimented using 10 estimators (trees) using both presence and frequency features. presence features performed better than frequency though the improvement was not substantial.

We present the summary results of our model's prediction as two line diagrams in Figure 3. The first line diagram is actual Bitcoin price movements in the above specified time period and the second one is a diagram made by our prediction model's output data.
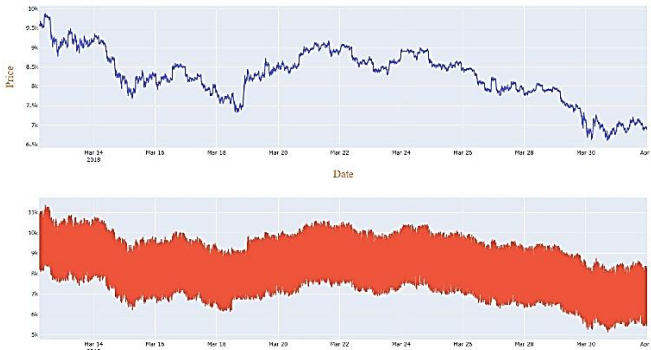

Figure 3. The first diagram is actual price movements and the second one is model's predicted movement diagram.

By considering defining the accuracy of the prediction we checked the model's outputs and made calculations for evaluating the model's performance. We tried to calculate the difference (γ) between every single prediction result (α) with the respective close price of Bitcoin (β) by subtracting to each other and taking an absolute value:

$$\gamma = |a - \beta| \qquad (1)$$

By looking at all (γ) results we found the best maximum error on prediction, the best minimum error on prediction, and the average of error. All results are shown in Table 2.

TABLE II.   STATISTICS OF ERRORS ON PREDICTION

	Definition	Value
1.	Number of tweets	92550
2.	Number of prediction data	80491
3.	Maximum value of error (%)	43.83
4.	Minimum value of error (%)	21.84
5.	Average of error (%)	37.52

Data described in Table 2 shows, during the prediction process our model lost more than 10000 data that can be useful for better performance. Module's predicted with 43.83% error most end with minimum 21.84% error accuracy.

V. CONCLUSION

Analysis of tweet sentiment is an important field of price forecasting studies. Because of the large amount of news updates per minute about Bitcoin, using Twitter in the sentiment analysis for Bitcoin is becoming a significant step for most researchers. Text mining and classification techniques on Twitter data are therefore needed which can predict the best sentiment.

This paper's main contribution is to find a partial correlation between the price fluctuation of Bitcoin and the fluctuation of the sentiment classes using a machine learning algorithm. Observed that there is a strong correlation between the Bitcoin percentage shift and Twitter sentiment.

As future research, developing a special Bitcoin sentiment lexicon could improve the connection between the sentiment analysis and the Bitcoin price fluctuation, considering other features such as hashtags, Twitter users, number of tweets, and emoticons.

REFERENCES

[1] Nakamoto S., "Bitcoin: A Peer-to-Peer Electronic Cash System", 2008.

[2] S & P Global web site.
https://www.spglobal.com/en/research-insights/articles/the-future-of-banking-cryptocurrencies-will-need-some-rules-to-change-the-game

[3] F. Mai, Q. Bai, J. Shan, "From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance", SSRN Electron. J 1-16, January 2015.

[4] Y. Kim, J. Kim, W. Kim, "Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies", PloS ONE 11(8), August 2016.

[5] Y. Kim, J. Lee, N. Park, "When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation", PloS ONE 12(5), May 2017.

[6] Xie, Peng. "Predicting digital currency market with social data: Implications of network structure and incentive hierarchy". Diss. Georgia Institute of Technology, 2017.

[7] Karalevicius, Vytautas, N. Degrande, J. De Weerdt. "Using sentiment analysis to predict interday Bitcoin price movements." The Journal of Risk Finance (2018).

[8] A. Phillip, JSK Chan, S. Peiris. "A new look at Cryptocurrencies." Economics Letters 163 (2018): 6-9.

[9] I. Georgoula, D. Pournarakis, C. Bilanakos, "Using time-series and sentiment analysis to detect the determinants of bitcoin prices." Available at SSRN 2607167 (2015).

[10] D. Garcia, F. Schweitzer. "Social signals and algorithmic trading of Bitcoin." Royal Society open science 2.9 (2015): 150288.

[11] V. Pagolu, K. Challa, G. Panda, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements", International conference on

Signal Processing, Communication, Power and Embedded System (SCOPES), 2016.

[12] M. Matta, I. Lunesu, M. Marchesi, "Bitcoin Spread Prediction Using Social and Web Search Media". Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management – 2015.

[13] E. Stenqvist, J. Lönnö, "Predicting Bitcoin price fluctuation with Twitter sentiment analysis" (2017).

[14] https://github.com/cjhutto/vaderSentiment