Exploring Patterns and Correlations Between Cryptocurrencies and Forecasting Crypto Prices Using Influential Tweets

Mohit Kumar, Gurram Sahithi Priya, Praneeth Gadipudi, Ishita Agarwal, and Saleti Sumalatha(✉)

SRM University, Amaravathi, India
sumalatha.s@srmap.edu.in

Abstract. The Crypto market, as we know, is a market full of various kinds of investors and influencers. We all know the pizza incident in 2010 where a guy purchased two pizzas at 10000 BTC, which ranges nearly around 80 million in current times. That describes how much the market has progressed in these 10–12 years. You can see drastic changes in the price of several coins in the past few years, which brings in many new investors to invest their money in this market. Crypto Market has highly volatile currencies. Bitcoin was around 5K INR in 2013, and by year 2021, it reached 48 Lakhs INR, which shows how volatile the market is. The dataset provides many fascinating and valuable insights that help us gather practical knowledge. As data scientists, we are very keen to understand such a market whose data is unstable and keeps changing frequently and making out new patterns with time. This introduction of new patterns with time makes this problem an interesting one and keeps on motivating us to find some valuable information. So, through this manuscript, we tried to analyze two specific crypto coins for a particular period, including more than 2900 records. We found several interesting patterns in the dataset and explored the historical return using several statistical models. We plotted the opening and closing prices of the particular coin by using NumPy, SciPy, and Matplotlib. We also tried to make predictions of the cost of the specific currency and then plot the predicted price line with the actual price line and understand the difference in the prediction model with the fundamental price mode. To do so, we used the Simple Exponential Smoothing (SES) model and performed sentiment analysis based on influencing tweets on Twitter. That makes our prediction more accurate and more reliable than existing techniques. Lastly, we used a linear regression model to establish the relationship between the returns of Ripple and Bitcoin.

Keywords: Crypto market · Cryptocurrency · Data mining · Data visualization · Simple exponential smoothing · Sentiment analysis · Linear regression

1 Introduction

A cryptocurrency is an encrypted string of data representing a unit of currency. It is overseen and hosted by a peer-to-peer network called the blockchain [1]. The cryptocurrency

N. Khare et al. (Eds.): MIND 2022, CCIS 1763, pp. 308–321, 2022.
https://doi.org/10.1007/978-3-031-24367-7_30

market is very similar to the stock market, where we can buy, sell and transfer digital coins instead of physical coins; these digital coins are considered digital currency [12]. Crypto is a decentralized currency which means the government of any particular country does not issue it. Cryptocurrencies are created using cryptographic algorithms that are maintained and validated through a process called mining. Cryptocurrencies have many coins, some of which are Cryptocurrencies that operate on a distributed public ledger known as the blockchain, a record of all transactions maintained and held by the owner of the currency. There are thousands of cryptocurrencies in the market; bitcoin, Ethereum, Litecoin, and Ripple are some of the famous cryptocurrencies.

Recently, we have seen the crypto market's growth peak, making this market a matter of curiosity to analyze and get to know the variations that occur there every minute. Here we have discussed data preprocessing, data reduction, finding patterns, historical results, and comparing crypto coins, namely bitcoin and ripple. We used various techniques like interpolation for cleaning data, Person correlation for data reduction, Single Exponential Smoothing for plotting predicted and actual values, and linear regression for performing a comparison of two coins. These techniques are used to analyze market data and the expected results, which we further compare with the actual data.

Bitcoin is a decentralized virtual foreign money created in January 2009 [13]. Bitcoin promises to decrease transaction costs than conventional online charge mechanisms do, and not like government-issued currencies, and its miles are operated with the aid of using a decentralized authority. Bitcoin may be very famous and has precipitated the release of masses of different cryptocurrencies, known as altcoins. Bitcoin is generally abbreviated as BTC while traded. Bitcoin makes use of the peer-to-peer generation for doing transactions.

Ripple is a financial system that functions as both a cryptocurrency and a digital payment network. Ripple's basic procedure is a payment settlement asset exchange and remittance system, comparable to the SWIFT system for international money and security transfers, which banks and financial intermediaries utilize. Ripple is a peer-to-peer decentralized open-source network that enables the frictionless movement of money in any form. It is a worldwide payments network with a customer base. XRP is employed in the company's products which allow rapid currency conversion.

We reviewed the already existing algorithms behind the proof of work which is a critical factor in mining Bitcoin/XRP, and we gathered the data and found the statistics of Bitcoin/XRP for the past five years. Then, we went to find interesting patterns in the dataset. After finding the patterns, we went to visualize closing prices and predicted vs. actual values of the coins. Later, we went on to forecast the price of crypto coins by using simple exponential smoothing and sentimental analysis model. Finally, we used an ML algorithm that is linear regression to find out the relationship between two coins, namely Bitcoin and XRP. Figure 1 describes the various tasks which we performed on the crypto currency dataset.

This document's contents are categorized as follows: In Sect. 2, we did the literature review on Bitcoin, XRP cryptocurrencies, and its mining process. It also consists of already existing algorithms and mining techniques at present. Section 3 contains a dataset description that is a detailed description of the dataset, which consists of 2 database tables. Section 4 consists of the implementation part. Section 5 is experimental result

and Analysis, which consists of graphs of data collected over several simulation results and analyses. In Sect. 6, we conclude the work with the collected results and discuss the shortcomings and future work.

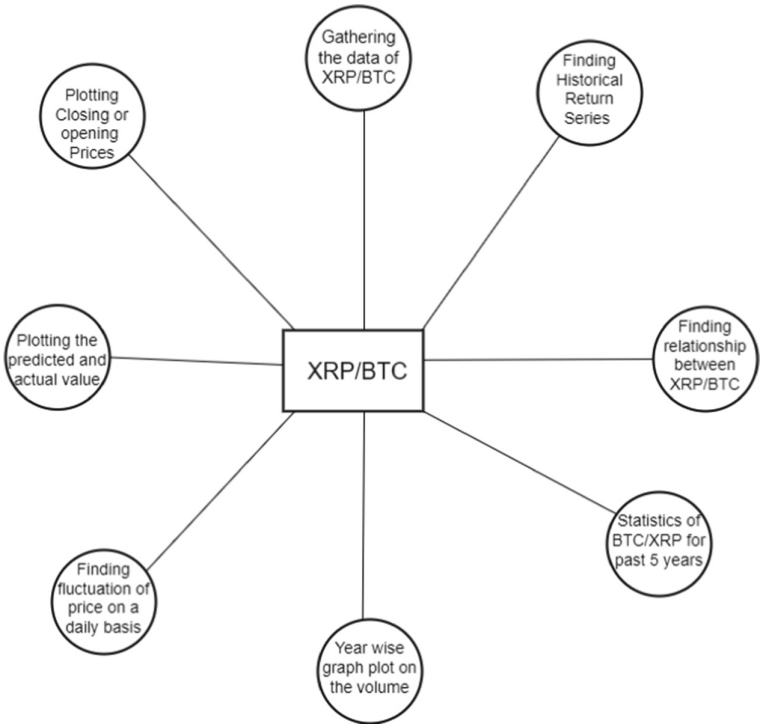

Fig. 1. Various analysis on cryptocurrencies.

2 Related Work

Blockchain technology [1] arose from the proposal of Distributed Ledger Technology (DLT) in the 1990s. Although DLT was introduced almost two decades ago, it gained traction with the deployment of Bitcoin [15]. In 2008, it was used as a crypto-currency for the first time. The data structure for the first was a hash-table was used to create a blockchain. However, as demand has grown, various other data formats have been proposed to address the constraints of the classic blockchain. Some writers advocated for the use of Directed Acyclic Graphs (DAGs) [15].

As it is scalable, lightweight, and decentralized, the Graph (DAG) keeps track of transactions. However, this option has some of the same issues as the first. At substantial sizes, the blockchain is used. RadixDLT's Tempo Ledger is designed to scale linearly in a limitless and efficient manner. In contrast to standard blockchain implementations,

where the entire ledger is stored at each node, the Tempo ledger allows each node to keep a shard of the entire ledger [15].

Bitcoin is described as a cryptocurrency that was established in 2009 as the first decentralised digital money. Bitcoin enables online payments to be made without the use of banks, as well as the purchase of products and services from one party to another without the need of a financial institution [12]. Bitcoin has just lately been a topic of economic study. In computer science, the issue has piqued interest for a long time. Computer scientists have authored a modest number of theoretical articles on incentives. Mining is not incentive-compatible, and that "selfish mining" can result in better revenue for miners that conspire against one another [4]. The anonymous online marketplace in cryptocurrency, is another example of computer science research [2]. Although there has been some work on Bitcoin published in legal journals, there is relatively little in the economics literature.

Blockchain technology is a relatively new emerging technology that has the potential to transform a variety of established businesses. Since the launch of Bitcoin, which is blockchain 1.0, blockchain technology has gotten a lot of attention, and a lot of user transaction data has been collected [1]. As a result, determining the degree of value, performance, and cost of a blockchain-based application requires a thorough understanding of what and how data is stored and altered. While blockchains improve data quality by offering a transparent, irreversible, and consistent data repository, they also present new issues in terms of data management. That's why we use data and web mining techniques and algorithms to resolve emerging issues [7].

A lot of research has been conducted on the correlation between Bitcoin and other financial assets. In one such study, they discovered data supporting the long-run correlation between Bitcoin and major stock indexes using the ARDL boundary test method [3], and they also found out that there is a relation between Bitcoin prices and the leading US market and China market, which can have a major impact in their long term investment decision process. Another study used the Monte Carlo simulation to assess the structure of Bitcoin [13] dynamically. One study shows that Bitcoin does not adhere to the one-price rule [9]. There is research that is based on GARCH models to study and determine the bitcoin volatility, and their study says Bitcoin is a very speculative market [6]. Applied rolling window approach to study the time-varying long-term memory in the Bitcoin market [5]. We can see that there is a lot of scope for more research in this field. So, in this paper, we tried first to explore and bring out the hidden patterns from the crypto dataset. We used two well-known technologies: simple exponential smoothing and sentimental analysis, to outperform the forecasting for cryptocurrencies. Using these both improves the accuracy and takes into account the current news and trends of crypto going across the globe.

3 Dataset Description

The dataset contains two distinct datasets each of them describing each coin (XRP, Bitcoin). The dataset has 2894, 2992 rows of XRP, Bitcoins with 10 features (or columns). We obtained this dataset from the Kaggle website, which was obtained from "coinmarketcap," an open-source, free-to-use data site. Since April 28, 2013, the cryptocurrency

price data has been gathered daily. This dataset has the historical price information of two top cryptocurrencies (XRP, and Bitcoin) by market capitalization. The features present in this dataset are the date of observation (Date), the opening price of the given day (Open), the highest price on the given day (High), the lowest price on the day (Low), the closing price of the day (Close), the volume of transactions on the given day (Volume), market capitalization in USD (Market Cap). There are a few redundant columns in this time series collection, and there is no room for missing values [11].

If we perform the statistical analysis on the bitcoin we can clearly see that the minimum value of the adjacent closing price of bitcoin in the span of 9 years is 68 dollars and the maximum value at the same time period is 63503 dollars which clearly indicates the high volatility of the bitcoin in the crypto market. Even at the seventy-fifty quartile, the value is 8576 which shows the sudden jump in the price of bitcoin within a short period of time, this creates the uneven distribution of data. Off all the 11 columns of data 7 of them are having float with 64 bit as its datatypes, 1 int (64 bit) datatype, and the rest with object datatype.

Whereas in the case of XRP or ripple the minimum value of the adjacent closing price in the span of 9 years is 0.0028 dollars and the maximum value at the same time period is 3.37 dollars. Compared to bitcoin the XRP is not too volatile in nature and it is easily predictable. Similar to that of bitcoin the off 11 columns, 7 have a float,1 int, and the rest object datatypes.

4 Preprocessing of Dataset

Data preprocessing converts data into a format that can be processed more quickly and efficiently in data mining. Before processing the data using many data mining techniques to find out the different patterns in the data, the initial step we are performing is data preprocessing.

Data preprocessing involves many steps:-

4.1 Importing the Required Libraries and Dataset

First, we imported the required libraries such as Numpy which contains mathematical functions and scipy which contains modules for interpolation, linear algebra etc. The second step is importing the dataset. The most common format for data sets is.csv. A CSV file is a plain text file that contains tabular data. To read a local CSV file as a data frame, we use the panda's library's read CSV method.

4.2 Data Cleaning

Now we perform data cleaning, in which our main aim is to remove the inconsistent data, fill in the missing values and ensure that the data is suitable for the analysis [10]. Suppose we take bitcoin, we have many values like high, low, open, close etc. If we have a missing or null value at some column, and if we are trying to remove it, there is a high possibility that it may alter our results as it is a time-series data, every data record is important for the analysis, as a result, we are using the interpolation to fill in the missing

or null values. So, in the data cleaning step, we are finding all the missing values and filling them by interpolation which is taking the average value to the above and below valid value (trying to fit in between the above and the below valid value).

4.3 Data Reduction

The dataset which we are working on is of 5 years data, so it would have some redundant data and also the attributes which are not required for our analysis. As the redundant data would cause us trouble in our analysis to give accurate results, we are dropping a few attributes which are redundant or have similarities with other attributes. In order to do so, we will use Pearson Correlation Coefficient between two columns. In our dataset, when we performed the Pearson correlation coefficient between "close" and "adjacent close", the result turned out to be 1 which means they are highly correlated. So, we dropped the adjacent close attribute from both the XRP and BTC.

5 Proposed Scheme and Implementation

5.1 Finding Patterns and Historical Return Series in the Dataset

Finding patterns is quite essential in understanding the coins and making up the investing plans. The historical return series provides us with the historical analysis of both gains and losses that occurred with both the coins XRP and BTC. We first preprocessed the dataset and computed the percentage change on the adjacent closing values to compute the historical return series. To better understand the percent change, just look at the Formula 1 mentioned below.

$$\text{pct\_change} = \text{close}_{\text{Today}} - \text{close}_{\text{PreviousDay}} \tag{1}$$

After computing the percentage change in the adjacent closing values, we plotted them graphically to understand better the trend of the returns. Along with this, we also explored the closing prices trend by plotting them using a line plot. Figure 2 shows the closing price curve for BTC, while Fig. 3 shows the closing price curve for XRP. While closing observing them, we could easily see how BTC started from somewhere around 70 USD and reached 60000 USD. Similarly, one can see XRP, which started from somewhere about 0.08 USD and reached 3.5 USD. It clearly shows that the crypto market is very volatile, and its users can find a lot of fluctuation as time passing.

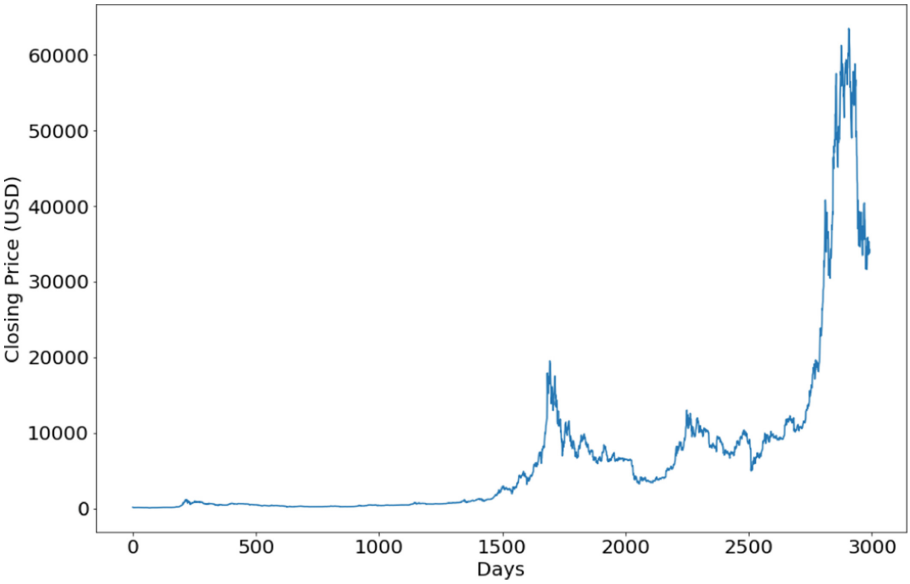

Fig. 2. Closing price curve for BTC.

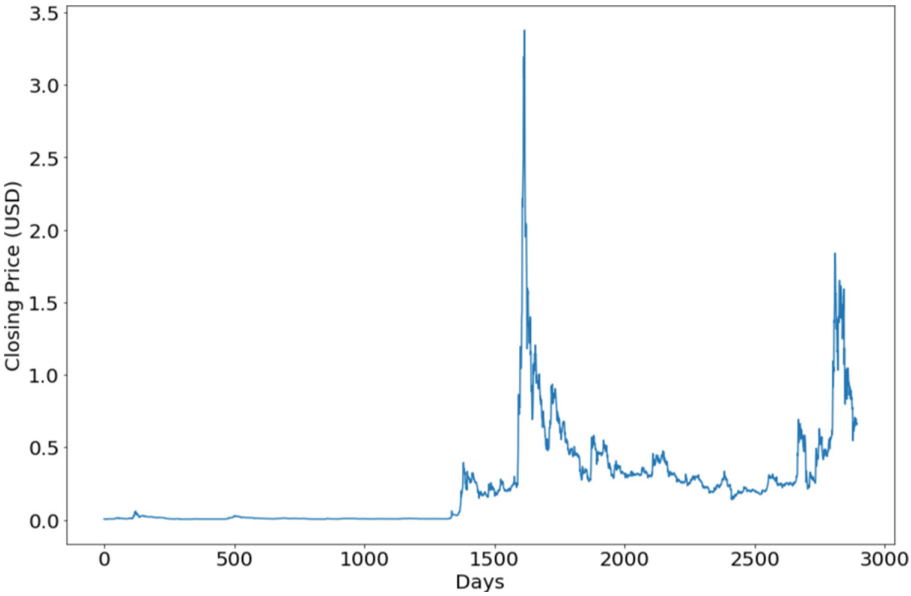

Fig. 3. Closing price curve for XRP.

5.2 Finding the Actual and Predicted Price Line Using Simple Exponential Smoothing and Influential Tweets

Simple exponential smoothing (SES) is used to predict time series when the data does explicitly not follow a trend in either an upward or downward trend, and seasonality (patterns caused by seasonal factors such as hours, days, and years. SES uses weighted averages to determine its rankings [8]. The average of the previous value and the current observation. Enormous weights are associated with recent observations, and the smallest weights are associated with older observations. The weight decrease is controlled by the smoothing parameter known as alpha or smoothing coefficient, which can be between 0 and 1. The forecast for future value is based on the average of past data. One means that the prediction of all future values is the final value.

Simple Exponential Smoothing (SES) is defined under the stats model library of python. SimpleExpSmoothing.fit (smoothing_level = None, optimized = True). The smoothing_level parameter is set, then the value is used for simple exponential smoothing. An optimized parameter of the boolean type is set to ask to go for optimizing the parameters as mentioned earlier automatically. We will be using.fit() to fit the model. Now, along with it we added sentiment analysis for each cryptocurrency to estimate the likelihood of a rise or fall in the price.

Bitcoin is a leading cryptocurrency in the crypto market. Its prices and returns are growing with the growth in the crypto market worldwide. On the other hand, Twitter has increasing recognition and predictive power for various programs related to the crypto world and the financial market [14, 16, 17]. So here we are examining how accurately public opinions on Twitter can be used to forecast Bitcoin returns. Using a sentiment analyzer on Bitcoin-related tweets and financial data, we observed that sentiment analysis on Twitter has predictive value for Bitcoin's results.

To begin with, we downloaded the historical prices of Bitcoin and the collection of Bitcoin-related Tweets, then analyzed the sentiment of the tweets. These tweets were gathered using several APIs and a little web scraping. In our dataset, we are examining the 92550 tweets which were posted virtually every minute, according to the statistics. Our objective is to use sentiment analysis to determine the subjective feelings or views regarding Bitcoin expressed in our collected tweets. We used the VADER (Valence Aware Dictionary and Sentiment Reasoner) for our processing.

Next, for evaluating our proposed model, we choose Random Forest regression. It is a type of machine learning algorithm which is effective when working with different kinds of inputs that are not related to each other at all. As inputs, we used the Sentimental Analysis score and history price of bitcoin and then implemented random forest analysis. When making predictions based on bitcoin-related tweet sentiment and historical bitcoin price, about 62.48% accuracy was observed. Price forecasting studies sometimes use sentiment analysis of tweets. Because of the enormous number of news updates per minute regarding Bitcoin, most academics use Twitter to analyze Bitcoin's sentiment [14, 17]. Similarly, we did the same analysis for XRP and tried to investigate the sentiment from XRP-related tweets.

5.3 Finding Relationship Between BTC and XRP Returns Using Linear Regression

We need to find the similarity between the two coins, BTC and XRP. To do so, we rely on a very well-known model, which is linear regression. Linear regression is one of the most fundamental regression models for predicting outcomes. It's used to build a relationship model between the independent and dependent variables. There is only one independent variable and one dependent variable in simple linear regression. One way to model the relationship between two variables is linear regression. Gradient equations are another name for equations. The equation is:

$$Y = a + bX$$

where Y is the dependent variable (that is, the variable plotted on the Y-axis), X is the independent variable (that is, the variable plotted on the X-axis), and b is the line. Gradient and y-intercept. So, keeping the above concept in mind, we applied Linear regression to analyze the relationship between BTC and XRP returns. Bitcoin is based on blockchain technology, while Ripple does not use blockchain but uses a distributed consensus ledger using a network of validation servers and a cryptocurrency token known as XRP. We kept BTC percentage returns as the independent variable while the XRP percentage returns as the dependent variable. Our ultimate goal was to analyze how the percentage return of BTC correlated to the percentage returns of XRP. As a first step, we computed the Pearson coefficient among the two variables and then calculated the slope and intercept for the best fit line. After getting the predicted y, we plotted the best fit line with the actual points. Doing so helped us to analyze the relationship between the two coins better.

6 Experimental Results and Analysis

We applied several statistical techniques and models to understand the crypto market trends better. While doing so, we examined several sub-sections like historical return series, forecasting closing prices, finding relationships between the two coins, etc. Now it's the time for us to explore all the results and draw conclusions from them. So, starting with the historical return series, we found that both XRP and BTC had comparatively similar kinds of return series except for a few outliers. Figure 4 and Fig. 5 show BTC and XRP historical return series, respectively.

We performed simple exponential smoothing on the time series crypto dataset to forecast the closing values for both BTC and XRP. Figure 6 and Fig. 7 show the plotting of both predicted and actual values of BTC and XRP, respectively. The red lines in the curves indicate the actual values, while the blue lines indicate the predicted value. We can observe that the predicted prices have coincided with actual prices for the first 1300 data rows. This means the model has performed very well at first, but the predicted values have varied. Though the model performed very well for low extremes, the model has failed to predict the prices when there is a sudden spike in the prices and at times of high volatility. One can easily conclude that the simple exponential smoothing model fits well to predict the closing values. However, there are a few moments where we see deviations in the

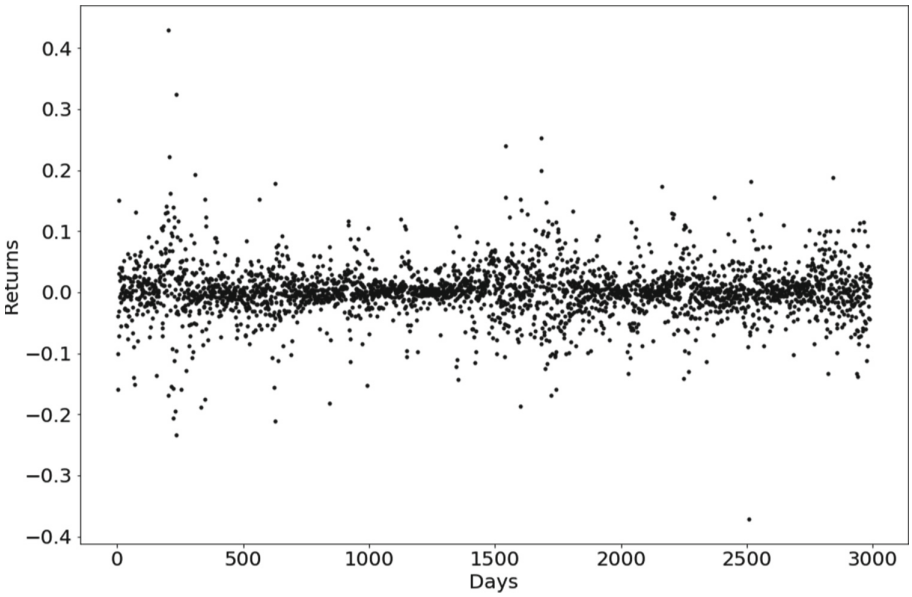

Fig. 4. Historical return series of BTC.

predicted and actual values, which can be further smoothed by considering some more parameters like current business news, war status, market volatility, etc. So, thinking the same we tried to examine influential tweets which can further improve the predictions.

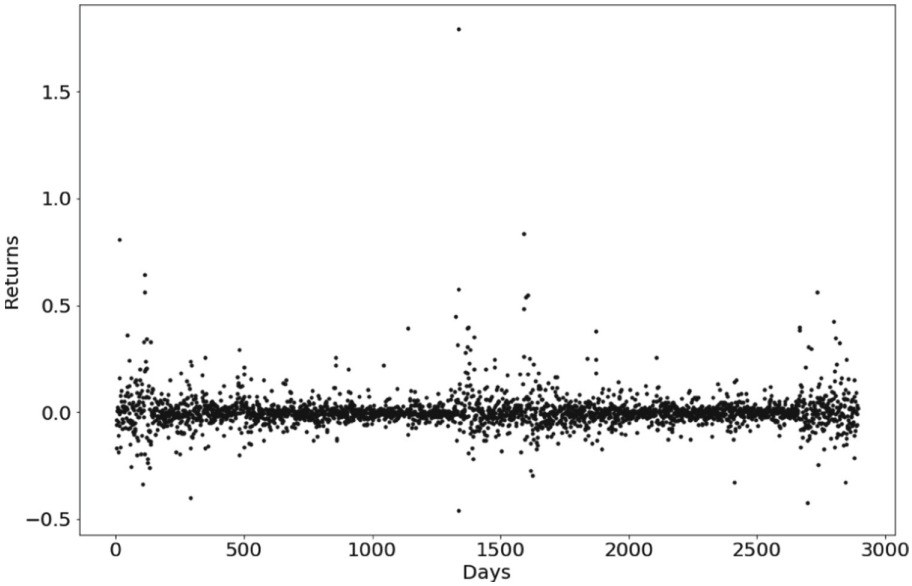

Fig. 5. Historical return series of XRP.

With simple exponential smoothing, we can easily get an exponential equation for the respective cryptocurrency, whether BTC or XRP. This equation is then used to make future predictions. We proposed a scheme that couples simple exponential smoothing with sentiment obtained from influential tweets to make the forecast more accurate. In other words, we use tweets that are related to crypto or tweets which are highly significant. To find out whether a tweet is influencing the market or not, we basically used hashtags and crypto-related keywords. We analyze the sentiment from those influential tweets and try to determine whether the tweet is affecting the market positively, negatively, or has no effect on the cryptocurrency. Figure 8 shows the obtained graph after performing sentiment analysis on influential tweets.

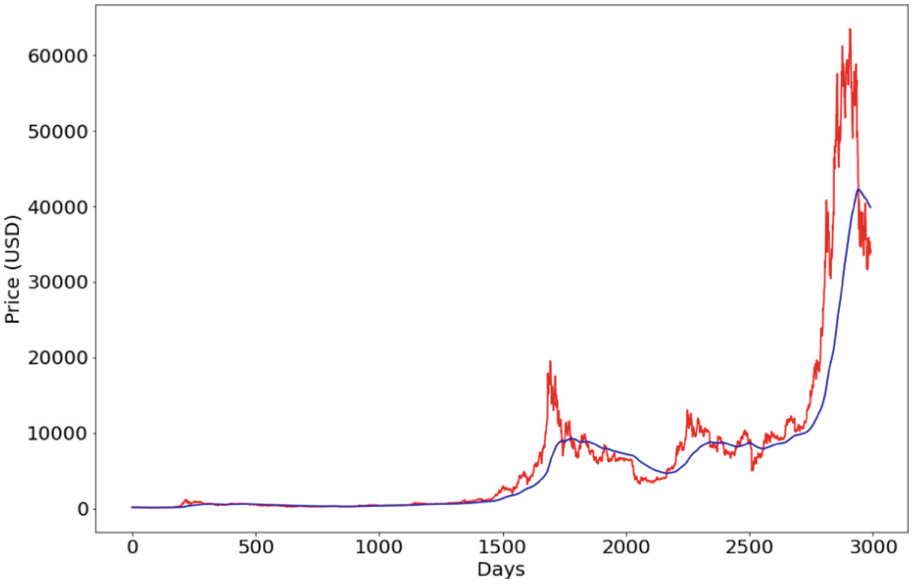

Fig. 6. Predicted vs actual values of BTC. (Color figure online)

Finally, we applied a linear regression model to analyze the relationship between the returns of BTC and XRP. We first computed the percentage change as usual and then used the seaborn regression plot module to plot the return series. Figure 9 shows the relation between the returns of BTC and XRP collectively. The X-axis consists of the BTC percentage return, while the Y-axis consists of the XRP percentage return. While closely observing the plot, we found that both BTC and XRP are highly correlated in returns and have the same percentage returns. However, if one aims to have higher profit margins, BTC shows significant fluctuations, which can be used as an advantage while trading BTC.

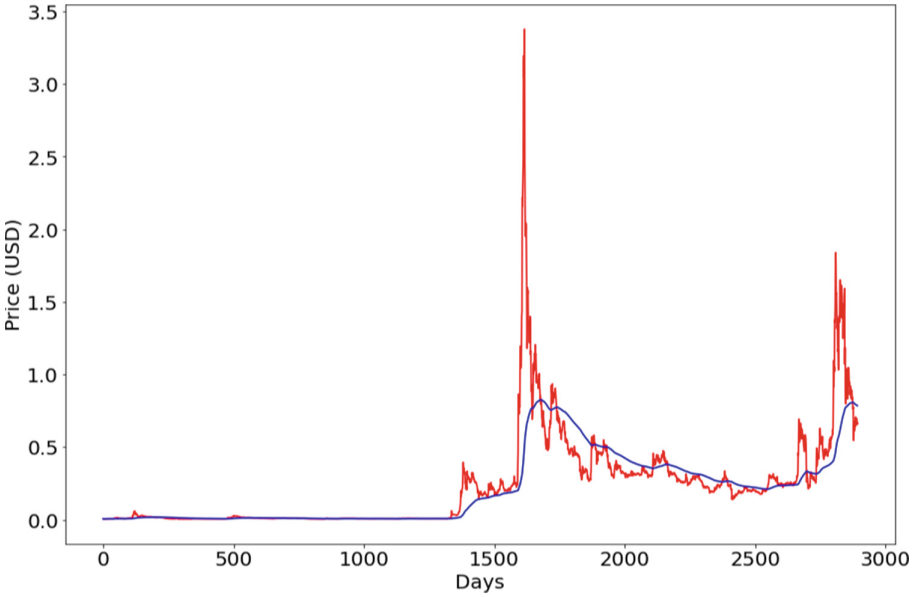

Fig. 7. Predicted vs actual values of XRP. (Color figure online)

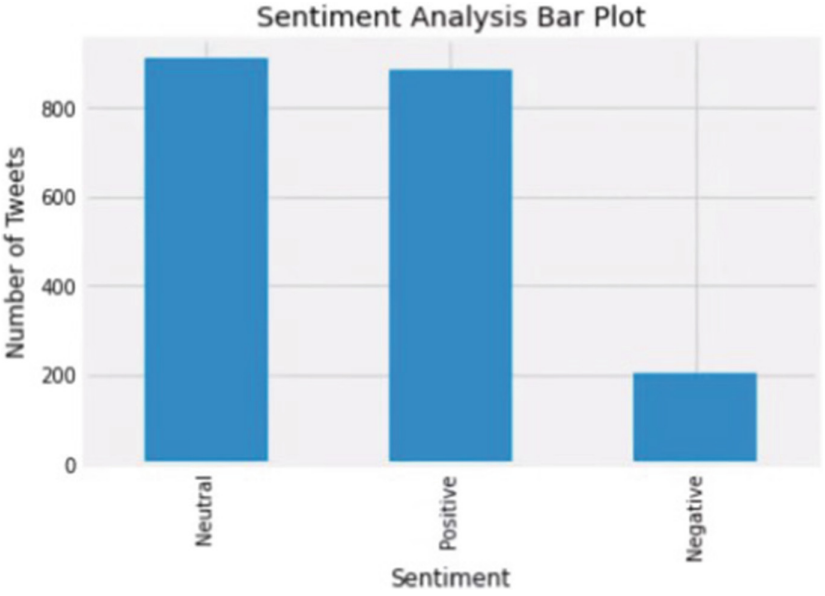

Fig. 8. Results of sentiment analysis over influential tweets.

7 Conclusion

The paper's experimental observations and results show that the crypto market is quite volatile and is susceptible to frequent fluctuations. The analysis performed on the BTC and XRP has shown a positive correlation and were following a similar trend in highly volatile times. The model built using VADER for processing and Random Forest for prediction has performed sentiment analysis on public opinions by tweets about how the subjective feelings were determining the prices of bitcoin has shown 62.48% of accuracy. Forecasting the prices helped us conclude that in the coming years, crypto market trading and the number of investors will rise. The simple exponential smoothing used to predict value has shown great results in low extremes, but it has some vulnerabilities when there is a sudden spike which we going to be tackled by adding various external factors. In the future, we would like to build a prediction model considering several factors and using current data and historical data to make better predictions of coin prices.

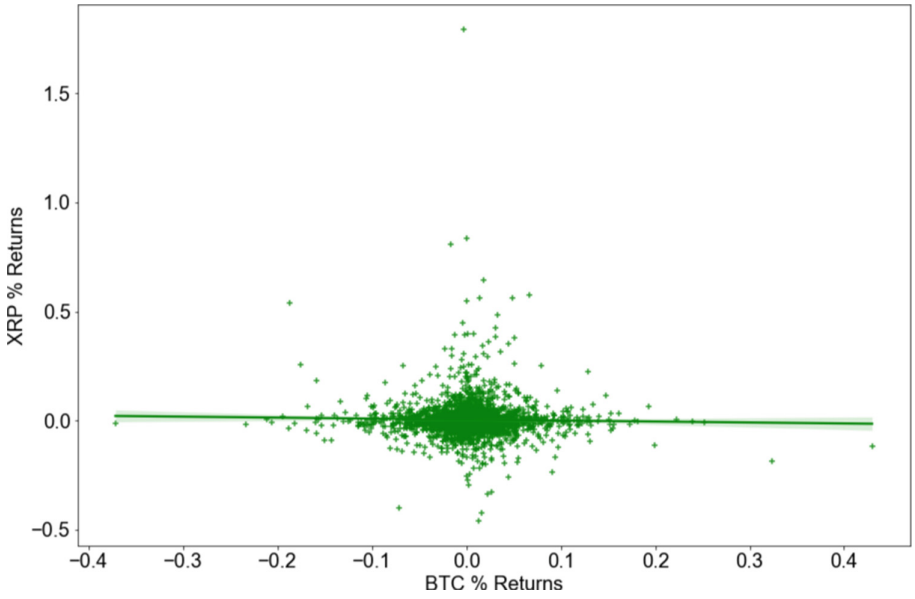

Fig. 9. Relation between returns of BTC and XRP.

References

1. Beck, R., Müller-Bloch, C.: Blockchain as radical innovation: a framework for engaging with distributed ledgers as incumbent organization (2017)
2. Christin, N.: Traveling the silk road: a measurement analysis of a large anonymous online marketplace. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 213–224 (2013)

3. Dirican, C., Canoz, I.: The cointegration relationship between Bitcoin prices and major world stock indices: an analysis with ARDL model approach. J. Econ. Financ. Account. **4**(4), 377–392 (2017)
4. Eyal, I., Sirer, E.G.: Majority is not enough: bitcoin mining is vulnerable. In: Christin, N., Safavi-Naini, R. (eds.) FC 2014. LNCS, vol. 8437, pp. 436–454. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-45472-5_28
5. Jiang, Y., Nie, H., Ruan, W.: Time-varying long-term memory in bitcoin market. Financ. Res. Lett. **25**, 280–284 (2018)
6. Katsiampa, P.: Volatility estimation for bitcoin: a comparison of GARCH models. Econ. Lett. **158**, 3–6 (2017)
7. Lee, S.J., Siau, K.: A review of data mining techniques. Ind. Manag. Data Syst. **101**, 41–46 (2001)
8. Ostertagova, E., Ostertag, O.: Forecasting using simple exponential smoothing method. Acta Electrotechnica et Informatica **12**(3), 62 (2012)
9. Pieters, G., Vivanco, S.: Financial regulations and price inconsistencies across bitcoin markets. Inf. Econ. Policy **39**, 1–14 (2017)
10. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23**(4), 3–13 (2000)
11. Rajkumar, S.: Cryptocurrency historical prices (2021). https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory
12. Raymaekers, W.: Cryptocurrency bitcoin: disruption, challenges and opportunities. J. Paym. Strat. Syst. **9**(1), 30–46 (2015)
13. Salman, A., Razzaq, M.G.A.: Bitcoin and the world of digital currencies. In: Financial Management from an Emerging Market Perspective, pp. 271–281 (2018)
14. Sattarov, O., Jeon, H.S., Oh, R., Lee, J.D.: Forecasting bitcoin price fluctuation by twitter sentiment analysis. In: 2020 International Conference on Information Science and Communications Technologies (ICISCT), pp. 1–4 (2020). https://doi.org/10.1109/ICISCT50599.2020.9351527
15. Yang, W., Garg, S., Raza, A., Herbert, D., Kang, B.: Blockchain: trends and future. In: Yoshida, K., Lee, M. (eds.) PKAW 2018. LNCS (LNAI), vol. 11016, pp. 201–210. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97289-3_15
16. Ye, Z., Wu, Y., Chen, H., Pan, Y., Jiang, Q.: A Stacking ensemble deep learning model for bitcoin price prediction using Twitter comments on bitcoin. Mathematics **10**(8), 1307 (2022)
17. Zaman, S., Yaqub, U., Saleem, T.: Analysis of Bitcoin's price spike in context of Elon Musk's Twitter activity. Glob. Knowl. Mem. Commun. (2022). https://doi.org/10.1108/GKMC-09-2021-0154