National College of Ireland

Bitcoin Price Prediction Using Time-series Analysis and Sentiment Analysis on Twitter Data in Cloud Environment

MSc Research Project
Cloud Computing

Siddharth Prashant
Student ID: x20154658

School of Computing
National College of Ireland

Supervisor: Dr. Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Siddharth Prashant
Student ID:	x20154658
Programme:	Cloud Computing
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Majid Latifi
Submission Due Date:	31/01/2022
Project Title:	Bitcoin Price Prediction Using Time-series Analysis and Sentiment Analysis on Twitter Data in Cloud Environment
Word Count:	XXX
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Bitcoin Price Prediction Using Time-series Analysis and Sentiment Analysis on Twitter Data in Cloud Environment

Siddharth Prashant
x20154658

Abstract

The motivation of the project came due to the sudden surge in the prices of Cryptocurrencies, especially Bitcoin. Bitcoin is considered an investment asset. The price of Bitcoin is very fluctuating, volatile and depends on the Crypto-currency market and its demand. In this research, I have performed time-series based analysis and sentiment analysis based on the Twitter data to predict the price of bitcoin. For time-series analysis I have obtained better outcomes using Long short-term memory (LSTM). On the other hand, for the Twitter-based sentiment analysis method, the sentiments of tweets have been identified and based on the sentiment score the bitcoin price has been predicted. For Twitter-based sentiments Recurrent Neural Network (RNN) architecture has provided better outcomes in terms of Bitcoin price prediction. The model with minimum MSE (Mean Square Error) and MAE (Mean Absolute Error) score has been considered as the optimal model for predicting the price of bitcoin.

1 Introduction

The advent of digital currencies has brought a better financial option to the people. In the current time, the people have adopted and sought after the emerging market of these digital currencies. These currencies have various edges over conventional currencies. The digital or virtual currency generally known as Cryptocurrency have no physical counterparts or in-store value. Cryptocurrencies consummately rely on the digital system and operations. The conventional currencies such as fiats currencies have in-store value but it is centralized. Authorities and third-party entities can have control over the fiat currencies and could be manipulated at any time. The only merit that exists for this is its balanced volatility. Present world economics and trades rely on the existing conventional currencies. But the majority of these challenges endured due to fiat currencies can be vanquished by the utilization of Cryptocurrencies. Being relied on the open-source digital system it works on decentralized mechanisms. Cryptocurrencies can therefore not be controlled or manipulated by multiple authorities, agencies or third-party entities. The novel technology behind the mechanism of Cryptocurrencies is the Blockchain system which makes it much more reliable to replace conventional currencies. But the significant downside over its advantages is its high volatility. Because of such volatility pace, it makes a great hurdle to replace conventional currencies as it may hinder cross-border or

inter trades. As a stable value cannot be predicted, this can become a downside for the trades

Currently, the most transacted and reliable Cryptocurrencies out of the heap of forty thousand Cryptocurrencies is Bitcoin (BTC). It is the most traded Cryptocurrency and the accumulative market value of Cryptocurrency is over one trillion dollars. Although, the complete adaptation of Cryptocurrencies is not achieved due to the higher volatility of the price trends. This builds a sense of fear and occlusion for losing the monetary value due to volatility. Not only this but the precise prediction of the price trends will help the investors and entities to take firm investment decisions and yield maximum growth output. Due to the unusual and complex trend of Cryptocurrencies, it becomes a multifarious task for the experts to predict the accurate trend of the price. Therefore, a mathematical and computational mechanism must be engaged to predict the future and upcoming price trends in Bitcoin which provides a sense of relief to the investors.

1.1 Research Question

Predicting the price of bitcoin or any Cryptocurrency is still an open area of research. Various authors and researchers have explored multiple ways of predicting the price of bitcoin. In this work, I try to solve the following research questions.

- How effectively the bitcoin prices can be predicted using time-series analysis? What is the impact of Twitter tweets on Bitcoin Prices?
- Which algorithm accurately predicts the price of Bitcoin based on time-series analysis?
- Which algorithm accurately predicts the price of Bitcoin based on Twitter Sentiment Analysis?

1.2 Research Objectives

1.2.1 Study of relevant Machine Learning and Deep Learning Models

In the existing approaches, the traditional and simple machine learning models of time-series analysis such as linear regression, Auto-Regressive Integrated Moving Average (ARIMA), Error Trend and Seasonality (ETS), is being utilized to predict the Bitcoin price trend. Multiple investors, researchers and entities utilized these approaches but these had various drawbacks. The conventional approach couldn't analyze the non-linear and complex price trends of Bitcoin. The machine learning time-series framework was prone to a large number of uncertain features for forecasting the Cryptocurrency (Bitcoin) price trends. Moreover, this technique fails to identify the long-term pattern. Therefore, the utilization and implementation of advanced deep learning algorithms can overcome the challenges of the conventional technique.

Most researchers are highly interested in recognising how Twitters tweets impact the fluctuation in bitcoin prices. Twitter has predictive power for a wide range of events especially in Cryptocurrency and financial markets. I have to recognise up to what level Bitcoin prices can be fluctuated by the tweets / public opinion. By analysing the sentiments of posted tweets related to Bitcoin (Cryptocurrency) I can forecast the price value of Bitcoin. Twitter sentimental analysis can be significant in the predictive analysis of Bitcoin Prices.

1.2.2 Proposing Research Model

Therefore, in this research, I have performed a time series analysis as well as Twitter sentiment analysis to predict the price of Bitcoin. For the Time series analysis, I have implemented advanced deep learning architecture such as Simple RNN, LSTM and Custom Model. These models have been evaluated based on the MSE (Mean Square Error) score to identify the best model for Bitcoin price prediction based on the time-series analysis. However, our second method utilizes the twitter-based sentiment analysis, where I have collected the 10 days of tweets about bitcoin and based on the sentiments of the people, I will try to predict the price of Bitcoin. In this work, I have implemented the 3 machine learning models to predict the price of bitcoin based on Twitter sentiments. The models are Random forest, XGBoost and RNN (Recurrent Neural Network). These models are evaluated based on the MAE score.

1.3 Document Structure

After this introduction, in Section 2I have reviewed the related work from various researchers in this domain. The section is divided further into subsections of Machine learning and Deep Learning Techniques. In Section 3, I have analysed the various model. This section is further divided into subsections for Time-series analysis for Bitcoin prices and Twitter sentiment Analysis on Bitcoin Prices. In the Design Specification section, I have presented the 2 best models (RNN and LSTM) and one custom model that I developed as the novelty of my research. In section 5, I have shared the details of the required resources and how I have trained the model. In section 6, I have evaluated all the trained models to choose the best model. In section 7, I have concluded my findings and future work.

2 Related Work

In this section, I reviewed the studies by various researchers on this domain. Furthermore, this section is divided into the following subsections of Machine Learning Technique and Deep Learning Technique.

2.1 Machine Learning Technique for Forecasting

Mangla (2019) proposed a method to predict the price of bitcoin using Machine Learning algorithms. An accurate estimation of the bitcoin price was done considering various factors which directly or directly affect the trends. For the dataset, the bitcoin's previous price and timestamps were utilized. The paper considered four algorithms to implement to the model which are Logistic Regression, Support Vector Machine, Recurrent Neural Network and ARIMA. These algorithms performed well for certain days but have poor performance when considered for long-term prediction. Furthermore, logistic regression, SVM, RNN, ARIMA achieved the accuracy of 47%, 48%, 53%, 50% respectively. Considerably the accuracy is poor which could be enhanced with the implementation of several other techniques. In another paper by the same author Mangla and Rathod (2018), the paper showed data analysis of unstructured data and processing through utilizing big data tools such as Hive and Machine learning algorithms such as Linear Regression. This

paper showed the working mechanism of Machine Learning for Big Data as it could extract the hidden and complex features of the data. Through this, future trends could be predicted. Furthermore, a big data tool such as the Hadoop Platform is utilized for a faster process of a large dataset. This platform uses the Map Reduce technique which ultimately enhances the pace. Finally, it also showed that the combined use of the Machine Learning Algorithm and Hadoop platform will make the model more efficient.

Raju and Tarif (2020) studied the real-time prediction and Sentimental Analysis of Bitcoin price using the Machine Learning Approach. The paper compared the extensive correlation between the bitcoin price trends and the comments extracted from the Twitter and Reddit posts. Through this several machine learning algorithms are implemented to predict the accurate future price trend of bitcoin. Two approaches namely ARIMA and LSTM are utilized and compare the future predicted price trend and the sentimental analysis. This comparative analysis showed that LSTM has higher performance than the ARIMA model. Moreover, the paper suggested the future possibilities of study on the hybrid LSTM model to build an autonomous Cryptocurrency trading platform.

Velankar et al. (2018) also proposed a machine learning approach to predict the bitcoin price trend. The paper compared two machine learning algorithms which are Bayesian Regression and Generalized Linear Model (GLM) Random Forest. The dataset for these models was obtained from sources such as Quandl and Coin Market Cap from which the sample was feature engineered. Several features such as block size, total coins, day high or low, transactions and trade volume Ire considered. Further, the sample was standardized to balance the inputs and enhance the efficiency of the model. After such implementation, the best and most efficient approach is implemented for the model.

Similarly, Reddy and Sriramya (2020) researched the machine learning algorithms for the projection of bitcoin prices. The paper utilized machine learning approaches such as Logistic Regression, K-Nearest Neighbour Naïve Bayes, Random Forest, Decision Tree, Least Absolute Shrinkage Selection Operator (LASSO) model. After implementation of these model, evaluation metrics like Residual Sum of Square (RSS) is computed to obtain the accuracy rate of each model. Finally, the paper concluded that the LASSO model is the best performing among all the models and also suggested the possibility of future study for a novel algorithm to enhance the profitability of bitcoin traders.

Jaquart et al. (2021) proposed an algorithm based on machine learning to predict short-term Cryptocurrency trends. The researchers in this paper predicted the bitcoin market trend over the period of one minute to sixty minutes. Various machine learning approach such as GRU, FNN, LSTM, LR, GBC, RF is being utilized and analyzed in this paper. The model extracted the extensive features from the samples and predicted the future trend. The paper showed that the recurrent-based and gradient boosting classifiers are the best performing for the prediction approach. The model precisely forecasted the rate of return on the given period and also predicted the negative returns if any. For short-term holders, it could calculate the returns including the transaction rate which majorly was negative.

Ho et al. (2021) explained the implementation of linear regression and the long short-term memory (LSTM) model for evaluating the price of bitcoin. The novelty of this study was that it utilized the combined approach of machine learning and artificial intelligence to forecast the bitcoin price. Through the approach of feature extraction, the best features of the dataset are extracted. Also, an impressive graphical user interface (GUI) is created for the convenience of the traders. The GUI is built using the Tkinter library in python. Overall, the accuracy of the model stood at 99.87% whereas the limited error rate stood

at 0.08%. This showed that the model is highly optimized.

Umer et al. (2019) showed a comparative analysis of various machine learning models for stock price prediction. This paper implemented ML models such as Linear Regression, Three Month Moving Average (3MMA), Exponential Smoothing (ES), and Time Series Forecasting presenting the performance of each model. Moreover, a statistical application for graphical and tabular representation which is Microsoft Excel is utilized. For the dataset, the sample trends of stock such as Amazon (AMZN), Apple (AAPL), Google (GOOGL) are sourced from the yahoo finance repository. With the implementation of these models, the stock price trend of one month is predicted. Further, the analysis showed that the ES model could predict the accurate price trend among all other models.

M. Obthong and Wills (2020) surveyed the machine learning approach for the stock price prediction. This paper precisely reviewed all the major machine learning frameworks for stock price prediction. Input, output, merits and demerits of each ML algorithm were discussed and suggested for multiple situations. Furthermore, the paper suggested a future task where the stock price can be predicted with the combination of sentimental analysis. As the stock price majorly depends on the dual factors of technical and fundamental of stock trends. This shows an emerging scope of the study for future enhancement of the machine learning framework.

Chen et al. (2019) proposed an approach for bitcoin price prediction utilizing a machine learning mechanism and the framework of sample dimension engineering. The price is first predicted by the machine learning approach and multiple highly dependent features of the trends are extracted for future prediction. Algorithms such as Logistic Regression and Linear Discriminant Analysis (LDA) are implemented.

Phaladisailoed and Numnonda (2018) also proposed and compared a machine learning model for forecasting the price trend of bitcoin. The price trend of the sample was sourced from the Kaggle website which was from the bitstamp repository. It included one-minute interval trade records from the previous time frames. After this, the requisite features were extracted from the sample. The machine learning algorithms such as Theil Sen Regression, Huber Regression, LSTM and GRU were utilized for the implementation of the model. Finally, the GRU model showed the best MSE and R2 scores which stood at 0.00002 and 0.992, respectively. But the computational time was lowest in the Huber Regression model. Although the endpaper also showed the importance of sentiment on the future price trends.

2.2 Deep Learning Technique for Forecasting

M. M. Patel and Kumar (2020) showed an approached method using a deep learning algorithm to predict the prices of cryptocurrencies for financial entities. The paper aims to propose a hybrid deep learning model considering LSTM and GRU models. Furthermore, the paper analyzes the price trends of just Litecoin and Monero. This model can conveniently extract multiple complex features among the samples. Also, it enhanced the outcomes of the LSTM model and reduced the error rate of the model. The proposed hybrid model outperformed all previous models. Moreover, the paper also suggested the future task to introduce a more complex model by integrating the sentimental data.

Lahmiri and Bekiros (2021) utilized a Deep Feed Forward Neural Network, a form of deep learning algorithm, to anticipate the Value of bitcoin using high-frequency sampling. To do this, the researchers gathered High-frequency sampling bitcoin intraday value statistics. The researchers utilized three distinct DFFN techniques: the Powell-Beale method,

the Resilient technique, and the Levenberg-Marquardt method. After conducting their experiments, they discovered that the Levenberg-Marquardt method outperforms the other techniques in terms of reliability, with an RMSE value of 1.4406.

Spilak (2018) also proposed a deep neural network (DNN) approach for forecasting the price of Cryptocurrency. Three DNN models which are deep feedforward networks, recurrent neural networks, long short-term memory networks are considered for the models in this paper. In the methodology, the features of the sample were fine-tuned and standardized to enhance the model and further trial and error mechanism was implemented. Furthermore, the models shown in the paper outcasted the performance of the conventional strategy of the CRIX index. However, these models were unable to predict based on the external factor effects. Therefore, the future works aim for the prediction with external factors such as financial policies and risk measures.

Biswas et al. (2021) suggested RNN, LSTM, and GRU methods to anticipate bitcoin prices. The researcher focused on two kinds of cryptocurrencies in this task: Litecoin (LTC) and Monero (XMR). They used pricing statistics from Investing.com to develop the algorithm over a three and half years of period. The information primarily consists of five characteristics: price, opening, closing, top, bottom, and quantity. The researchers primarily assessed the algorithm utilizing Root Mean Square Error (RMSE) while forecasting bitcoin prices for the subsequent week. Following their study, researchers discovered that the GRU and LSTM models dominate bitcoin value forecasting due to their capacity to recall and retain the features of relatively transitory information.

In a paper, Mahmud and Mohammed (2021) surveyed a deep learning approach for time series forecasting. Various literature reviews from multiple domains have been discussed in the paper. Classical methods of time series such as ARIMA, NARMA, SARIMA, etc and novel approaches such as Artificial Neural Network (ANN), Convolutional Neural Network (CNN), etc are precisely discussed and explained. Each application using time series deep neural network is shown. Also, tuning of various parameters and modifications in the model further enhanced the outcome. Finally, the paper concluded that deep learning algorithms showed better performance and accuracy for solving the time series problems. A similar survey for time series forecasting utilizing a deep learning approach was done by Torres et al. (2020). Primarily the paper generated a time series forecasting problem incorporating the mathematical fundamentals. Generally, other researchers used the conventional deep learning approach to predict the time series problems but in this paper, algorithms such as feed-forward networks, RNN and CNN are utilized. Furthermore, suitable features are extracted and the parameters are tuned to enhance the model. Moreover, this paper showed the merits and demerits of each algorithm for multiple applications. The accuracy and performance of each model are also analyzed. The researcher also identified the drawbacks in the previous tasks and suggested some of the tasks for future work.

I. E. Livieris and Pintelas (2021) proposed an advanced CNN-LSTM model for Cryptocurrency price trend prediction. The research paper relied on the multi-input framework for the accurate prediction of Cryptocurrency price trends. Initially, the samples are sourced and complex features are extracted from them. Then it is implemented over the multiple LSTM and Convolutional layers. Through this framework, the paper proposed a multi-input deep neural network mechanism called MICDL. Furthermore, classification metrics such as Graphical Model (GM) and Sen x Spe are considered to evaluate the performance of the model. Additionally, the paper discussed the limitation that occurred due to conventional approaches as these only depended on predicting the performance

rather than the development of the performance.

Albariqi and Winarko (2020) conducted a study to forecast the value of Cryptocurrency. The researcher of this study uses recurrent system topology to forecast short-term and long-term Cryptocurrency value fluctuations. The researcher employs Multi-layer Perceptron (MLP) and Recurrent Neural Network (RNN) methods. The information for the value of Cryptocurrency was gathered from blockchain.info, and researchers deemed Fourteen important characteristics for forecasting out of the 35 distinct aspects of the ledger and Cryptocurrency value. The researchers investigated a set of characteristics tweaking as well as specific variables in their investigation. Over several experiments, the researchers discovered that long-term forecasting had higher efficiency than short-term forecasting. The MLP algorithm had the best efficiency of 81.3%.

In the paper by Politis et al. (2021), he showed an alternative approach for predicting the price of ether utilizing advanced deep learning algorithms. The paper also proposed the work to overcome the challenges followed by the black-box approach. In the proposed framework, a method of feature engineering was considered to extract complex characteristics. Based on these features, a batch of deep learning algorithms was built such as utilizing LSTM, GRU and TCN. Through this model, the existing value of ether could be predicted with both the short-term and long-term trends.

Similarly, Chandrasekaran (2018) predicted the price of Litecoin using the algorithms such as ARIMA and LSTM. The data source for this study was of the period five and half years on which these models were implemented. Further, these models were then evaluated by metrics such as MAPE, ME, MAE, and RMSE. Finally, the paper showed LSTM as the best performing model when utilized with MAPE achieving an accuracy rate of 5.759%.

2.3 Sentiment Analysis in Twitter

In a study by Li et al. (2019), the author did a sentiment-based prediction of alternative Cryptocurrency price fluctuations. For this, advanced learning algorithms are utilized which are Gradient Boosting Tre Model. Here for the implementation, the tweets were extracted from the Twitter database for the duration of 3 and half weeks every hour. Then the author implemented the method of indexing to develop a weighted and an unweighted index from the obtained tweets with multiple sentiments. Furthermore, to train the model, the labels were made such as positive, negative, and neutral. Then the model of extreme gradient boosting regression tree model was implemented to train the model. The model showed an effective result. In another paper by Sattarov et al. (2020), the bitcoin price was forecasted using the technique of sentimental analysis in Twitter posts. To forecast the price fluctuation, various time-series algorithms were utilized. The technique of valence aware dictionary and sentiment reasoner (VADER) was implemented in the model. Furthermore, the dataset was obtained from the Quandl repositories. In the final take, the model was able to achieve an accuracy of 62.48% using the sentimental analysis of the tweet.

The opinion of the people was evaluated using the Twitter sentimental analysis on the bitcoin by the author named Hassan et al. (2021) The data was scrapped using the RStudio where around 15,000 tweets on bitcoin were obtained. Using various machine learning algorithms, the sentiments were evaluated. A detailed report on the dataset was then obtained which provided the sentiment of each tweet in the threshold. In addition to these, various other factors influencing the fluctuation of the prices were discussed in

Table 1: Comparison of Different Key Works for this Study

Paper Title	Author and Year	Method	Advantages	Future Scope / Disadvantages
Bitcoin Price Prediction Using Machine Learning	Mangla (2019)	Logistic Regression, Support Vector Machine, Recurrent Neural Network and ARIMA	Compared the Machine learning approach	Efficiency could be enhanced
Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis	Raju and Tarif (2020)	ARIMA and LSTM	Comparative analysis showed that LSTM has higher performance than the ARIMA model	No disadvantage found.
Bitcoin Price Prediction Using Machine Learning and Artificial Neural Network Model	Ho et al. (2021)	LR and LSTM	GUI was incorporated to the model	No disadvantage found
An Advanced CNN-LSTM Model for Cryptocurrency Forecasting	I. E. Livieris and Pintelas (2021)	CNN-LSTM	the paper proposed a multi-input deep neural network mechanism called MICDL	depended on predicting the performance rather than the development of the performance
Ether Price Prediction Using Advanced Deep Learning Models	Politis et al. (2021)	LSTM, GRU and TCN	the work to overcome the challenges followed by the black-box approach	existing value of ether could be predicted with both the short-term and long-term trends
Prediction of Litecoin prices using ARIMA and LSTM	Chandrasekaran (2018)	ARIMA and LSTM	these models were then evaluated by metrics such as MAPE,ME, MAE, and RMSE.	the best performing model when utilized with MAPE achieving an accuracy rate of 5.759% only

the paper. The paper provided an overview of the Cryptocurrency environment.

The comparative analysis for the various studies by different researchers is shown in Table 1.

3 Methodology

This Research is aimed to select the most appropriate deep learning model based on the time series for predicting the price of Bitcoin. The time-series analysis is performed by analysing the past trends observed in historical data. Other than that, this research also aims to find the impact on the value of bitcoin-based on Sentiments of tweets. To achieve this I have proposed a multi-step framework that includes accessing and cleaning of data set, data pre-processing, visualisation and exploratory data analysis, selection and extraction of features, training of model and prediction. In this section, each step is covered in detail. There are two different methods I have approached for predicting the price of bitcoin, in the first approach I have performed the time-series based analysis for bitcoin prices and in the second approach, I identified the impact of Twitter sentiments on the bitcoin price.

3.1 Predicting Bitcoin Price based on the Time-Series Analysis

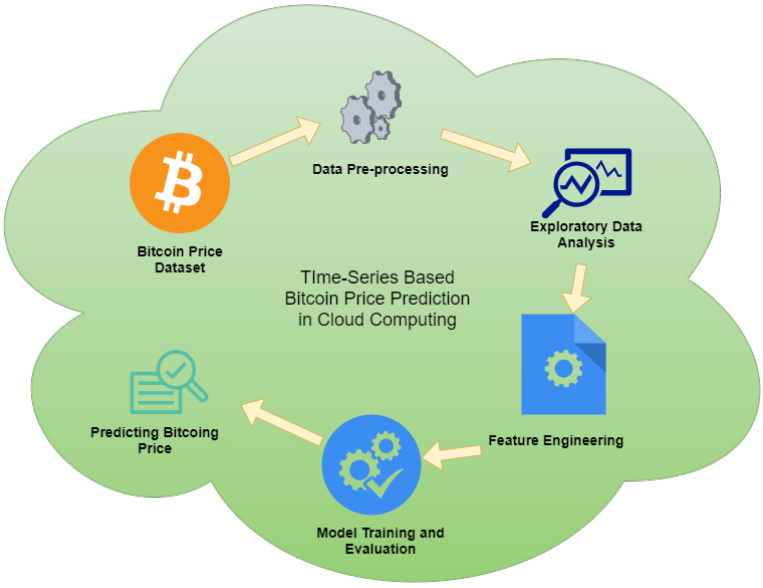


Figure 1: Time-series based Framework for Predicting Bitcoin Price

3.1.1 Data Set Description

The Bitcoin data set is collected from open source for the period of Jan 2012 to March 2021, with a minute to minute updates. This data set contains 8 columns as Timestamp, open, high, low, close, volume(BTC), volume(currency), weighted Price. These features are in numerical format except timestamp which is in UNIX time format. The total size of the data set is around 318 Mb.

3.1.2 Data Pre-processing

In data pre-processing I first change the timestamp feature into the required date-time format which is followed by data cleaning in which I eliminated null values from each column. The names of the columns have also been changed for better understanding. A detailed explanation about each feature is described in the next subsection.

3.1.3 Exploratory Data Analysis

The pre-processed data contains 3613769 rows × 9 columns and I performed data analysis and visualisation. On analysis, I found that the value of open price increases rapidly for initial stamps i.e, up to 10000 while it becomes constant after 20000-time stamps as shown in figure 2.

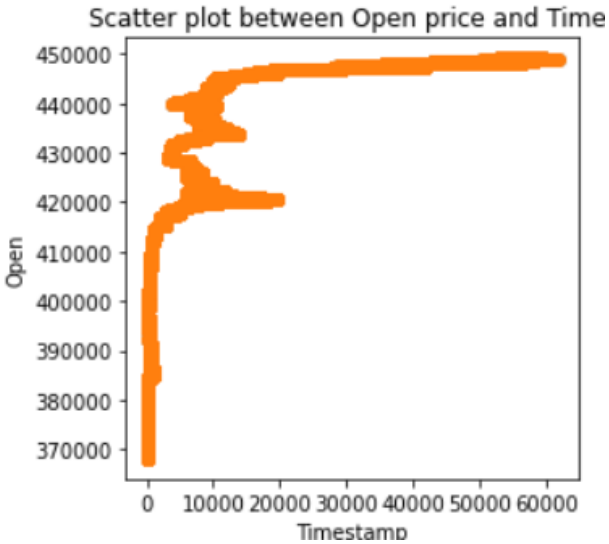


Figure 2: scatter plot between open price and time

High Price feature has shown interesting trend that for values up to 500, percentage of increasing is high which sets to lowering down trend till 1250 and set to negligible change after this value as shown in figure 3.

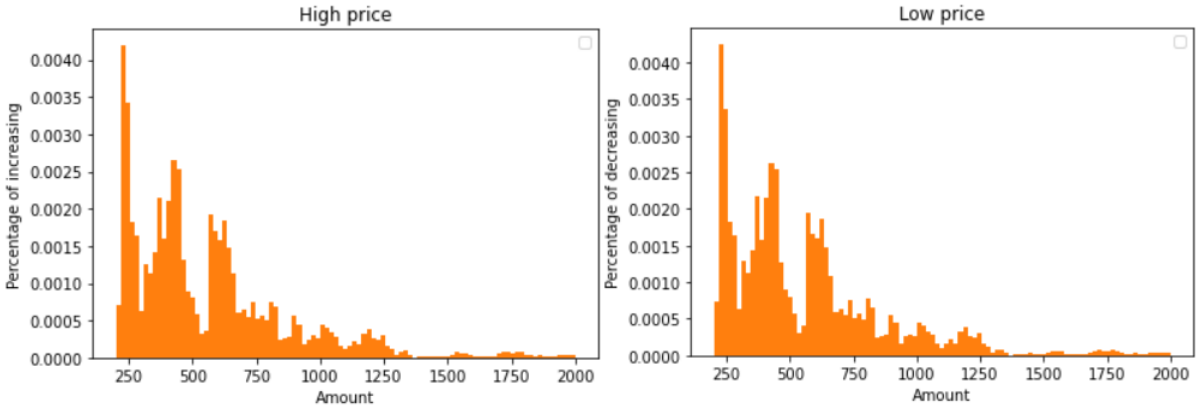


Figure 3: Plot for High price and Low price versus percentage of Increasing

A similar analysis is performed for the Low-value feature which shows that for initial values rate of percentage of increase is slightly higher than that of the High Price feature as shown in figure 3. For open feature, I have observed that since date up to the year 2017 open price was constant and in the year 2018 a spike has been observed and after this, an increasing trend has been observed as shown in figure 4.

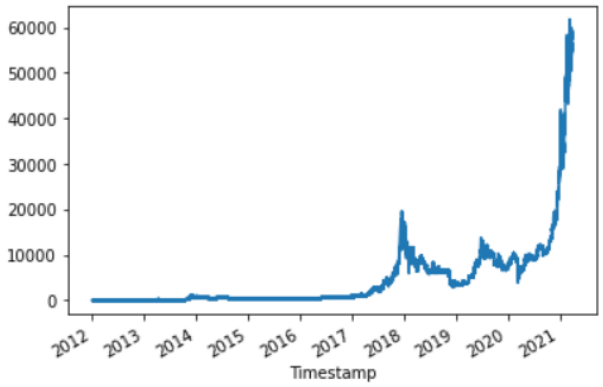


Figure 4: scatter plot between open values and year

In all these years I found that some spikes which are not common and can be considered as outliers so I have to eliminate them from the data set as shown in figure 5.

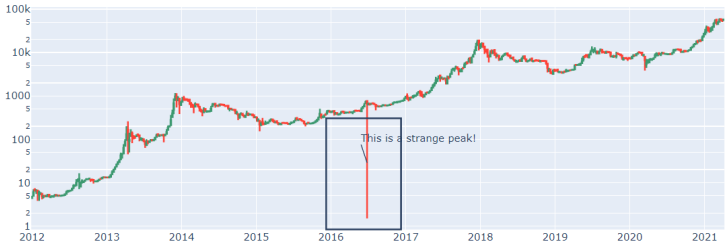


Figure 5: outlier/spike in bitcoin trend

Figure 6: eliminated outlier plot of bitcoin

Rolling averages help us track and analyze progression levels for just about any given metric. It gives a more realistic picture than the standard average, enabling us to take corrective action as and when needed. Rolling averages are especially useful in forecasting revenue. I have applied a rolling window of 5 and observed the following trends as shown in figure 7.

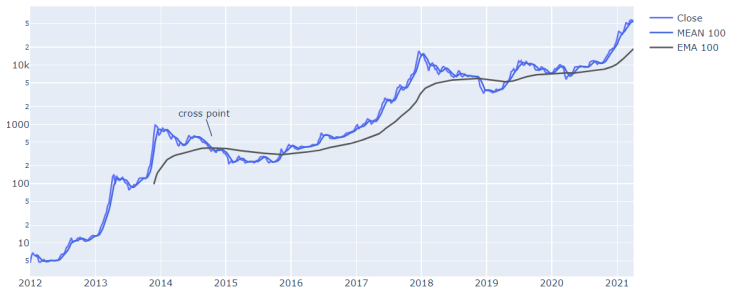


Figure 7: Rolling mean plot

3.1.4 Feature Engineering And Data Scaling

For such a task it is important that I should perform feature extraction and feature selection which comes under feature engineering. As timestamp column is UNIX format which has been changed into the date-time format. Scaling of data is important in such tasks so I have used MinMaxScalar within range 0 to 1 after changing values of data to float32 format. The close feature has been assigned as label/target variable and the remaining all are considered as input variables for the model.

3.1.5 Model Training and Testing

In this section, I have deployed 3 deep learning models for comparative analysis. First, the dataset is spitted into training and testing datasets in 80 to 20 ratio respectively. Three models i.e. simple RNN, LSTM and a custom model which is developed using Keras tuner are initialised with the help of Keras and TensorFlow libraries and trained for 100 epochs.

3.1.6 Model Evaluation

since this is time series prediction work and continuous values have to be predicted hence it becomes a regression problem. So I have used Mean Square Error(MSE) as the evaluation metric. The best model can be selected based on the model with the lowest MSE. I have also plotted a line plot between real bitcoin prices and predicted bitcoin prices for better understanding and visualisation.

3.2 Impact of Twitter Sentiments on Bitcoin price

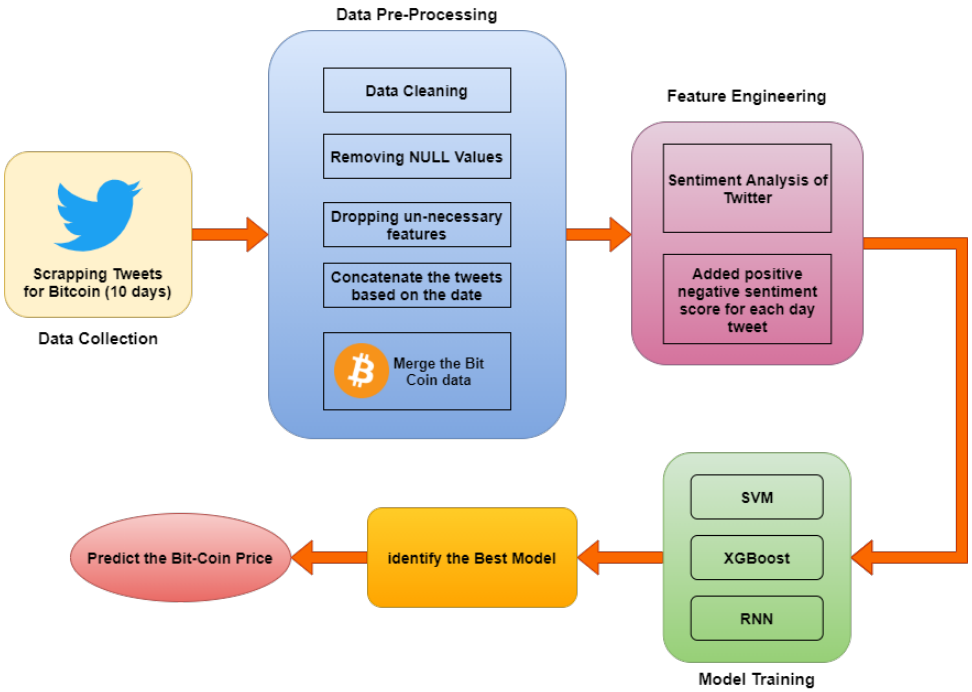


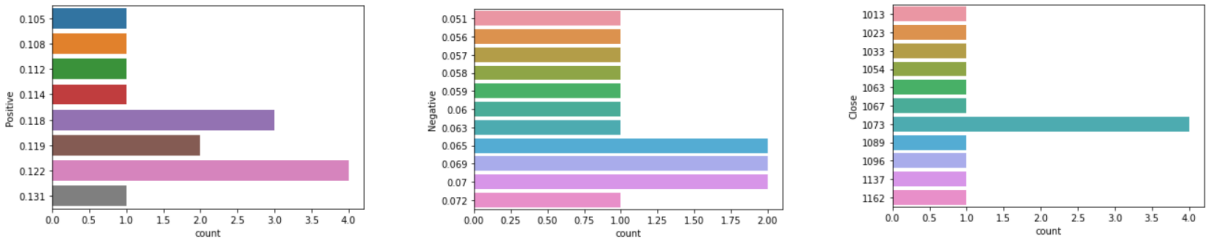
Figure 8: Proposed Flowchart For Predicting Bitcoin Price based on Twitter sentimental analysis

3.2.1 Data Set Description

For this work, I have downloaded two data set separately which includes one as tweets and the other as bitcoin values corresponding to the same date. Twitter data has been web scrapped from the last 10 days of tweets from Twitter. This data set contains 6 different features, which include keyword, content, Conversationld, date, retweet-Count, tweet-URL while The second data set has been downloaded from Investopedia.com. This data set contain 6 different features, which includes Date, Price, Open, High, Low, Volume and percentage change.

3.2.2 Data Pre-processing

In sequence to eliminate the noise from the dataset and for better analysis, data Pre-processing is a necessary step. The sequence of data pre-processing is incorporated in the cleaning of data, generalization of data, null values removal, etc. Some of the features obtained from Twitter data are non-useful such as Conversational, date, retweet-Count, tweet-URL. Therefore, the first step is to drop unnecessary columns then change the name of columns accordingly. Further, I have checked for the null value and eliminated null values. After all these steps, the type of Date column is converted into a date-time format to extract only dates from each column. At last, I concatenated all tweets of the same date into one row. After pre-processing of Twitter data, I have cleaned the bitcoin price data such as noise removal, changing the date into Date Time format followed by assigning price value to Twitter data according to the same dates. In the third step, the



(a) *PositivePlot* (b) *NegativePlot* (c) *ClosePlot*

Figure 10: Count plot of Positive, Negative and Close features

task is to fill adjacent null values of bitcoin prices according to bitcoin price trend and add columns for sentimental analysis in the processed dataset.

3.2.3 Feature Engineering

Features that are used in supervised learning are converted from raw data into useful information. To perform a better prediction, the unnecessary features that need to be removed such as Keyword, conversation, retweet-count, tweet-URL and from bitcoin price dataset, I have used the only close column as the target variable. I have concatenated all tweets of the same date and developed a new feature as per the requirement and merged both datasets corresponding to the same date. I have used NLTK library (Vader Lexicon) for extracting the sentiments from the tweets and based on which three new columns have been added (Negative, positive, Compound) in the data to store the sentiments extracted from the tweets as per the dates as shown in figure 9.

	Date	Prices	Comp	Negative	Positive
0	2021-11-10	64932	1.0	0.066	0.107
1	2021-11-08	67527	1.0	0.045	0.113
2	2021-11-09	66904	1.0	0.048	0.119
3	2021-11-12	64134	1.0	0.057	0.117
4	2021-11-11	64806	1.0	0.052	0.113
5	2021-11-07	63273	1.0	0.047	0.121
6	2021-11-14	65508	1.0	0.051	0.12
7	2021-11-13	64398	1.0	0.055	0.126
8	2021-11-15	63597	1.0	0.049	0.11
9	2021-11-16	60089	1.0	0.053	0.114

Figure 9: Processed Data after sentimental analysis of tweets

3.2.4 Exploratory Data Analysis

Since I have 5 features in the final dataset I have analyzed the frequency of all values in the dataset because I have data of 10 rows only for 10 days. This is important to analyze because it will prevent models from generating biased results if the frequency of any one value is high in numbers. I have plotted bar plots to visualize the following features. These all plots represents the frequency measure of values in the feature variables. as shown in Figure 10.

3.2.5 Model Training and Testing

In this work, I have used different ML and DL algorithms for prediction purposes. For each algorithm, 80% of the data samples were used for training and the remaining 20% of the data is used as a test data set. Algorithms such as Support Vector Machine (SVM), Xgboost, and Recurrent neural network (RNN) has been used to predict the price of bitcoin.

3.2.6 Model Evaluation

To select the best model for bitcoin price prediction based on Twitter sentimental analysis, the evaluation for each model has been performed. Predicting the price of bitcoin-based on sentimental analysis of tweets makes this a regression problem. I have used MAE (Mean Absolute Error), which refers to the magnitude of difference between the prediction of observation and the true value of that observation. Test MAE value is used to compare all implemented models.

4 Design Specification

Since I have deployed 3 different models for analysis, hence the architecture of all models are different. I have used Simple RNN, LSTM and a custom Model. In further subsections, I have described all models briefly.

4.1 Recurrent Neural Network Model

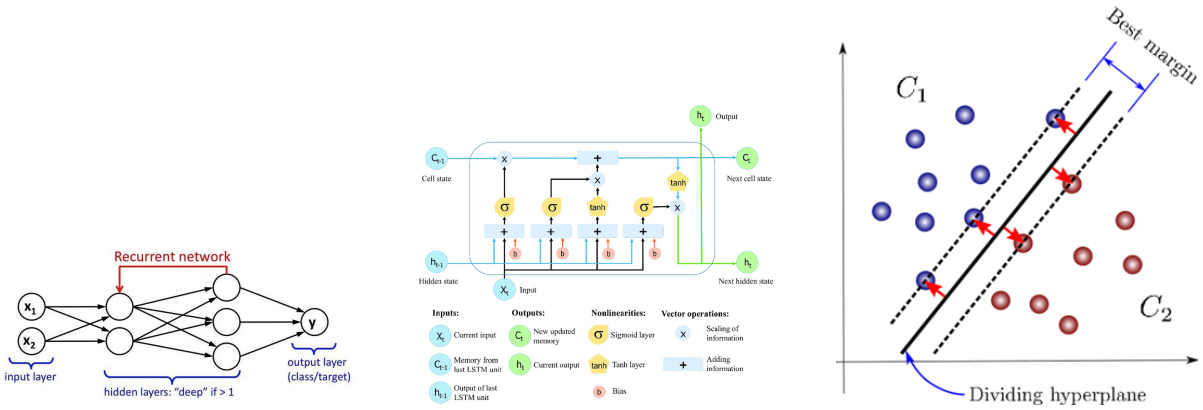
It is a class of neural networks that is good at modelling sequence data such as time-series data. RNN layer uses a for loop to iterate over the time steps of a sequence while maintaining an internal state that encodes information about the time steps it has seen so far. The network is first provided with a single time step. After this calculated state along with the current state a new state is calculated and this continues till forwarding propagation of the network. The output is calculated when all time steps are completed and then this output is compared to the actual output and an error is generated which is then backpropagated to the network to update weights and in this way, RNN is trained. The architecture is shown in figure 11a.

4.2 Long Short Term Memory Model

Long short term memory(LSTM) model is the same as that of RNN, its training is also the same as of RNN but its architecture is different as concepts of gates such as forget gate is introduced which is used to retain only important information from long sequences of information which produces the problem of gradient vanishing as in simple RNN networks. These networks are state of art models and have wide application areas, especially in time series data. The architecture of the LSTM model can be shown in figure 11b.

4.3 Support Vector Regressor

To perform the prediction of bitcoin values impacted by tweets I have used this machine learning algorithm because of fewer data. Support vector regression is a supervised



(a) *RNN* (b) *LSTM* (c) *SVR*

Figure 11: Architecture Model of RNN,LSTM and SVR

learning algorithm that is used to predict the best fit line for a dataset. SVR is based on SVM that stands for Support Vector Machines. These are used to divide the data in such a way that it categorises the data into n-dimensional space so it becomes easier to insert new data in the correct category. The boundary used to categorise the data is known as the hyperplane. The dimensions of the hyperplane depend on the parameters of the data. For example, if there are 2 parameters then the hyperplane will be a straight line, if there are 3 parameters then it will be a 2D plane. There can be multiple hyperplanes for the same dataset, the best one is the one that has the maximum margin which means the maximum distance between the two data points. The algorithm is shown in the figure 11c.

4.4 Custom Model

To make a custom model with LSTM, dense and dropout layers C have used Keras tuner. By use of such a method, I can figure out what number of layers, neurons are in each layer and which activation should be used to get the most optimized result on the same data. After deploying Keras tuner with training data by selecting the lowest Mean Absolute Error as evaluation metric I obtained the model as shown in figure 12.

Layer (type)	Output Shape	Param #
lstm_23 (LSTM)	(None, 1, 128)	71168
dropout_17 (Dropout)	(None, 1, 128)	0
lstm_24 (LSTM)	(None, 256)	394240
dense_13 (Dense)	(None, 30)	7710
dropout_18 (Dropout)	(None, 30)	0
dense_14 (Dense)	(None, 1)	31

Total params: 473,149
Trainable params: 473,149
Non-trainable params: 0

Figure 12: Custom Model Architecture

4.5 XgBoost Algorithm

It stands for Extreme Gradient Boosting, it is an open-source library that is quite efficient in predicting numerical data like prices, population, numbers, and heights. Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modelling problems. Ensembles consist of various decision trees and these trees are added individually to correct the mistakes of the previous one. Gradient descent algorithm is used for optimization purposes and initially, a loss function that is differentiable is chosen randomly to fit the model. As the loss gradient function is minimized when the model is fitted which gives it's the name "gradient boosting,".

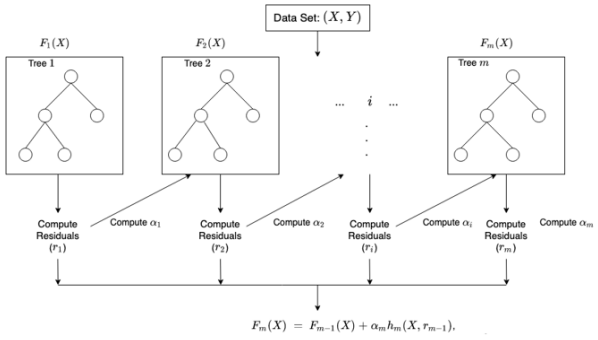


Figure 13: XgBoost Algorithm

5 Implementation

I have implemented three deep learning models (RNN, LSTM and Custom model) and the best model has been selected whose value of Mean Squared Error is lowest for the same data.ALL Models with ADAM optimization function and MSE as loss function is used and trained on the training dataset. All models are fitted with training data and trained for 100 epochs and tested on a test dataset. There are a number of libraries that have been utilized in order to implement the proposed work which includes pandas, NumPy, matplotlib, seaborn, TensorFlow, Keras, sklearn etc. The whole experiment is performed on Google Colab for training purposes with python as the programming language. The following specification was required in order to implement the model.

- Operating System : windows 10
- Random Access Memory (RAM) : 12GB (Provided by Colab)
- Hard disk : 15GB (Provided by Colab)
- Languages : Python, HTML
- Cloud Platform : Google Colab
- Python libraries : numpy, Pandas, matplotlib, tensorflow, numpy, seaborn and keras.

6 Evaluation

Evaluation of models is necessary because I have to choose the optimal model which can be only decided by the evaluation of models. Different metrics are used to evaluate the model. Since it is a regression analysis problem, I have used MAE and MSE metrics for evaluating the performance of the model.

6.1 Experimentation for time series analysis and prediction of Bitcoin

A model with minimum MSE score will be considered as the optimal model.There are two type of MSE i.e., training and testing MSE. However I will look into testing MSE value to consider best model.In each experiment, the number of epochs are used as 100 for training each model. The dataset has been splitted into the ratio of 80:20 where 3028 number of rows are used for training purpose and remaining 338 number of rows are used for testing purpose (validating the model performance).After this I have plotted a line plot which represents the Actual and predicted price of bitcoin.

6.2 Experiment 1 / Evaluation of RNN model

In the first experiment, I have used Recurrent Neural Network (RNN) to predict time series data of bitcoin. I have trained this model for 100 epochs on the training set of 3028 rows and the calculated value of training MSE.ADAM optimizer is used to optimize the model with a learning rate of 0.001. Model is tested on test data of 338 rows and I calculated the value of test MSE but to get a better look I plotted real value of test data against predicted value on test data by model on a line plot as shown in figure 14.

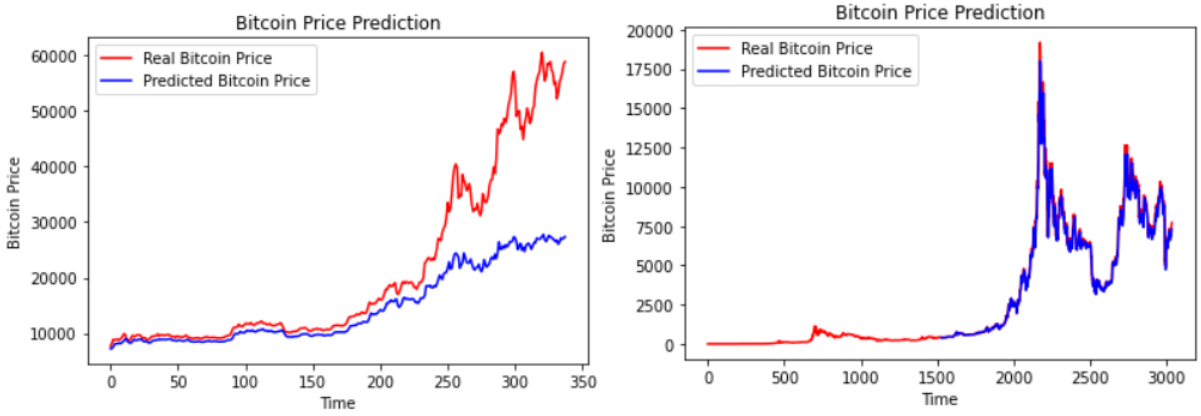


Figure 14: Line plot between Real and predicted Bitcoin Price Value for RNN and LSTM model respectively

If the GAP between the Predicted and Actual bitcoin price is higher, the model is not generating accurate results. Therefore, for an efficient model, the GAP between the actual and predicted price difference should be minimal.

6.2.1 Experiment 2 / Evaluation of LSTM model

In the second experiment, I have used the Long Short term Memory (LSTM) model to make predictions. This model is trained on the same training dataset comprised of 3028

rows for 100 epochs with an ADAM optimizer on a learning rate of 0.001. After complete training train score was found to be 33169.49 MSE. Model is tested on test data of 338 rows and MSE found to be 68940.73. I have also plotted a line plot to visualise the difference between predicted and real values of test data as shown in figure 14. In this graph, I observed that the difference between the predicted and actual bitcoin prices is minimum. Thus, I can say that the prediction obtained using the model LSTM is quite accurate.

6.2.2 Experiment 3 / Evaluation of custom model

In the third experiment, I have to build a custom model with LSTM, dropout and dense layers. I used Keras tuner to tune the number of neurons, activation function in each layer of the model so that I can find the optimized model with a lower RMSE value without overfitting the model. This has been trained on the same dataset with Adam optimizer and tested on the same test data. I found the train MSE value as 58812.18 and the test MSE value as 122377.11. To visualise the difference between the predicted and true value of test data I have plotted a line plot as shown in figure 15.

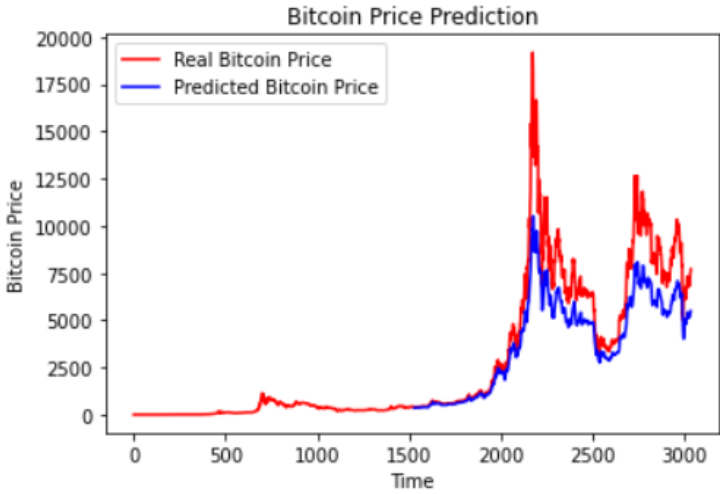


Figure 15: Line plot between Real and predicted Bitcoin Price Value (Custom Model)

.

The gap between the true and predicted values are higher as compared to the LSTM Model but minimum than the RNN model. Thus, I can say that Custom model predictions are better than RNN.

6.3 Experimentation for sentimental analysis of tweets and prediction of Bitcoin

6.3.1 Experiment 1 / Evaluation of SVR model

In the first experiment to predict bitcoin values impacted by tweets, I have implemented a machine learning model i.e., Support Vector Regressor. The training dataset is converted into a NumPy array with corresponding target values and fitted into the model for the training task.The parameters for model is decided as degree=3, gamma='scale', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache-size=200, verbose=False,

max-iter=- 1. After the training, the model is tested on the testing data and test MAE is calculated which is around 2237.47 and I printed predicted values to look at the deviation from real values.

6.3.2 Experiment 2 / Evaluation of XgBoost model

In the second experiment, I have implemented another machine learning model which is XgBoost Regressor. This is an ensemble method with boosting technique that is capable to perform better for such data sets. Model is trained with training dataset for 100 estimators for n-jobs=8 with learning-rate=0.300. After complete training, the model is tested on test data and MAE is calculated which is found to be 3108.19. MAE value for this model is higher than Support vector regression. hence I can conclude that the SVR model is better than the XGBoost model.

6.3.3 Experiment 3 / Evaluation of RNN model

In this experiment, I have used a deep learning model which is generally used for such type of sequential data. A recurrent Neural Network is proposed for the task. RNN model with Adam optimizer and MAE as loss function is trained on training dataset for 100 epochs. For every epoch, the training MAE is calculated using the RNN algorithm. Till 25 epochs loss decreases after this loss follow the up-down path which represents over-fitting in the model. This may be because I have a small dataset (10 days) for such a deep learning model. The training MAE has been shown in Figure **??** with the help of line graphs. The model is tested on test data and I have found MAE=1523.29, which is the lowest as compared to all the models.

6.4 Discussion

In this research, I have performed two different types of analysis to predict the price of Bitcoin, the methods are Time-series analysis and Sentiments analysis on Twitter data. After a certain set of experimentation for time series analysis, I have observed that the LSTM model outperforms as compared to all the models with minimum MSE score. The price prediction graph of LSTM mostly coincides with the line of real value price which justifies that our model is working well. Custom model build shows better results than RNN model which is clearly justified by plotted lines. Other than that, I have also calculated the MSE score for each model. The model having the minimum MSE score will be considered as the optimal model for bitcoin price prediction. After analysing the graph as shown in Figure **??**. I can say that the MSE score of LSTM is minimum, followed by a custom model and RNN. Thus, I can say that for time-series analysis the RNN performs poorly as compared to the LSTM and custom model. 16.

Considering the sentiment analysis on Twitter, after a certain set of experiments and comparing the MAE over the test dataset, I can say that RNN is the best performing for predicting bitcoin price as it achieves the minimum MAE score. But, it is not as always necessary that RNN provide the optimal results as our current dataset is very small and contains only 10 days of tweets. For the large dataset, the more complex architectures can provide better outcomes. On comparing the MAE score of the other two models, it has been found that the SVR model outperforms than XGBoost with an MAE Score of 2237.48 and XGBoost in our case poorly performs on the 10 days twitter data for bitcoin

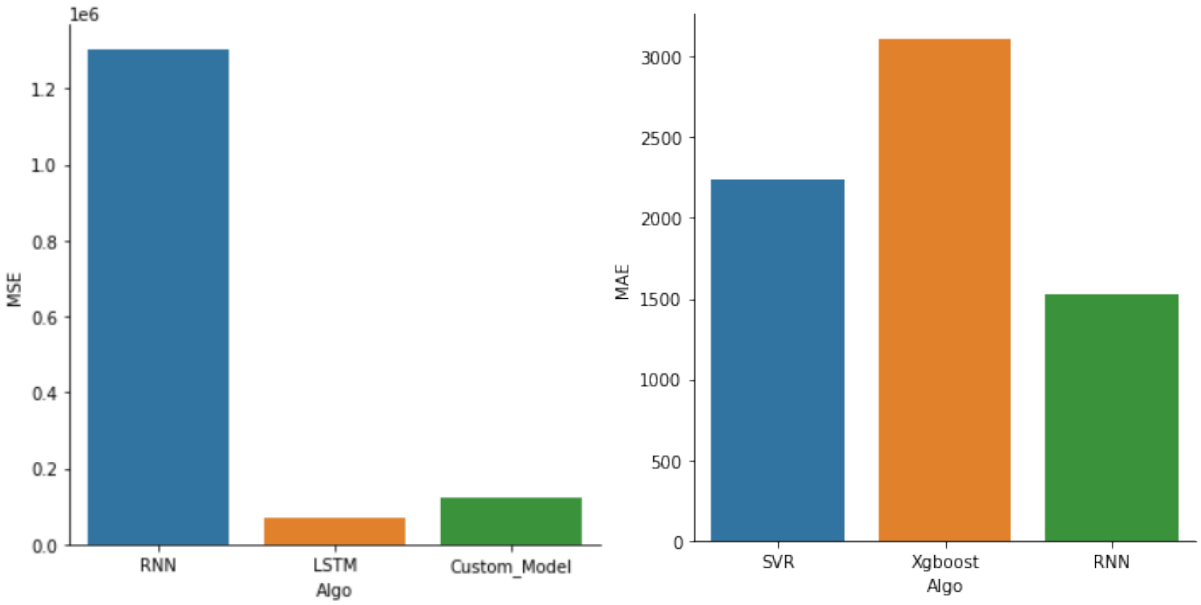


Figure 16: MSE Comparison for Time-series Analysis and MAE Comparison for Sentimental Analysis on Twitter data

price prediction. The obtained MAE for Twitter sentiment analysis is shown in Figure 16.

7 Conclusion and Future Work

Predicting the price of bitcoin is a challenging task and still, an open area of research as bitcoin prices are very much fluctuating and is solely based on the underlying terms of the Cryptocurrency market and its demand. Many researchers have proposed different models and techniques in order to accurately predict the price of bitcoin. In this research, I have attempted to predict the price of bitcoin using time series analysis and Twitter sentiment analysis. For the time-series analysis, the last 10 years of bitcoin price has been collected and prediction is made. For the twitter-based sentiment analysis, the last 10 days of tweets have been collected and the price of bitcoin is predicted based on the sentiments score. For both approaches, I have used 3 different algorithms. After a certain set of experiments, it has been found that for time-series analysis LSTM algorithm outperform in terms of prediction as compared to the RNN and custom model. For the twitter-based sentiments analysis, I have obtained better results using Recurrent Neural Network. Current research has been performed only over the 10 days of tweets data. However, for better prediction, more days of data can be collected using Twitter API. Collecting the data using Twitter API is another challenging task, as the response from Twitter API is slow and it takes many days to collect the data. Twitter mainly informs about the opinion of people about the bitcoin, by extracting the sentiments the financial industries can get an insight about the market behaviour and it can be useful for them to gain better profits.

References

Albariqi, R. and Winarko, E. (2020). Prediction of bitcoin price change using neural networks, pp. 1–4.

Biswas, S., Pawar, M., Badole, S., Galande, N. and Rathod, S. (2021). Cryptocurrency price prediction using neural networks and deep learning, pp. 408–413.

Chandrasekaran, R. (2018). Prediction of litecoin prices using arima and lstm.

Chen, Z., Li, C. and Sun, W. (2019). Bitcoin price prediction using machine learning: An approach to sample dimension engineering, *Journal of Computational and Applied Mathematics* **365**: 112395.

Hassan, M., Hudaefi, F. and Caraka, R. (2021). Mining netizen's opinion on cryptocurrency: sentiment analysis of twitter data, *Studies in Economics and Finance* **38**.

Ho, A., Vatambeti, R. and Ravichandran, S. (2021). Bitcoin price prediction using machine learning and artificial neural network model, *Indian Journal of Science and Technology* **14**: 2300–2308.

I. E. Livieris, N. Kiriakidou, S. S. and Pintelas, P. (2021). An advanced cnn-lstm model for cryptocurrency forecasting, **61**.

Jaquart, P., Dann, D. and Weinhardt, C. (2021). Short-term bitcoin market prediction via machine learning, *The Journal of Finance and Data Science* **7**.

Lahmiri, S. and Bekiros, S. (2021). Deep learning forecasting in cryptocurrency high-frequency trading, *Cognitive Computation* **13**.

Li, T., Chamrajnagar, A., Fong, X., Rizik, N. and Fu, F. (2019). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model, *Frontiers in Physics* **7**: 98.

M. M. Patel, S. Tanwar, R. G. and Kumar, N. (2020). A deep learning-based cryptocurrency price prediction scheme for financial institutions, *Journal of Information Security and Application* **55**.

M. Obthong, N. Tantisantiwong, W. J. and Wills, G. (2020). A survey on machine learning for stock price prediction: Algorithms and techniques, *in Proceedings of the 2nd International Conference on Finance, Economics, Management and IT Business, Prague, Czech Republic*, p. 63–71.

Mahmud, A. and Mohammed, A. (2021). *A Survey on Deep Learning for Time-Series Forecasting*, pp. 365–392.

Mangla, N. (2019). Bitcoin price prediction using machine learning.

Mangla, N. and Rathod, P. (2018). Unstructured data analysis and processing using big data tool-hive and machine learning algorithm-linear regression.

Phaladisailoed, T. and Numnonda, T. (2018). Machine learning models comparison for bitcoin price prediction, **6**: 506–511.

Politis, A., Doka, K. and Koziris, N. (2021). Ether price prediction using advanced deep learning models, pp. 1–3.

Raju, S. M. and Tarif, A. (2020). Real-time prediction of bitcoin price using machine learning techniques and public sentiment analysis.

Reddy, L. S. and Sriramya, D. P. (2020). A research on bitcoin price prediction using machine learning algorithms, p. 5.

Sattarov, O., Jeon, H., Oh, R. and Lee, J. (2020). Forecasting bitcoin price fluctuation by twitter sentiment analysis, pp. 1–4.

Spilak, B. (2018). *Deep Neural Networks for Cryptocurrencies Price prediction*, PhD thesis.

Torres, J., Hadjout, D., Sebaa, A., Martínez-Álvarez, F. and Troncoso, A. (2020). Deep learning for time series forecasting: A survey, *Big Data* **9**.

Umer, M., Awais, M. and Muzammul, M. (2019). Stock market prediction using machine learning(ml)algorithms, *ADCAIJ: ADVANCES IN DISTRIBUTED COMPUTING AND ARTIFICIAL INTELLIGENCE JOURNAL* **8**: 97.

Velankar, S., Valecha, S. and Maji, S. (2018). Bitcoin price prediction using machine learning, pp. 1–1.