Bittooth: Bitcoin Value Prediction using Twitter Sentiment and Statistical Analysis

Chayisara Sakunkoo, Julian Kikuchi, Kevin Chakornsiri, Nuttaset Pattanadee

This manuscript was compiled on March 15, 2022

Bitcoin price has become volatile in recent years. A single Tweets from a single person can see a rise of 20% on the price of Bitcoin. There have been research on the correlation of Twitter sentiments and how it may affect the value of Bitcoin. But none have yet combined it with statistical analysis and developed a prediction model. We use data from two main sources: Twitter Tweets for sentiment analysis and Trading View for data analysis. The model uses data from 2020 to 2021 as a training set.

Machine Learning | Data Scraping | Prediction Model | Bitcoin

Bitcoin is a decentralized digital currency that has played a big role in the world of finance in recent years. The market value of cryptocurrencies in May 2021 is about 2.4 trillion, up from around 200 billion in 2019 (1). During this time, Bitcoin value fluctuates a lot depending on the investor's opinion on them; therefore, our team would like to find the correlation between people's opinions and the value of Bitcoin at any given time.

Our team will first try to predict the bitcoin value with only machine learning model. We will then try to add sentiment value from Twitter through NLP to see whether there is an improvement of the overall model. The knowledge of NLP is required for identifying the market sentiment from texts about Bitcoin. The knowledge of Machine Learning is used to predict the value of Bitcoin based on both market sentiment and past Bitcoin value. The core knowledge comes from the courses that we have studied in the past three years at Chulalongkorn University and also from the classes that we have attended during the exchange program. We will also need to consult multiple research papers to develop an appropriate set of features used to determine the prediction. Data also must be cleaned and analyzed before using them as training data for the model.

Our goal is to create a Machine Learning/Natural Language Processing-Powered tool for predicting future Bitcoin value.

Current Status of Research

There is a wide range of ways to predict Bitcoin prices. Most existing tools for cryptocurrency price prediction are available online via many platforms.

Fundamental Analysis. Fundamental analysis (FA) is an approach that determines an asset's value by looking at that asset's information. Typically, for stock prices, fundamental analysis can be done by evaluating indicators such as earnings per share or price-to-book ratio. On the other hand, cryptocurrency such as Bitcoin does not have such data. To decide whether a coin is overvalued or undervalued, one must understand where the coin derives its value from. Examples of such metrics can be the number of transactions, transaction values, active addresses, the fee paid, hash rate, the amount staked, whitepaper, liquidity, volume, and supply mechanism. More parameters for FA can be market capitalization, tokenomics, coin's team, token utility, and community size. The more market cap the coin has, the more stable the coin is. A high market cap means the coin is healthy and not prone to be manipulated by whales, which are people who own huge amounts of assets. An example of this would be the pump and dump scheme. Coin's team plays an important role as well. The management team of a coin creates trust and reliability of that coin. A large community size means more coin users in the system, which creates a large network. Since cryptocurrencies use blockchain technology, the more users there are, the more secured the network is (2).

Significance Statement

The purpose of this study is to find the correlation between Twitter sentiment and the value of Bitcoin and hopefully develop a prediction model that can accurately predict the future value of Bitcoin. The results of this study may help inform investors before doing any investment in Bitcoin.

N.T. conceived of the presented idea. K.C and N.T scraped and pulled data to perform analytics. C.S. cleaned and extract sentiments from the data. N.T. and J.K. optimize and develop machine learning model for prediction

We, the undersigned, hereby declare that this submission is entirely our work, in our own words, and that all sources used in researching it are fully acknowledged and all quotations properly identified.

All authors contributed equally to this work.

Pros	Cons
It is very suitable for long-term investment in cryptocurrencies. Since FA relies on the real value of an asset, it takes a long time for the price to meet the FA's prediction.	Metrics of fundamental analysis on cryptocurrencies are very new, unlike metrics for stock markets. There could be more metrics that people have never considered before. As a result, the value can be miscalculated. Understanding the coin's purposes and its project thoroughly is very time-consuming, not to mention that the coin's information is not public and easy to find. It is not recommended for short-term investment at all costs. Cryptocurrency prices fluctuate very much. This market has high volatility.

Table 1. Pros and cons of fundamental analysis (3)

Technical Analysis. Technical analysis (TA) is an approach to predict an asset's value by studying its price and volume over time series. There are two well-known common ways of TA: chart patterns and statistical indicators. The chart patterns are shapes implemented within a price chart that can suggest what the price will be. The chart patterns are based on studying shapes that have been forming in every asset. Examples of the patterns are triangle, head and shoulder, and cup and handle. The chart patterns can be applied to find the resistance and support levels on a chart so that a prediction can be made whether the price would rise or decline. The statistical indicators are the result of a mathematical calculation on price, volume, and time. Examples of the indicators are Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Bollinger Band, and Stochastic. The indicators give traders extra information. Most traders use both the chart patterns and the statistical indicator to predict the price movement. TA can be applied to any asset that has a price chart, including all cryptocurrencies. Since TA studies solely on price movement, most traders apply this on Bitcoin frequently (4).

Pros	Cons
TA requires fewer metrics than FA. It only requires a price chart over time, with corresponding trading volume. Since most traders do TA, there are a lot of practices and methods available to follow.	TA can be overwhelming for beginners. Even though it requires only a price chart, there are a lot of chart patterns and indicators available to use. Novice users can get overwhelmed and frustrated by mixing signals from indicators. It requires an area of expertise to understand and be able to selectively use appropriate tools. It is very vulnerable in volatile cryptocurrency markets. The prices fluctuate too much and indicators keep changing their results. Consequently, the predictions change.

Table 2. Pros and cons of technical analysis (3)

Estimation from Social Statistics. There is a system from the CoinMarketCap website that predicts cryptocurrencies prices from user-generated data. Their price estimate feature allows their users, who are mostly traders, to participate in this prediction. The system collects data, estimates prices from users, and calculates the estimated mean and median. It has estimations of every period, day, and month. It also shows the historical estimations from the past as well as the accuracy of the estimation compared to the real price. Users can take a look at the estimation online at any time.

Pros	Cons
Getting to know the prediction from traders can give a hint about how other traders in the cryptocurrency market think instantaneously.	Results are from users of a particular website, therefore, it is not a significant number. Since the sample size is too little, the prediction is not accurate enough for investment.

Table 3. Pros and cons of estimation from social statistics

Proposed Model. A prediction model that uses Twitter sentiment and statistical analysis to predict the future value of Bitcoin. Tweets will be analyzed using Natural Language Processing to determine the sentiment of Bitcoin. Past Bitcoin values and metrics will be used as features for the machine learning model. The time frame of the data that is used to train will be from 2020 to 2021. The model will also need to answer the following research questions:

- **Prediction Model:** How accurate is our machine learning model using statistical analysis?
- **Twitter Sentiment:** Which type of users (verified vs unverified) have more impact on the value of Bitcoin? Is there a trend between Twitter sentiment and the differential value of Bitcoin? Can the text regarding 'money' in Tweets be analyzed to predict the value of Bitcoin?
- **Combined Model:** Does adding Twitter sentiment improve the performance of the model? Does adding the 'money' improve the performance of the model?

Model Development Methods

Data Acquisition. There are two main sources of data that our team uses: Twitter and Trading View. Twitter data are acquired through the use of an advanced open-source Twitter scraping tool written in Python called Twint. Trading View data are acquired through an open-source API interface called TvDatafeed. Table 4 shows how each data source is used in the project.

Table 4. Data usage

Data Sources	Twitter	Trading View
Dataset Name	Scraped Tweets	Historical Data, Block-chain Data
Data Interval	2020 to 2021	
Unit of Analysis	Per Tweets	Per Day
Fields/Variables	id, tweet, metrics	open_Bitcoin, open_eth, open_bnb, open_ada, value_number_transaction, value_number_address, value_transaction_second, value_total_Bitcoin, value_hash_rate, close_Bitcoin

Feature Selection for Twitter Data. People use Twitter to give their opinion on Bitcoin prices. In each tweet, they might say that the price is going up or going down. The first feature of the model, sentiment data, analyzes the words and context of each tweet and computes the compound score which is the sentiment score for each tweet. Moreover, people also tweet the prices of Bitcoin in a number form for example, "Bitcoin will go up to $60000". The number data can be extracted as a second feature called Bitcoin price opinion.

- **Sentiment scores:** the scores identifying how much the tweet is negative neutral or positive.
- **Bitcoin Price Opinions:** the average money extracted from each tweet

Sentiment Data. To clean the sentiment data, first, the duplicated tweets were dropped because they might be advertisement tweets. Then, the tweets were filtered to only the English language and turned into lower cases. URLs, mentions, non-word characters, numbers, and stopwords were removed. The character sequencing of more than 3 was reduced to 3, and, 2 or more spaces were reduced to a single space.

Before Cleaning	After Cleaning
Bitcoin the worst decision i made this decade	worst decision made decade
#Bitcoin rally begun in 2013 and it reached a peak of $20,000 in 2017 https://t.co/W3ghBhpfMX	rally begun reached peak
Bitcoin faces uncertain 2022 after record year https://t.co/76aqPx4l1Q https://t.co/76aqPx4l1Q	faces uncertain record year
@Mike_Rooker I don't give investment advice here bud, but never gamble what you can't afford to lose and crypto in my opinion is highly speculative, but I do own a load of Bitcoin.	give investment advice bud never gamble afford lose opinion highly speculative load
Bitcoin up 2.58% over the last 24 hours https://t.co/iuY90oolUI	up last hours

Table 5. Tweets before and after the cleaning process

After cleaning the data, the data was ready for analysis. The SentimentIntensityAnalyzer from vaderSentiment.vaderSentiment library was used to analyze the cleaned tweet data. However, the lexicon words in the library were not enough for analyzing Bitcoin prices. Therefore, the Bitcoin-related words were added, for example, up, down, green, and red. The results from the analysis were tweets and the compound scores for each tweet.

ID	Tweet	Sentiment Score
1212239143687741440	Bitcoin the worst decision i made this decade	-0.6249
1212267316789952512	#Bitcoin rally begun in 2013 and it reached a peak of $20,000 in 2017 https://t.co/W3ghBhpfMX	0.1027
1477138318982725633	Bitcoin faces uncertain 2022 after record year https://t.co/76aqPx4l1Q https://t.co/76aqPx4l1Q	-0.2960
1213501927058792450	@Mike_Rooker I don’t give investment advice here bud, but never gamble what you can’t afford to lose and crypto in my opinion is highly speculative, but I do own a load of Bitcoin.	-0.3634
1213085771567108097	Bitcoin up 2.58% over the last 24 hours https://t.co/iuY90oolUI	0.0

Table 6. Tweets and sentiment scores

Bitcoin Price Opinions. The second feature from Twitter data was Bitcoin price opinions. According to Table 7, The price opinions were extracted from each tweet using regex.

Tweet	Bitcoin Price Opinions
Bitcoin the worst decision i made this decade	[]
#Bitcoin rally begun in 2013 and it reached a peak of $20,000 in 2017	[2013, 20000, 2017]
Bitcoin faces uncertain 2022 after record year	[2022]
@Mike_Rooker I don’t give investment advice here bud, but never gamble what you can’t afford to lose and crypto in my opinion is highly speculative, but I do own a load of Bitcoin.	[]
Bitcoin up 2.58% over the last 24 hours	[2.58, 24]

Table 7. Tweets and extracted Bitcoin price opinions

After extracting the number from each tweet. the price opinions were filtered based on the open price on that day. According to figure 1, the price opinions and the open prices on each day have a very high correlation. To avoid bias from the filtering, the price opinions were normalized before using them as a feature in the machine learning model.

Tweet	Bitcoin Price Opinions
Bitcoin the worst decision i made this decade	[]
#Bitcoin rally begun in 2013 and it reached a peak of $20,000 in 2017	[20000]
Bitcoin faces uncertain 2022 after record year	[]
@Mike_Rooker I don’t give investment advice here bud, but never gamble what you can’t afford to lose and crypto in my opinion is highly speculative, but I do own a load of Bitcoin.	[]
Bitcoin up 2.58% over the last 24 hours	[]

Table 8. Tweets and filtered Bitcoin price opinions

Feature Weighting. Features will need to be weighed in according to the importance of Tweets. Each Tweet has the following key metrics that need to be analyzed:

- Reply count
- Like count
- Retweet count

To calculate the weight of a tweet, we analyze the distribution of "Sums of Metrics", which is the sum of the metrics mentioned above for each Tweet, against the "Frequency of Tweets", which is how often the Tweet with that amount of "Sums of Metrics" appears in the data set. Figure 2 shows that the distribution is decreasing exponentially; therefore, we decided to first normalize the y-axis so that the value is from 0 to 1 only, and then find the complement of all values. The complemented value is used as the weight for the feature of that Tweet.

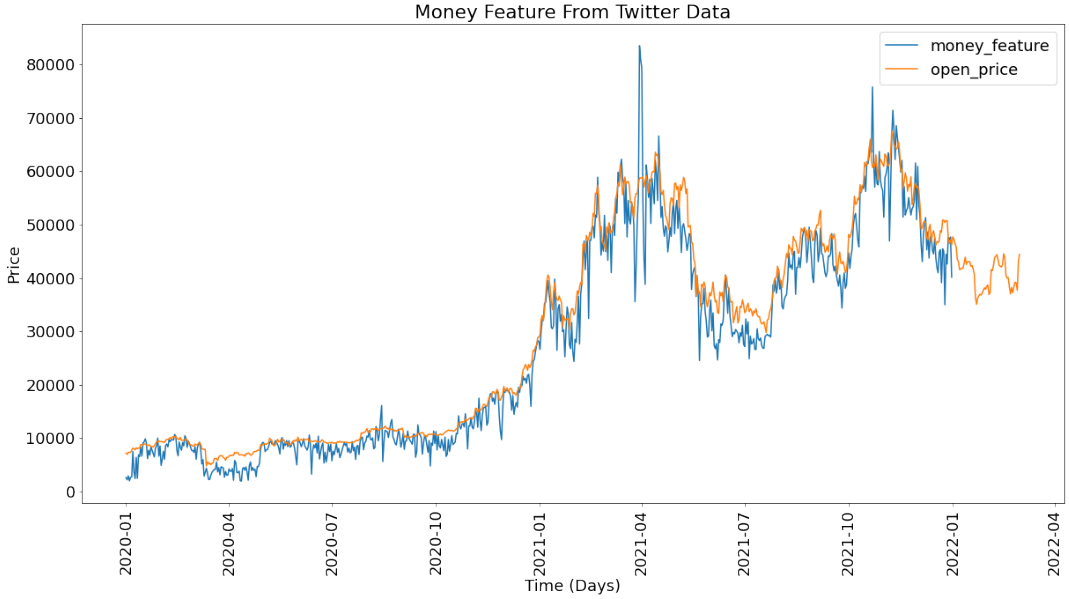


Fig. 1. Graph between average Bitcoin price opinion and Bitcoin open price on each day

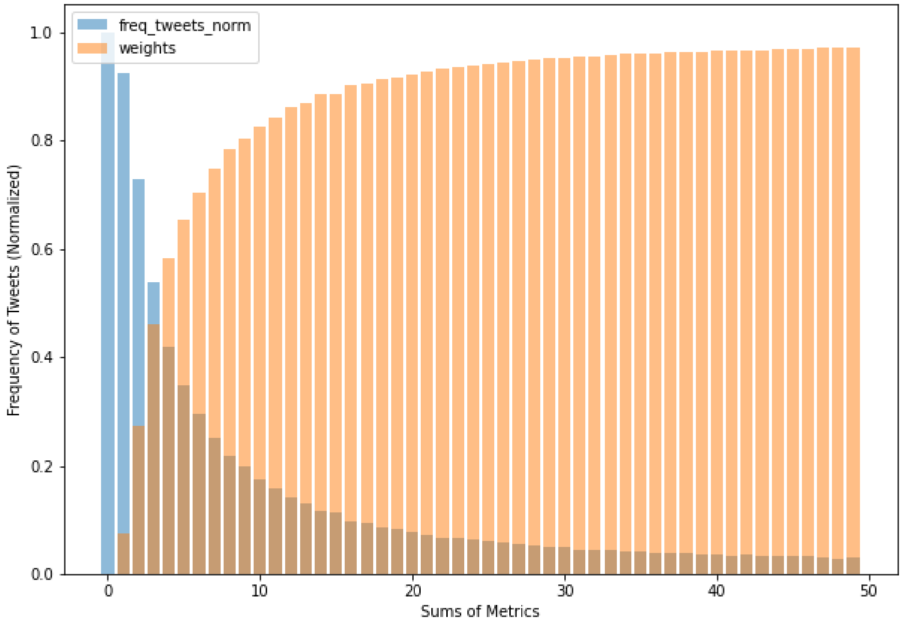


Fig. 2. Distribution of "Sum of Metrics" and "Frequency of Tweets"

Features Analysis and Visualization. With data cleaned and extracted, we are ready to compare how both features may somehow be related to the value of Bitcoin. To visualize how both features might correlate, we plot the following plots:

- Weighted Sentiment Value vs Change of Bitcoin Value
- Weighted "Money" vs Change of Bitcoin Value

The reason why we use the "Change of Bitcoin Value" instead of the value of Bitcoin is because we want to analyze the direction of the price corresponding to the feature values.

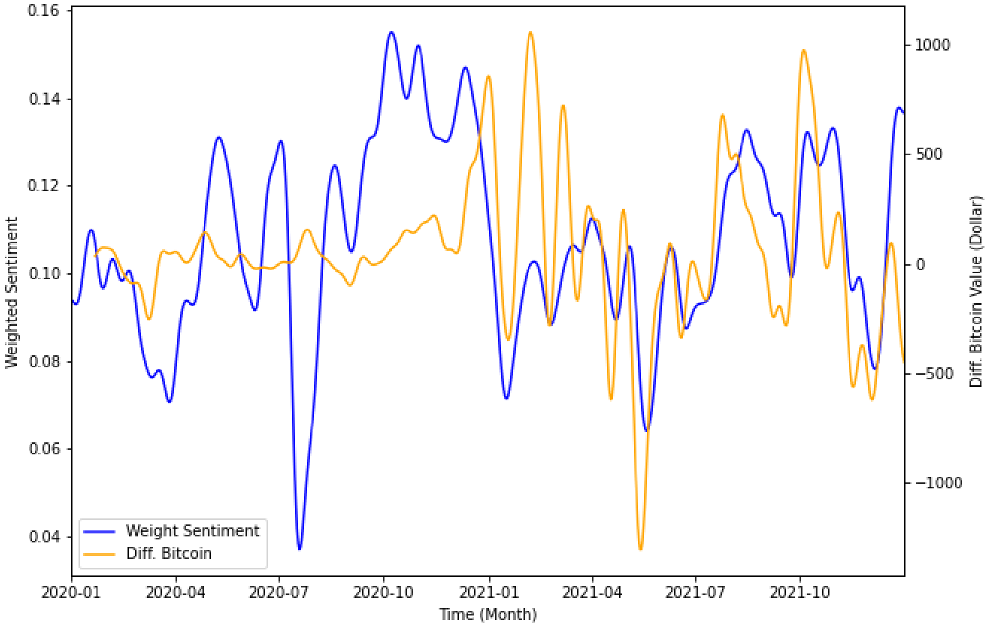


Fig. 3. Time vs Sentiment for 2020 to 2021

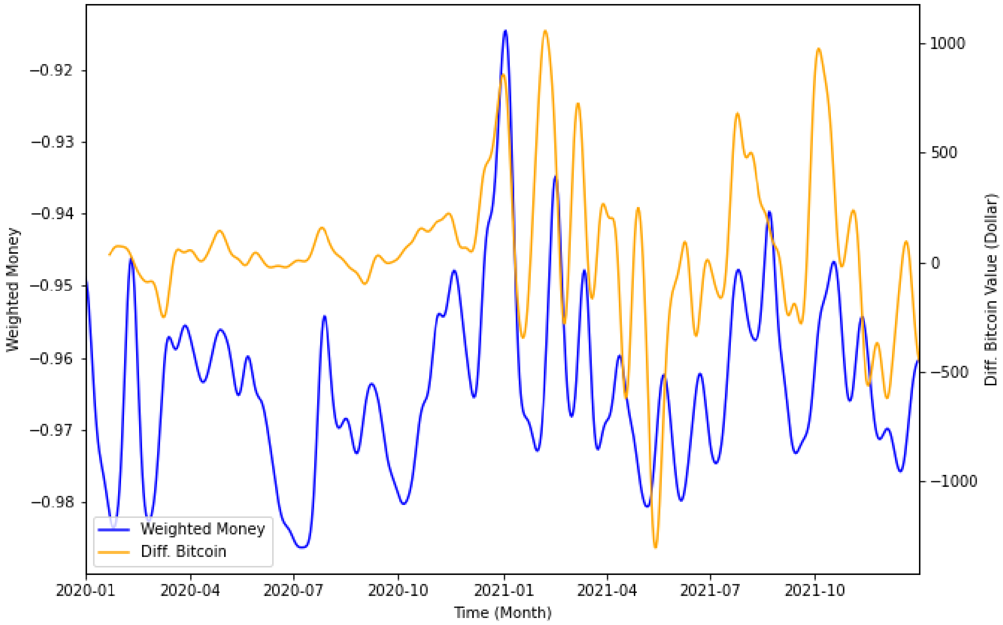


Fig. 4. Time vs Money for 2020 to 2021

Figure 3 and figure 4 show the correlations between the features and "Change of Bitcoin Value" from 2020 to 2021. In 2020, the feature values do not have much correlation to the value of Bitcoin, but in 2021, which is the year where Bitcoin saw an uproar in price and volatile behavior, is where the line seems to have an interesting pattern. To further analyze the data, we filter the plot to only 2021 to see a clearer picture of how this pattern matches.

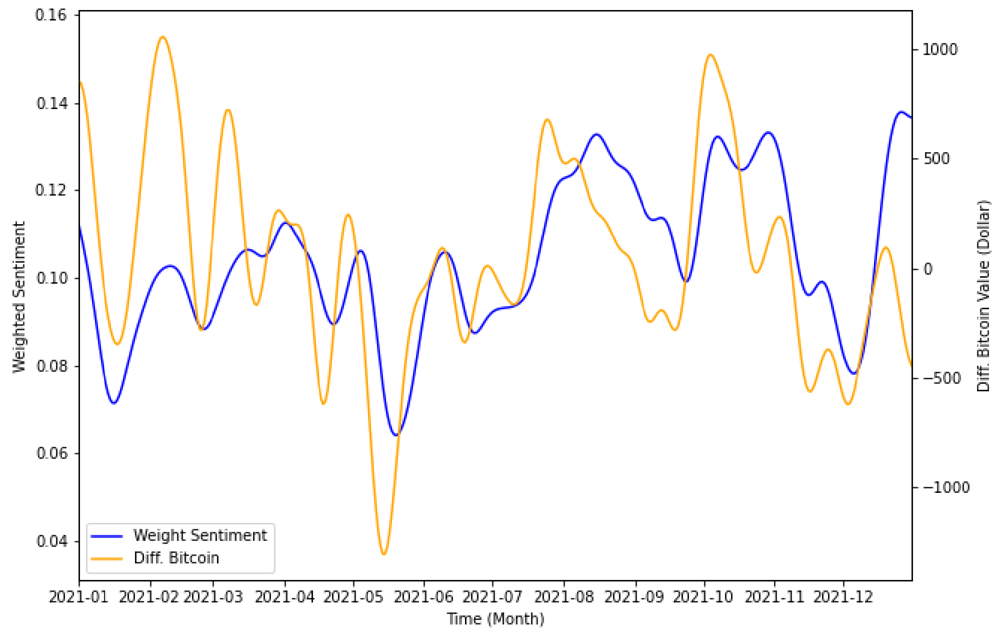


Fig. 5. Time vs Sentiment for 2021

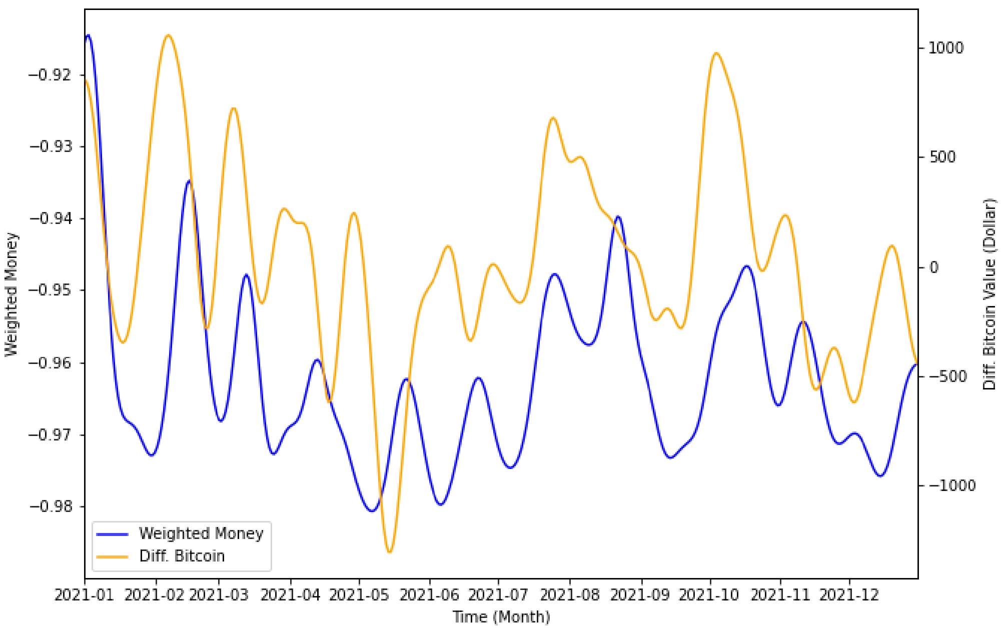


Fig. 6. Time vs Money for 2021

Figure 5 and figure 6 show the correlations between the features and "Change of Bitcoin Value" only in 2021. We can see a very interesting correlation between the two lines inside both plots. Both lines seem to follow the trend of the Bitcoin value. Although sometimes lag behind a bit, the value can theoretically be a good indication of whether the value would go up or down in many ranges. In conclusion, adding these two features to the model may result in higher accuracy in prediction.

Exploratory Data Analysis. In this section, we provide a brief exploratory data analysis for the classical linear regression that will follow. Although the purpose of the article is to understand whether Twitter sentiment such as text sentiment and price opinions are useful in predicting the future price of Bitcoin, we would also examine if there exists a statistically significant relationship between Bitcoin price and sentiment. Because machine learning is used as a prediction tool, the statistical significance of each feature is often neglected in practice. However, we also believe that it is important to discuss the statistical significance of the features and Bitcoin price movements to see if we observe a correlation between the two.

For exploratory data analysis, we merged Bitcoin price data, Twitter sentiment data as well as Twitter Bitcoin price opinion data. They were merged by exact merge using outer join based on the date-time.

As classical linear regression assumes independence of the dependent variable, we examined the autocorrelation of Bitcoin prices where autocorrelation is defined as $r_k = \frac{\sum_{t=k+1}^{n}(y_t-\bar{y})\left(y_{t-k}-\bar{y}\right)}{\sum_{t=1}^{n}(y_t-\bar{y})^2}$. Here, k represents the lag of observation while t represents the time of observation.

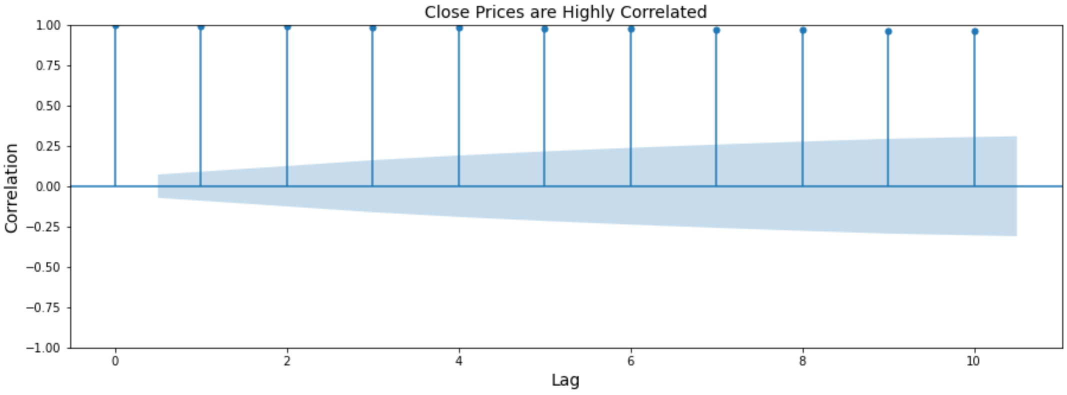


Fig. 7. Autocorrelation of Bitcoin Price

As we can observe from figure 7, the price of Bitcoin is highly correlated for different lags from one to ten. As this can be problematic in running a linear regression where the purpose is to determine the existence of statistically significant correlation, it is necessary to conduct transformation such that the values are no longer autocorrelated. To address the issue, we transformed Bitcoin price to returns where return is defined as $\%\Delta y_k = \frac{y_{t+k}-y_t}{y_t}$ where k represents the lag of return. Specifically, we calculated a one-day return for ease of analysis.

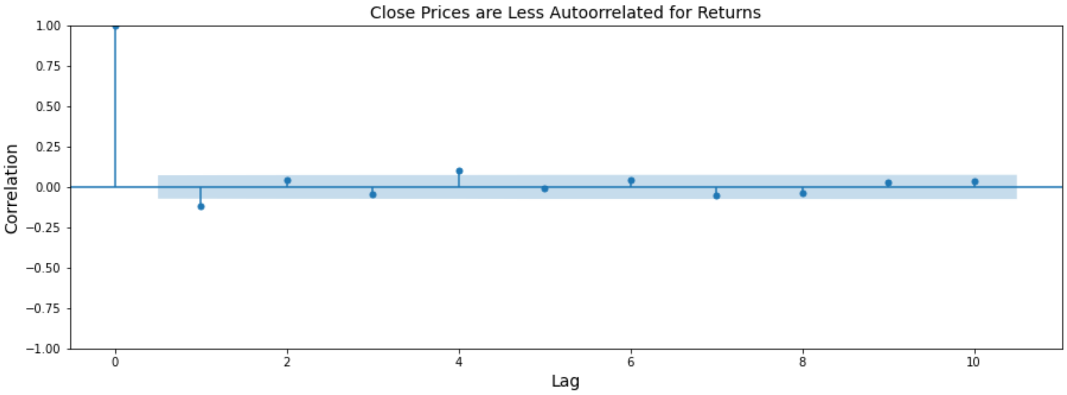


Fig. 8. Autocorrelation of Bitcoin One-Day Return

As we can see from figure 8, the autocorrelation for one-day return is much smaller and thus we use one-day return as an independent variable for the classical regression that follows.

Classical Linear Regression. In this section, we examine the correlation between Twitter sentiment, Twitter Bitcoin price opinion, and Bitcoin return.

Bitcoin return and Twitter Sentiment. It is reasonable to assume a positive correlation between Bitcoin return and Twitter sentiment as people talk more positively about Bitcoin, more people may become interested in possessing Bitcoin or increase their holdings by buying more Bitcoin in the market. This will increase the demand for Bitcoin and will lead to a higher price, thereby increasing return. Thus, we examine the hypothesis using the following regression model.

$$\%\Delta y_t = \alpha + Sentiment_{t-1} \qquad [1]$$

where $\%\Delta y_t = \frac{y_t - y_{t-1}}{y_{t-1}}$ represents the return on Bitcoin on day t. Here, y_t represents the closing price of Bitcoin on day t.

From table 9, we can observe that weighted sentiment and return of Bitcoin are indeed correlated at the 5% significance level.

Bitcoin return and Twitter Price Opinion. As for Twitter price opinion and Bitcoin return, we can assume that the higher the opinion on Bitcoin price is on Twitter, the higher the return. The hypothesis is built on the assumption that tweet's regarding the price of Bitcoin are usually the reflection of their thoughts on what the price will be. However, it may be that most of the tweets regarding Bitcoin price are about the current price and not their projections on what the prices will be in the future. In this case, tweets regarding Bitcoin price may not be informative of Bitcoin return. Nonetheless, we examine the correlation between Bitcoin return and Bitcoin price opinion using the following equation.

$$\%\Delta y_t = \alpha + price\_opinion_{t-1} \qquad [2]$$

However, from table 9, we can observe that Bitcoin price opinion and Bitcoin return are not correlated even at a 10% significance level, suggesting that they are not correlated to one another.

From the analysis, we observe that Twitter sentiment indeed seems to be correlated with Bitcoin return while price opinion does not. The correlation was indeed in line with the hypothesis that Twitter sentiment will be positively correlated with Bitcoin return. However, our result may change greatly depending on feature transformation and data sample period. Moreover, although the analysis shows a positive correlation between Bitcoin price and Twitter sentiment, it does not claim the existence of a causal relationship between the two.

Table 9. Regression Result

	Dependent variable: Bitcoin Return	
	(1)	(2)
Intercept	-0.005	0.003*
	(0.004)	(0.002)
money_opinion		-0.000
		(0.004)
weighted_sentiment	0.079**	
	(0.039)	
Observations	730	723
R^2	0.006	0.000
Adjusted R^2	0.004	-0.001
Residual Std. Error	0.041(df = 728)	0.041(df = 721)
F Statistic	4.118** (df = 1.0; 728.0)	0.010 (df = 1.0; 721.0)
Note:		*p$<$0.1; **p$<$0.05; ***p$<$0.01

Machine Learning Ridge Regression. Moving on from classical linear regression analysis, we perform statistical analysis using machine learning. The model for this analysis is ridge regression. Ridge regression is ordinary least squares regression with an L2 penalty term on the weights in the cost function. To begin, we gather two main sources of data which are historical data and block-chain data using API interface. Then, we merge the two data frames. To get rid of null values, we apply linear interpolating. After the data is well cleaned, we categorized the data into x and y data frames. Dataframe x are the features which are open_Bitcoin, open_eth, open_bnb, open_ada, value_number_transaction, value_number_address, value_transaction_second, value_total_Bitcoin, and value_hash_rate. Dataframe y is the output that we want to predict which is close_Bitcoin. Following, we split the data into 60% training, 20% validation, and 20% testing set chronologically because our data is time-series. If we randomly split the data, there would be a leakage which is training the model with future data.

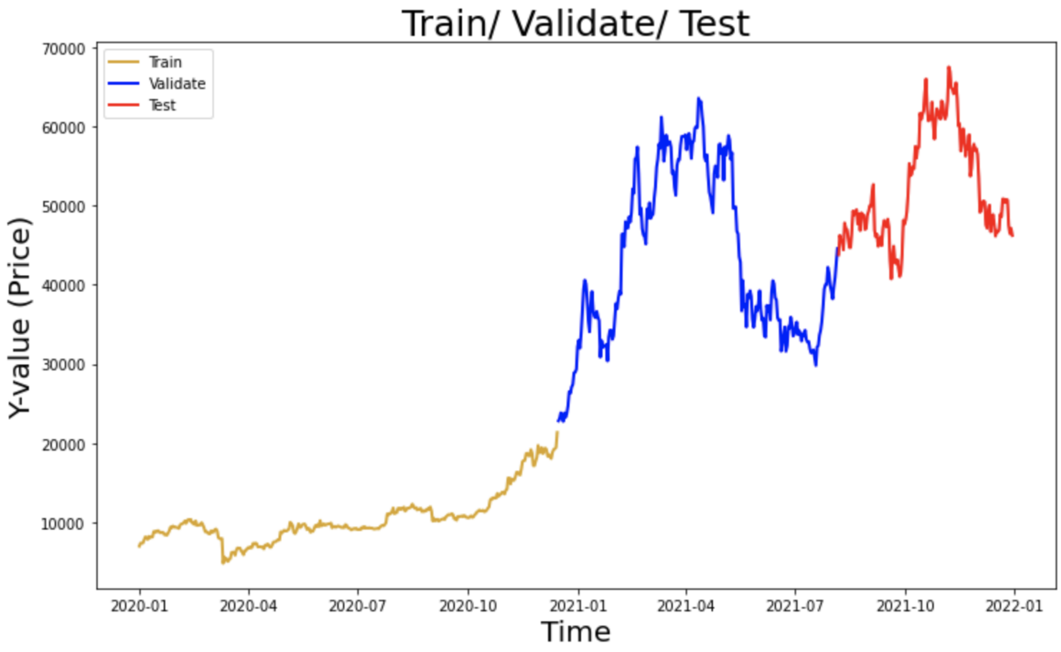
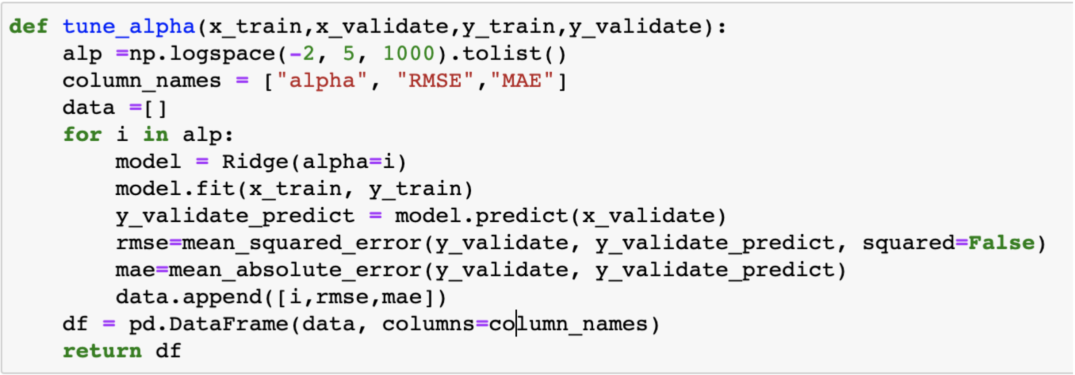


Fig. 9. Time vs Money for 2021

Figure 9 shows how the data is split chronologically. Since our data is time-series, if we randomly split the data, there would be a leakage which is training the model with future data.

Figure 10 shows our function to tune alpha in ridge regression.

After this, we standardize the x train, x validate, and x test data frames. We keep the scaler from the training set to transform validating and testing set. In our ridge regression model, there is a hyperparameter which is alpha. It indicates how much we penalize our model's parameters in the cost function. If the alpha is zero, it is just an ordinary least square cost function. To tune the alpha, we create our function. For each alpha, we create a model with our training data and use our validating set to calculate the error. Consequently, we select the alpha that gives out the lowest error on the validation set.



```
def tune_alpha(x_train,x_validate,y_train,y_validate):
    alp =np.logspace(-2, 5, 1000).tolist()
    column_names = ["alpha", "RMSE","MAE"]
    data =[]
    for i in alp:
        model = Ridge(alpha=i)
        model.fit(x_train, y_train)
        y_validate_predict = model.predict(x_validate)
        rmse=mean_squared_error(y_validate, y_validate_predict, squared=False)
        mae=mean_absolute_error(y_validate, y_validate_predict)
        data.append([i,rmse,mae])
    df = pd.DataFrame(data, columns=column_names)
    return df
```

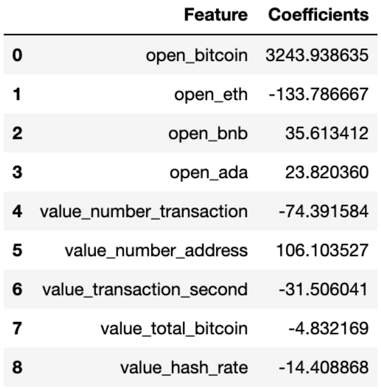
Fig. 10. Hyperparameter tuning function

Figure 10 shows our function to tune alpha in ridge regression.

Subsequently, we have a tuned alpha which is 0.5646. We create a ridge regression model using the tuned alpha to predict our testing data set. We then create a table of each feature along with its coefficient. Last but not least, we evaluate our model by calculating Root Mean Squared Error, Mean Absolute Error, and R2 score.

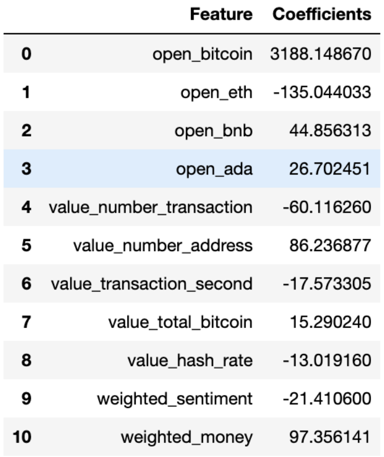
We have performed the statistical analysis using machine learning on statistical data already, however, we have not combined the result from sentiment analysis which is weighted_sentiment and weighted_money. Next, we create a model using all the data which are historical data, block-chain data, sentiment, and money. We do not go over step by step of this model because all the steps are the same as before. In this second model, the tuned alpha is 0.6424 which is a bit higher than the first model.

Results



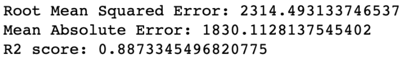
	Feature	Coefficients
0	open_bitcoin	3243.938635
1	open_eth	-133.786667
2	open_bnb	35.613412
3	open_ada	23.820360
4	value_number_transaction	-74.391584
5	value_number_address	106.103527
6	value_transaction_second	-31.506041
7	value_total_bitcoin	-4.832169
8	value_hash_rate	-14.408868

Fig. 11. Feature and its coefficient



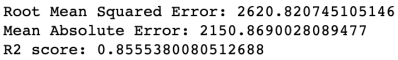
	Feature	Coefficients
0	open_bitcoin	3188.148670
1	open_eth	-135.044033
2	open_bnb	44.856313
3	open_ada	26.702451
4	value_number_transaction	-60.116260
5	value_number_address	86.236877
6	value_transaction_second	-17.573305
7	value_total_bitcoin	15.290240
8	value_hash_rate	-13.019160
9	weighted_sentiment	-21.410600
10	weighted_money	97.356141

Fig. 12. Feature and its coefficient



```
Root Mean Squared Error: 2314.493133746537
Mean Absolute Error: 1830.1128137545402
R2 score: 0.8873345496820775
```

Fig. 13. RMSE, MAE, and R2 score



```
Root Mean Squared Error: 2620.820745105146
Mean Absolute Error: 2150.8690028089477
R2 score: 0.8555380080512688
```

Fig. 14. RMSE, MAE, and R2 score

Feature and Its coefficient. Figure 11 and figure 12 shows the features of each model along with the feature's coefficient.

It can be seen that the dominating parameter for both models is the open_Bitcoin parameter. It affects our prediction the most. The more magnitude the coefficient has, the more impact it has on our model. For example, if the open value of Bitcoin is very high, the prediction from our model tends to be high as well. The least important factor in the first model is the total Bitcoin that has been mined. The least important factor in the second model is the total hash rate.

To compare both models, we can see that the weighted_sentiment and weighted_money have a magnitude of their coefficients larger than some of the block-chain data and some of the historical data.

RMSE, MAE, R2 score. Figure 13 and figure 14 shows our model's performance which are RMSE, MAE, and R2 scores.

R2 score tells us how well our model performs. The higher R2 means the predictions fit well with the true values. If the R2 is low, the prediction and the actual value would be far away from each other. According to our R2 score, our model performance is significantly high.

For the RMSE and MAE, they are used to measure the error of our data. If we had not added the L2 penalty term and not tuned its alpha, the differences between training error and testing error would be much higher than this due to overfitting.

To compare both models, it can be seen that the second model performs a bit worse than the first model. The latter has more error and less R2 score than the former.

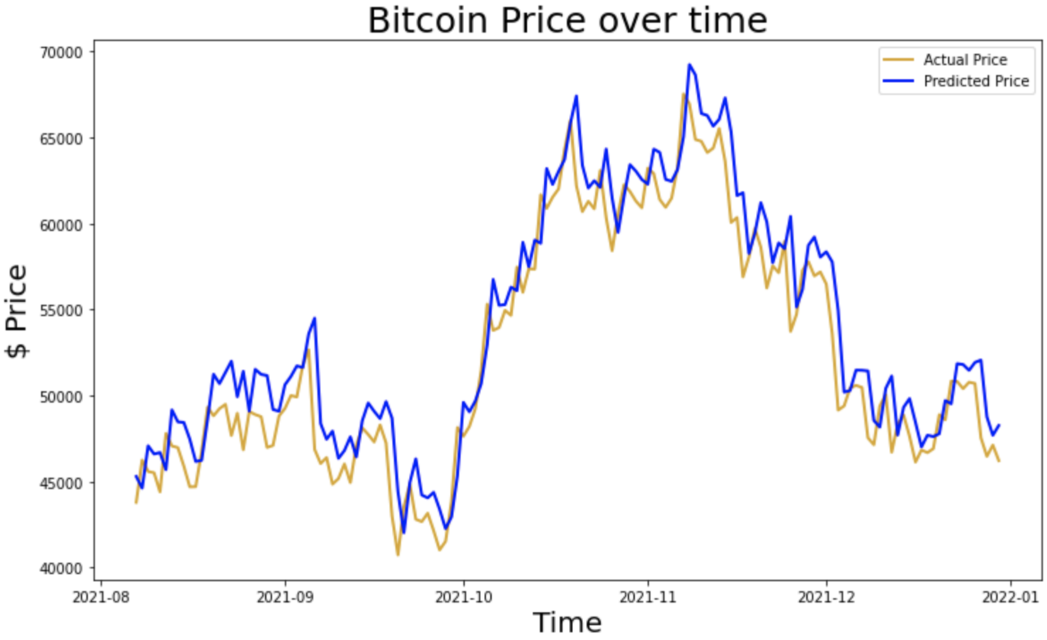


Fig. 15. Bitcoin Price over Time without sentiment analysis on the test data set

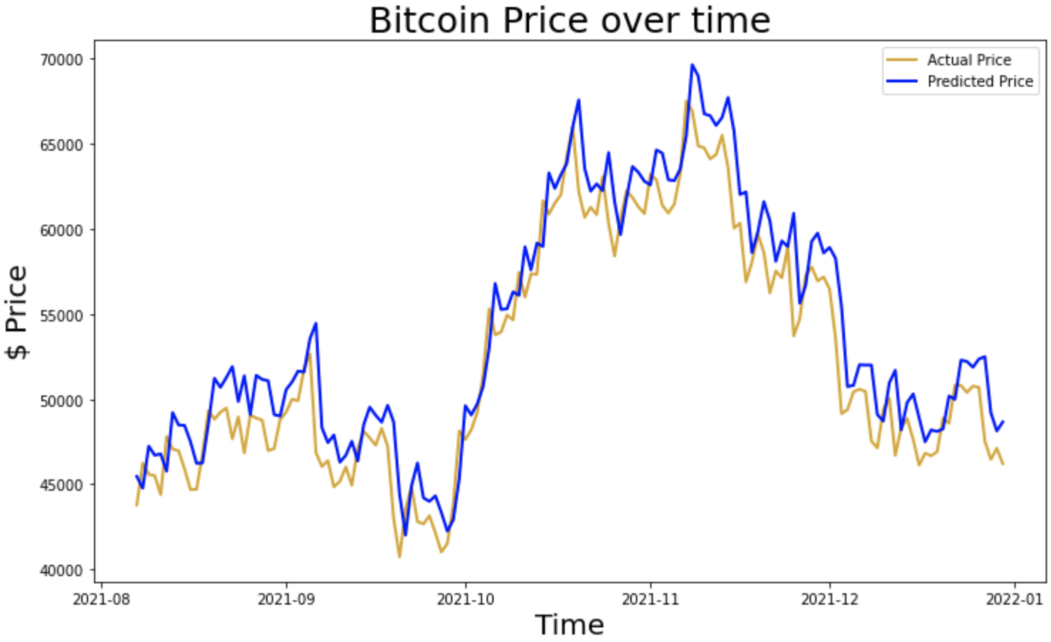


Fig. 16. Bitcoin Price over Time with sentiment analysis on test data set

Bitcoin Price Over Time. Figure 15 and figure 16 shows the prediction of daily close Bitcoin price over a two-year period from our two models and the actual close price.

We can see that our prediction daily close price (blue line) is mostly above the actual daily close price (gold line). We can interpret from the result that our models over predict the daily close price of Bitcoin.

Improvement. We believe that our models can be improved if we use return which is the percentage of changes in the daily close price instead of just the daily close price.

Conclusion. In conclusion, we have developed a machine learning algorithm using various features related to Bitcoin and other cryptocurrencies and supplemented them with non-traditional features from Twitter. Although the model with Twitter sentiment and Twitter price opinion features performed slightly worse than the model without them, classical linear regressions result suggests that Twitter sentiment and Bitcoin return is correlated. Therefore, it is necessary to conduct a more in-depth analysis regarding the relationship between them to understand whether Twitter sentiments including price opinions are indeed predictive of future Bitcoin price.

ACKNOWLEDGMENTS. We thank Professor Rebecca Johnson for assistance and recommendations throughout the process of development.

1. C McFall, Seven charts that explain the current state of crypto (2021).
2. B Learn, How to analyze a cryptocurrency using fundamental analysis (2021).
3. V Ugochukwu, Fundamental vs technical analysis in crypto (2021).
4. E Genç, Technical analysis explained: Elementary concepts in trading cryptocurrency (2021).