Prediction of Bitcoin Prices Using Deep learning and Sentiment Analysis Based on Bitcoin Tweets

MSc Research Project
Data Analytics

Adebayo J Toriola
Student ID: x19192118

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Adebayo J Toriola
Student ID:	x19192118
Programme:	Data Analytics
Year:	2020-2021
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	16/08/2021
Project Title:	Prediction of Bitcoin Prices Using Deep learning and Sentiment Analysis Based on Bitcoin Tweets
Word Count:	6400
Page Count:	21

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:	20th September 2021

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):	

Prediction of Bitcoin Prices Using Deep learning and Sentiment Analysis Based on Bitcoin Tweets

Adebayo J Toriola
x19192118

Abstract

Price fluctuation and volatility of Bitcoin currency in the last few years has caught the attention of many researchers. Many researches are still ongoing to identify and define the sum of exact factors that influences Bitcoin price. Speculation expressed on sentiments has been identified by many researchers as a significant factor of the price volatility. Twitter is one of such platforms where users express their sentiment on issues. However, these data is a time series because it is recorded at regular interval, hence it is necessary to look at the data overtime to predict the next value therefore, in this research VADER sentiment analyser is used to analyse and assigned sentiment scores to Bitcoin-related tweets, the score is merged with historical price. Thereafter, ARIMA and LSTM models were applied to analyze the merged data in order to predict the price movement. Time series analysis is performed on the merged data and it reveals that there is a positive correlation between the Twitter sentiment and the bitcoin price. Finally, the execution time of these two models were evaluated on both local machine and cloud environment and the LSTM model achieve a good RMSE of 0.014 within 1.09 minutes for per minutes data on GPU and and RMSE of 0.018 within 1.29 minutes for for the per hour data.

Keywords:- ARIMA, Bitcoin, Cryptocurrency, LSTM, Sentiment Analysis, VADER, Twitter, Tweets

1 Introduction

The effect of financial markets is significant and important for the smooth operations of any economy, there are different types of financial markets e.g. Stock market, Money market, Forex market, over the counter (OTC) market, Derivatives, Commodities and recently Cryptocurrency market. Bitcoin is the first cryptocurrency introduced in 2009 (Nakamoto; 2008), ever since then hundreds of cryptocurrency has been created and traded across different crypto exchanges by swapping one crypto for another or for fiat currency. The unique value proposition for cryptocurrency market is that its operates on peer-to-peer (P2P) trading and does not have central authority control which contributed to the volatility of the market.

1.1 Background

The adoption of Bitcoin by the public in the last decade is overwhelming due to its propagation from the media thereby resulting to disruption in the financial sector especially within the payment system. Bitcoin has remained dominant of all cryptos. Figure 1 shows the historical price obtain from the dataset used for this research, Bitcoin price rises from 333.75 USD to around 65,000 USD indicating the all time high price between 2015 and 2021. Furthermore, apart from being used as a form of exchange Bitcoin also serves as a form of investment hence the need for this research to predict price of Bitcoin in order to guide investors decisions.

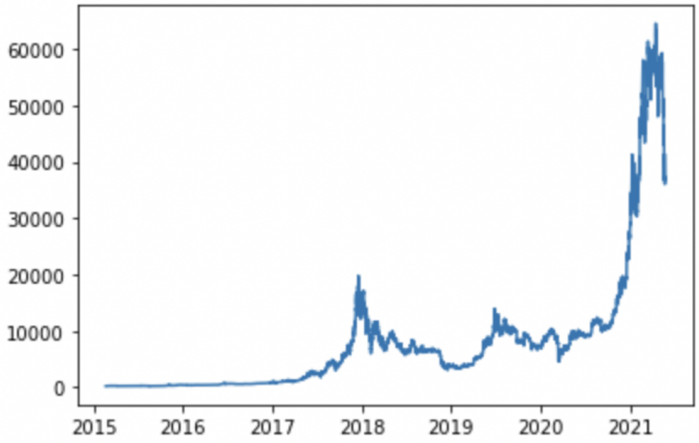


Figure 1: Bitcoin Historical Price

The volatility and surge in Bitcoin price in the last three years and its adoption by large number of users informed and motivate the needs for this research. Though there are quite a number of research on machine learning models on time series prediction of Bitcoin price there is a limited study on how to efficiently predict the price using correlation with factor such as speculations on social media platforms like Facebook, Reddit, Instagram, Blogs, Twitter etc. Large amount of Bitcoin related tweets are generated daily by millions of Twitter users', this data together with the historical price can be used to study the patterns and predict Bitcoin price using statistics, machine and deep learning, sentiment analysis, time series forecast, and natural language processing.

The first section of this report states the background and the research objective, section two covers the related work, we have limited literature in this area however, cryptocurrency and stock market prediction techniques used by researchers were discussed. Section three describes the design and methodology used for this research, in section four the design and specification used in this research was explained. We elaborates on project implementation in section 5 while the evaluation and the interpretation of the model was done in section six, lastly section seven focused on the concluding part and future work in the area.

1.2 Research Questions

Research Question: How accurate can ARIMA and Deep Learning LSTM model predict Bitcoin price using Sentiment Analysis on Bitcoin related tweets

Sub Research Questions: Is there a difference in the execution time when the models is run on Local machine and Cloud environment

1.2.1 Research Purpose

The aim of this research is to determine how accurate the price of Bitcoin combined with Bitcoin tweets using sentiment analysis and LSTM and ARIMAX models can be predicted, in addition we compare the models speed on the local machine and cloud using Google Colab, although Azure cloud environment had initially been proposed but due to resource constrained and readily available Colab has been selected. Financial markets generally including the cryptocurrey market are regarded to be volatile, however cryptocurrency is considered to be more volatile and fragile due to lack of government regulation and central authority, this research is focused on predicting the price based on the volatility factors. Existing studies have shown that there is a correlation between financial market and news or speculation, this correlation will be analysed using LSTM and ARIMA on local machine and on google colab in order to answer the research questions.

In addition, this work can be used to measure the execution time of machine learning models which is the new and core idea introduced to this research. For this research, the models execution time will be measured using the same resources allocation to each of the environments i.e. (Local Machine and Cloud) that is being used for this work. The reason why it is significant to evaluate the response time is because of the high frequency of the data whose price attribute changes per second and decision is based on the price factor.

Finally, the result of this research will offers opportunity for investors to make profits this is why we are combining the bitcoin related tweets which serves as customers reviews, an important and integral part of any business. Prior to the emergence of web forum Business collect customers feedback and opinions through survey, in a similar way that financial markets harnessed public view data through investor survey, the amount of data generated online is big and cannot be analysed through the traditional methods hence the reason why deep learning is being considered for this research (Li et al.; 2016).

1.3 Bitcoin

Over the years, payment and financial services has been dealing with the issues of mistrust and double spending and in order to resolve this issues and possible manipulation, Satoshi Nakamoto (Nakamoto; 2008) in 2008 created Bitcoin using the block-chain technology which stands on cryptography protocols, a peer-to peer networks and decentralised system which is aimed to eliminate third parties' influence. Block-chain technology is used to store information in a distributed digital ledger and verified through cryptography by nodes across the entire computer networks in a block thereby making it difficult to alter the information stored on the network. Since its introduction in 2009 Bitcoin, its study as a time series data has received wide analysis, Bitcoin price varies just as stock but differently because of parameters are responsible for the volatility however the basis is on speculation hence leveraging on machine learning algorithms to forecast Bitcoin price direction Angela and Sun (2020) is ideal.

Generally, financial market performance over a period is viewed as being negative (bearish), positive (bullish) or neutral Yang and Zhang (2015). Harvesting sentiment or opinion is a common phenomenon in financial market however with the advent of internet so much data is being generated thereby categorising sentiment for market prediction has become important task for researchers, therefore any investment decision taken without considering investors opinions could led to bad investment decision.

1.4 Bitcoin Price Prediction Techniques

Bitcoin price prediction is a complicated and challenging task due to dynamics number of variables involved, according to Derakhshan and Beigy (2019); Lim and Yeo (2020) the approach for stock markets analysis is divided into two namely technical and fundamental. The technical analysis is categories into four – statistical, pattern recognition, machine learning, and sentiment analysis, though there is a hybrid approach which combines the technical and fundamental analysis approaches to predict market prices using machine learning it offers great opportunities to investors and researchers (Shah et al.; 2019). Figure 1 shows the classification of stock market prediction techniques.

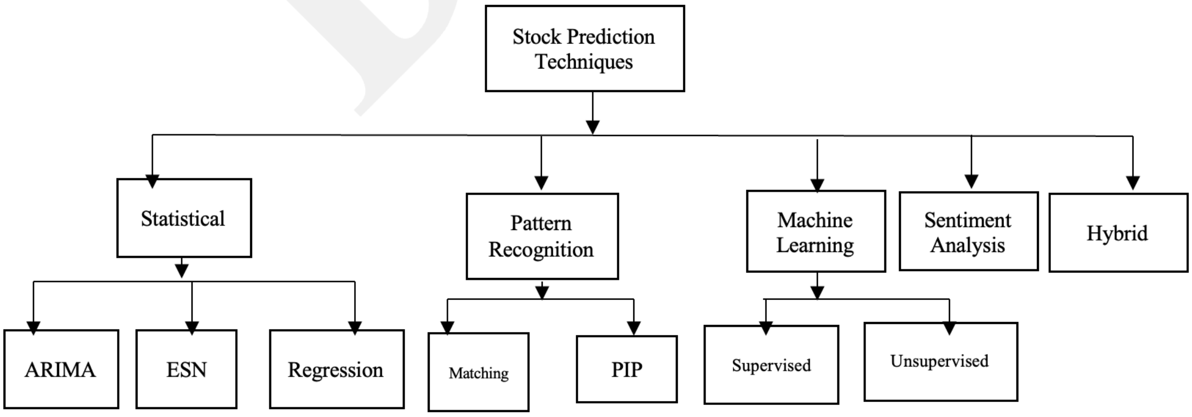


Figure 2: Prediction Techniques

2 Related Work

In this section we discussed related works in financial markets and how the advancement in technology is being used to predict stock market based on user's expression and opinion of the stock.

2.1 Payment System and Cryptocurrency Market

The global financial crisis in 2008 combined with the new advanced technologies in the finance sector paved the way for emergence of digital currencies and a new form of financial institutions, financial institutions is considered to be one of the leading sectors in the implementation of technologies and innovation (Zhu et al.; 2019) and financiers has expressed concerns about this inevitable disruption that is being caused in their sector. In the last one and half decades different electronic means of payments has emerged this include cryptocurrency. However, the popularity of cryptocurrency in today's financial market is great, in the last few years the coin has received more attention from public and government and their agencies, (Liang et al.; 2019) this is why it is important to research and develop models that will reveals the patterns accurately predict the stock price.

According to Liang et al. (2019) cryptocurrency market value is estimated to be in several hundreds of billion dollars, hence its impact on the global economy cannot be overlooked, therefore it is important to predict Bitcoin price and determine the how speculation contributed to it price fluctuation, although it may be difficult to determine the exact number of cryptocurrency users an estimation of the number can be obtained by analysing the number exchange sites' users. The coin has gained so much relevant to the extents that it is now being accepted as a payment method by many businesses, in fact many now sees bitcoin as investment opportunity. The anonymity of bitcoin transactions is a contributing factors to its adoption (Alzahrani and Daim; 2019), bitcoin transactions is consummated as a decentralised payment systems, initiated as a P2P bitcoin has seen to be used for international payment. Due to its lack of regulation bitcoin has not been widely accepted as an alternative to fiat currency because of its volatility which remains high at about 74% despite its awareness level. According to Liang et al. (2019) cryptocurrency market is fragile than the traditional markets, while the foreign exchange market is more stable, the stock market is less stable and the cryptocurrency market is fragile, its fragility is because of the increasing public interest on Bitcoin which is driven by speculation and sentiments without regulation.

Statistical technique is a process of summarising the entire data set and provides meaningful information by conducting analysis to discover insights, this technique is used before the advent of machine learning to analyse time series in stock market, its analysis is based on a chronological collection of observations which assumes linearity, stationarity, and normality. Auto-Regressive Moving Average (ARMA), the Auto-Regressive Integrated Moving Average (ARIMA), the Generalized Autoregressive Conditional Heteroskedastic (GARCH) volatility, and the Smooth Transition Autoregressive (STAR) models are examples of statistical approaches use for predicting time series data. However, ARIMA model is the most common technique used for stock market analysis to forecast future points for a fitted time series data because its combine AR (AutoRegressive) and MA (Moving Average) performances. It convert stationary series from a non-stationary series Zhong and Enke (2017). According to McNally et al. (2018), ARIMA model performed poorly when compared with deep learning models, however highest accuracy achieved for deep learning when optimised RNN was implemented with LSTM was 52%.

Pattern recognition and machine learning are in a way the same but the techniques are applied differently on stock market data, according to (Zhong and Enke; 2017) stock markets patterns are recurring sequences that are found in Open-High-Low-Close (OHLC) indicators which is based on historical data of buy and sell signals. Most technical analysis is based on patterns seen in stock data which can be viewed over time on charts that shows variations in volume and price. The most common pattern recognition techniques are Matching and Perceptual Important Point (PIP), matching technique uses pictographic image to match a stock pattern for price identification while PIP technique is used to store the significant points by reducing the data point dimension (time series reduction).

Machine Learning (ML) is the study of computer algorithms that learn automatically from data, and discover patterns for decisions making with minimal human intervention Pahwa and Agarwal (2019). Machine learning approaches are broadly categories into Supervised Learning, Unsupervised Learning, and Reinforcement Learning, ML is used extensively in the studies of stock markets prediction Vats and Samdani (2019), Madan et al. (2015) achieved an accuracy of 98.7% when machine learning techniques was applied for the prediction of Bitcoin price but this experiment was performed on daily price for a period of five years and achieved accuracy between 50-50% when the the time interval is 10 minutes. (Ballings et al.; 2015) gave different examples of machine learning models used for predicting stocks price includes Random Forest, Support vector machine, Logistic regression, and neural networks. However in recent times, deep learning has become popular and dominant for the analysis of financial market due to the nonlinear, multivariate analysis, data driven, and its ease to generalize characteristics. (Serafini et al.; 2020) analyse the predictive power of network sentiments using deep learning and statistics techniques to predict the future price of Bitcoin, the study shows the significant factor of sentiment in predicting Bitcoin price. Serafini et al. (2020) compare Recurrent Neural Network and Auto-Regressive Integrated Moving Average with eXogenous input (ARIMAX) models for Bitcoin time-series prediction while considering sentiment feature and achieve optimal results with Mean Squared Error (MSE) of 0.4%.

Sentiment analysis is an interesting area for researchers and investors, opinions and emotions expression of customers (internet users) has become popular and an acceptable means for reviewing and analysis of users opinions. Twitter, Facebook, Instagram are example of social media platforms used for collecting sentiment data for analysis. The primary objective of implementing this techniques is to detect and extract emotions through natural language processing technique in spoken or written format. Kaminski (2014) concluded that Bitcoin price depends on sentiment of tweets but for a limited period from November 23rd, 2013 to March 7th 2014 with 160,000 tweets. However to confirm this hypothesis other factors has to be considered for the price prediction, (Kalra and Prasad; 2019) describes sentiment analysis as techniques to analyse text corpuses used in predicting stock markets direction, the text corpuses could be from any news feeds e.g. Twitter or any social media platforms Ahuja et al. (2015). There are series semantics analysis techniques for polarity classification, however Valence Aware Dictionary and sEntiment Reasoner (VADER) has proven to be consistent and outperformed the other techniques in analysing social media text, capable of detecting the intensity and polarity in text by combining lexicon and rule-based sentiment techniques, VADER is designed to address the difficulty in analysing symbols, slang, abbreviations, language, text styles specifically in the social medial platform. Ruths and Pfeffer (2014), VADER requires no training data and does not suffer severely from speed-performance tradeoff. According to Mittal et al. (2019) after applying LSTM, RNN, Polynomial regression the worst results was achieved on tweet sentiment while the tweet volume and Google trends predicted accuracy of 77.01% and 66.66% for the Bitcoin direction respectively, however, the limitation of this research was that news from other social media platform were not taken into account. The hybrid approach is the combination of multiple different approaches to achieve improved and better performance, for example a combination of statistical and machine learning techniques or a hybrid of statistical and pattern recognition approaches. Several hybrid approaches have been applied to stock market prediction for instance, Markowska-Kaczmar and Dziedzic (2008) implemented the machine learning and pattern recognition approaches using a supervised feedforward neural network to detect patterns in stock data and PIP technique by reducing the dimensionality to determine only the important points of the patterns. According to (Wang et al.; 2012) the implementation of a Proposed Hybrid Model (PHM) that combines three models namely are the Exponential Smoothing Model which is a smoothing technique applied on a time series data, ARIMA model and a Backpropagation Neural Network model while relying on each model to predict weekly stock prices. The hybrid model when tested on Shenzhen Integrated Index and DJIA the result shows that the performances of the hybrid model is better than the individual constituent sub-models with directional accuracy of 70.16%.

3 Methodology

This section describes various steps that are follows in order to answer the research question with an objective in mind. Knowledge Discovery in Database (KDD) methodology is used for this research, figure 3 shows the steps involved, Zhong et al. (2019) explained the concept behind knowledge discovery in database as a method of searching, cleaning, transforming, and refining meaningful data from a raw database in order to uncover hidden patterns and interpret the result to get meaningful insights.

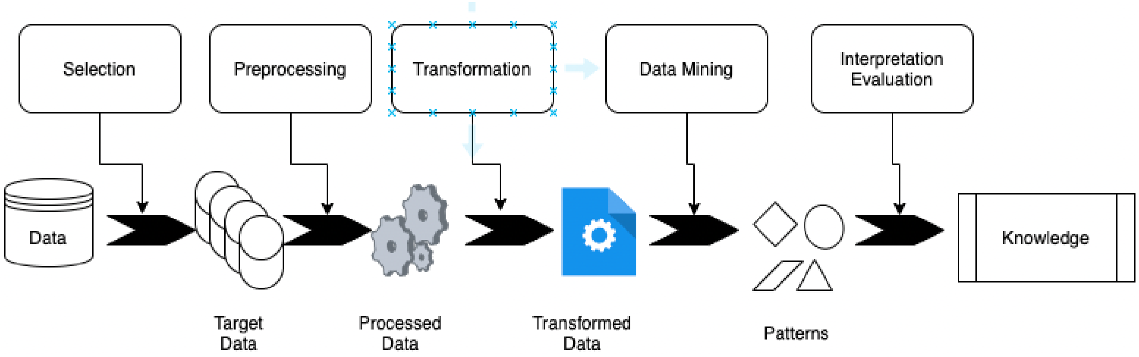


Figure 3: KDD Methodology

3.1 Data Mining

Data mining technique is used for this research because it helps to select the models and used for categorization and prediction of data patterns. This process enables us to make in-depth inquiries into the historical price and tweets data thereby given us the opportunity to make sense of the data. The process is used to dig into enormous set of data to produce meaningful data, used for studying and predicting future trends, the process is a computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. This technique has proven to be successful in the statistical world, its major elements consists of Extract, Transform and Load (ETL), storage and management of the data, data accessibility and its presentation in a useful format. Zengin et al. (2011).

3.2 Data Collection

This is the phase concerns with the acquisition of the historical price and Tweets data, based on preliminary knowledge of the domain, this technique is performed on the historical dataset of bitcoin price and tweets data source. The tweets data is obtained from Kaggle [1] and its consist of 800,165 records with a period range between February 10, 2021 and June 20th, 2021. For this research, we are analysing a sample of the tweets because it is within a limited time, the data is collected using keywords such as $Bitcoin, #BTC, #bitcoin, #cryptocurrency. The users opinion about bitcoin is expressed through

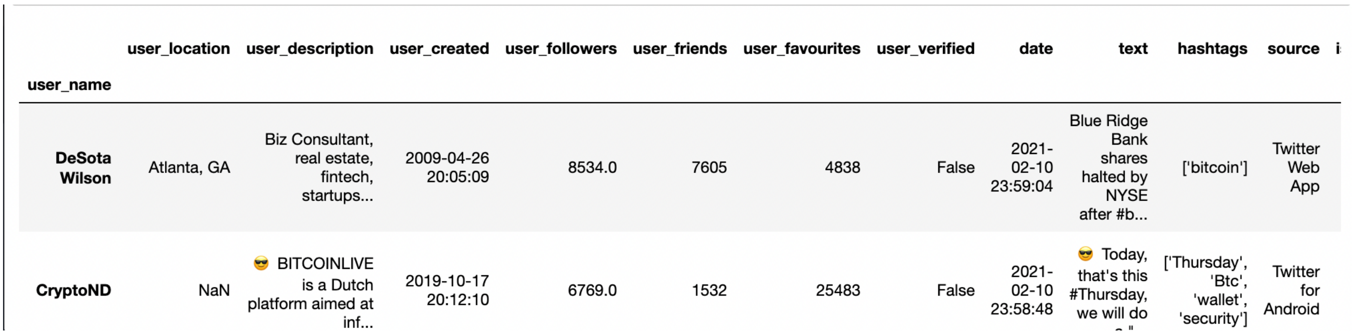
[1] https://www.kaggle.com/kaushiksuresh147/bitcoin-tweets

the different platforms, however we decided to analyse data collected from Twitter platform, this forms a significant part of our research Elbagir and Yang (2019). In similar manner the bitcoin hourly and per minutes dataset is collected from cryptodataworld [2], cryptodataworld is repository of crptocurrency coin historical data collected from different exchanges. The period of the historical data collected is from May 2015 to July 2021.

3.3 Data Preprocessing

The cleaning and removal of irrelevant information from our data is performed in this phase, this techniques is deals with the removal of noise and outliers. We identify null values present in our dataset and replaced them. The raw data is noisy with irrelevant and inconsistent data, this noise is handled by filling the missing values and by removing the duplicate records and the outliers furthermore, the tweets were filtered to remove non English tweets and duplicates. Data preprocessing step is performed to facilitate the training and testing process for transforming and scaling of the entire dataset, this steps is required to save the processing time and retain the quality of the data for data mining. Data pre-processing is necessary because it prepares the data in a way that is meaningful for the subsequent detailed analysis Al-Jabery et al. (2019).

Shown below is the screenshot of tweets sample collected prior to preprocessing.



	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	i
user_name												
DeSota Wilson	Atlanta, GA	Biz Consultant, real estate, fintech, startups...	2009-04-26 20:05:09	8534.0	7605	4838	False	2021-02-10 23:59:04	Blue Ridge Bank shares halted by NYSE after #b...	['bitcoin']	Twitter Web App	
CryptoND	NaN	😎 BITCOINLIVE is a Dutch platform aimed at inf...	2019-10-17 20:12:10	6769.0	1532	25483	False	2021-02-10 23:58:48	😎 Today, that's this #Thursday, we will do	['Thursday', 'Btc', 'wallet', 'security']	Twitter for Android	

Figure 4: Bitcoin Plot

The tweets dataset consists of data like emoticons, URL, misspelling, unwanted punctuation, and pictures which are not required for our analysis hence require series of preprocessing steps, the tweets data is filtered in three stages: Special Characters Removal, Stopwords removal, Tokenization of the text data contain in the tweet.

Special Characters Removal: Regular Expression (Regex) matching is used to remove special character, URL, hashtags #, and @ which is used to address other users. For example, #Bitcoin is replaced with Bitcoin and @Elonmusks is replaced with the USER name.

Stopword Removal: Using the text sentiment analyser, words that do not express any emotion in the tweets text are removed after splitting the tweets, examples of such words are the, is, with, a etc. which are all removed from the the list of words.

Tokenization: This is the process of splitting a paragraph, sentence, phrase of the text data into individual words or terms called tokens. The token provides understanding and give background to developing the model for the natural language processing, it provides

[2] https://www.cryptodatadownload.com/data/bitstamp/

interpretation to the analyzed text. After splitting the tweets data based on the space into individual words, a list of individual words is the then formed from the text.

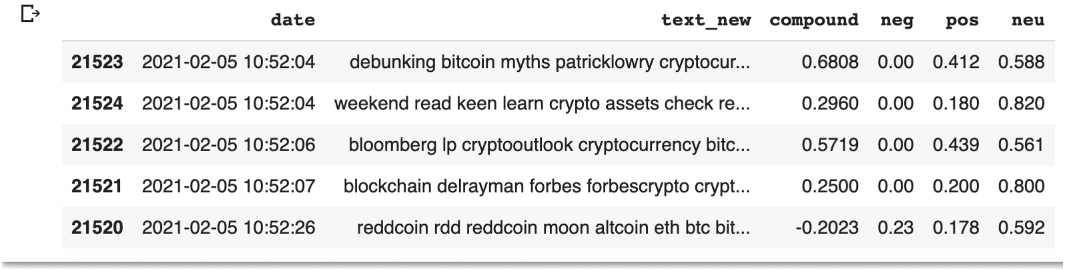
The steps used in preprocessing the text data includes:

- Tokenization
- Stop words Removal
- Filtering
- Strip Non-ASCII characters
- Tokenization, stop-words removal, stemming
- Remove URLs of type (http://)
- Remove the Emoticons

3.4 Data Transformation

This is the stage that the data is transformed from its raw form after the cleaning process for mining purposes, the importance of this process is to make the data to be better organized. A well transformed brings about improved data quality and guide against possible null values, duplicate records, incorrect indexing. We applied NLTK's VADER analyzer as the scoring method of our data, the tweets data which is originally in text format is transformed to numeric using the Valence Aware Dictionary for Sentiment Reasoning (VADER) model. VADER is used to measure the attitude, sentiments, evaluations and emotions of the user based on the intensity of the text, it is used to assign the polarity using the NLTK package. The sentiment analysis used for this research (VADER) is intelligent enough to understand the basic context of the text data Elbagir and Yang (2019).

Extract below is tweet sample and the probability of polarity assigned by VADER analyser and word cloud below shows the frequent words contains in the text.



	date	text_new	compound	neg	pos	neu
21523	2021-02-05 10:52:04	debunking bitcoin myths patricklowry cryptocur...	0.6808	0.00	0.412	0.588
21524	2021-02-05 10:52:04	weekend read keen learn crypto assets check re...	0.2960	0.00	0.180	0.820
21522	2021-02-05 10:52:06	bloomberg lp cryptooutlook cryptocurrency bitc...	0.5719	0.00	0.439	0.561
21521	2021-02-05 10:52:07	blockchain delrayman forbes forbescrypto crypt...	0.2500	0.00	0.200	0.800
21520	2021-02-05 10:52:26	reddcoin rdd reddcoin moon altcoin eth btc bit...	-0.2023	0.23	0.178	0.592

Figure 5: Tweets Polarity

3.5 Feature Engineering

Feature engineering is the process of extracting the most valuable variables from the raw data for analysis using the domain knowledge, and helps to reduce the number of variables in the data set, the process guides in achieving good prediction. Most predictive models used features, it is used to get the enough out of the data for the models to work by influencing the predictive models in other words the better the features the better the results of our prediction.

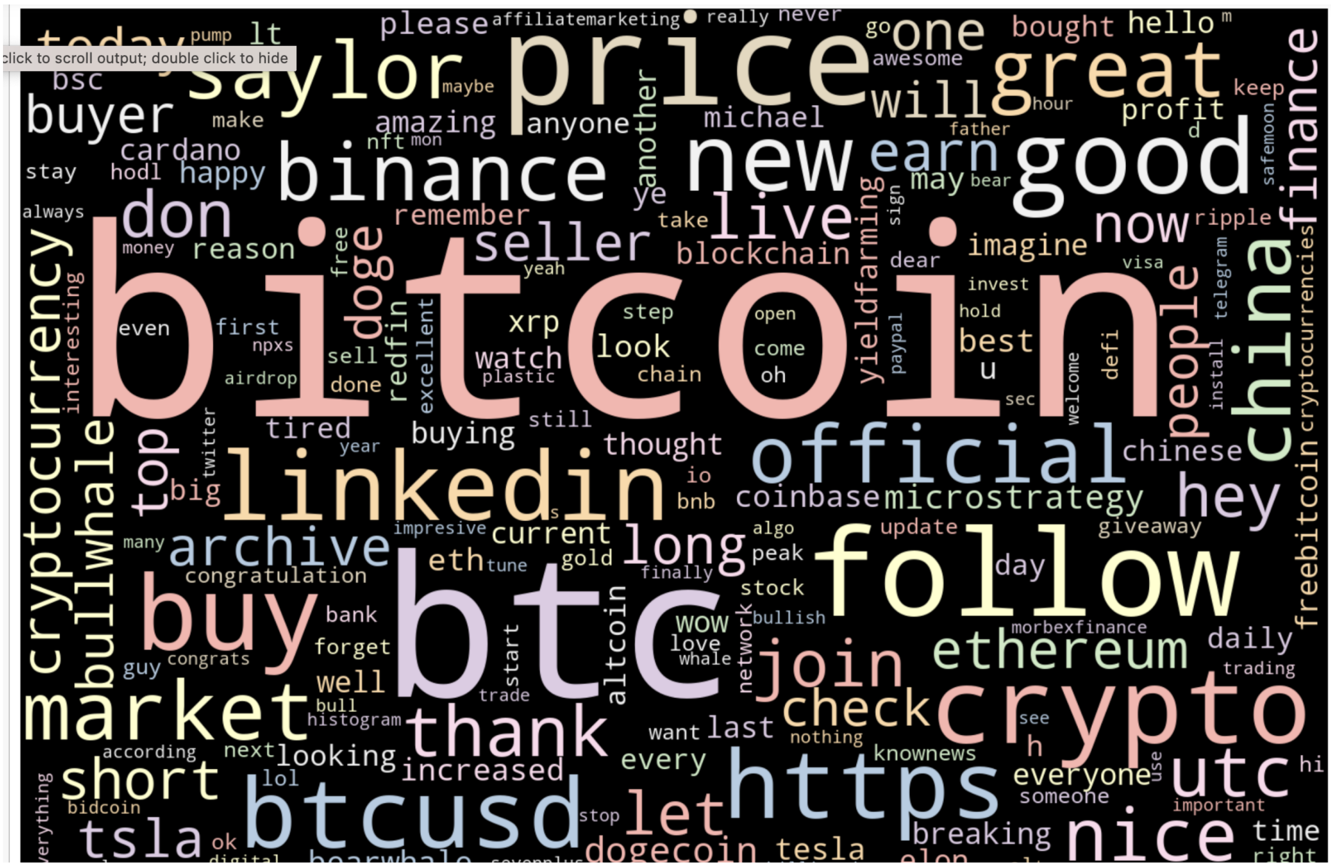


Figure 6: Word Cloud

3.6 Evaluation

This process is used to determine the performance of our models and how it could be improved to achieve a better result, it provides a systematic approach to this research and how well to achieves our goals. The result of the models used for the work is evaluated and interpreted with a focus on the understanding and usefulness of the induced model Maimon and Rokach (2005). The resources of the machines upon which the models are run noted and the time it takes to execute each model are noted. The response time or speed of these for models (ARIMA and RNN(LSTM)) are measured besides observing how well the changes in Bitcoin price (rises and falls) correlates with opinion expressed on tweets. In this research, the data is split into train and test in ratio 80:20 respectively, the model is then evaluated using RMSE.

3.7 Knowledge

This is last phase of this research where the novel on this research is explained, it is helps to identify the potential usefulness of this work, it guides in understanding the relationship within our dataset in order to make important decisions. The success of the entire KDD methodology is determined by the knowledge and insight gained from the research, this research helps to gain interesting insight on how the system environment performs, incorporating the knowledge into another system shows a generic answer is obtained Hutter et al. (2021).

3.8 Exploratory Data Analysis

We use exploratory data analysis (EDA) to analyse our data before making any assumptions, with the EDA obvious error within our data is identify, it helps to detect outliers or events that are unknown in our data and find interesting relations in our variables. The chart below show the distribution of the tweet polarity for us to understand the distribution within our data.

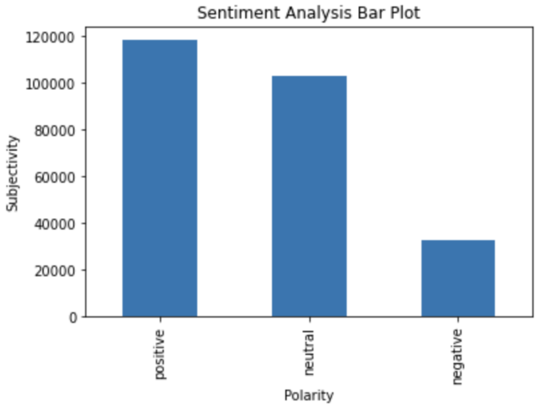


Figure 7: Bar Chart of Tweet polarity

4 Design Specification

In fulfilling the objective of this research project the design specification is significant because it helps to proceed from the abstract to the concrete solution. This description of the design specification used in this project is stated below, the workflow framework is divided into three layers namely the client, business logic and the data layer shown in figure 4:

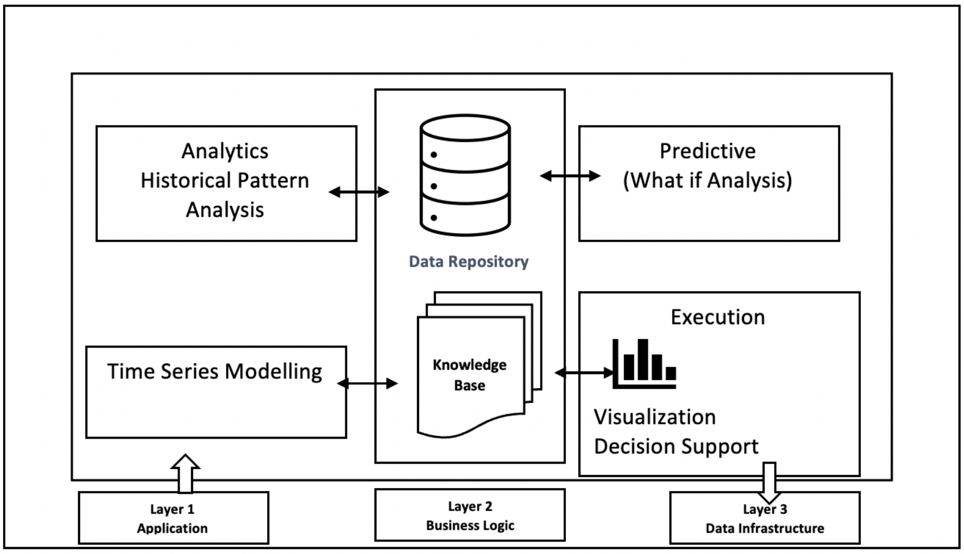


Figure 8: Design Framework

Layer 1 (Application) - The historical price and tweets data is collected from the data source. The data collected is a time series data which is suitable to apply statistical and machine learning techniques to analyse and predict over a period of time. Data preprocessing, features extraction, features engineering and data transformation are performed in this layer. We applied time series models, Recurrent Neural Network using LSTM, ARIMA as the based statistical and ARIMAX including sentiment analysis to analyse the text in the tweet. These models are used to analyse and study the patterns each dataset and merged in order to predict and evaluate the model performance.

Layer 2 (Business logic) - In the second layer of the design, this is the point that we run our models on the data, for the purpsoe of prediction and measuring of the execution time of the models ARIMA, ARIMAX, and LSTM models. VADER analyser is used to extract opinion from the tweet data using the Natural Language Processing (NLP) toolkit to analyse the text data. The sentiment analysis helps to understand opinion of users through information extraction from a bunch of words, a rule based approach known as sentiment lexicon is used to classify words into positive, neutral and negative value to identify subjective, polarity score or the opinion. The lexicon-based approach is used to determine the semantic orientation of the text document using tokenization and stemming approach.

Layer 3 (Data Infrastructure) - This is the layer that produce the graphic representation of the data based on the results of the models, using series of python this layer helps in communicating and support decision making process. The diagram below id the design framework and its consists of Time series modeling, Analytics, Predictive and Execution and a unified data repository.

Time series data modeling is used for monitoring and sorting by time structured and unstructured of data collected, this is then stored in the repository.

Analytics module leverage on statistical and machine learning models to analyse the historical data while monitoring the run-time resource usage such as CPU utilization. In this module cross validation is used to evaluate accuracy of predictions of the data collected while it is split into training and testing.

Prediction helps to make decisions on resource configuration changes. Emphasis is placed on what is happening now and next at a particular time interval in order to make decision, with higher precision the models learned incrementally with the new data.

Execution This module in the design focuses on the data visualization and decision support in order to adjust resources while keeping the capacity and performance in check with minimal resource consumption.

Data repository interacts with other module and provides a uniform data store for collecting data and generating patterns and prediction from the models.

Figure 5 represent the embedded workflow that is operated with design specification.

5 Implementation

This research work is implemented on Jupyter notebook using Python programming language, a general-purpose programming language used by Data Scientist. The methodology described in section 3 was followed, we initiated various data preprocessing techniques described in section 3.3. The process flow for this research is shown in figure 10, we started with the data selection and carried out series of data preprocessing. We used VADER Sentiment analyser on the text data to get the polarity and intensity of each text

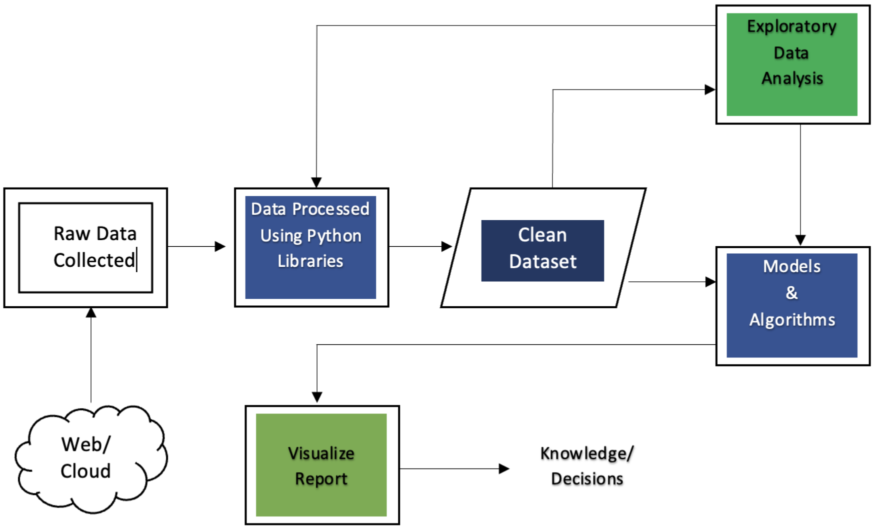
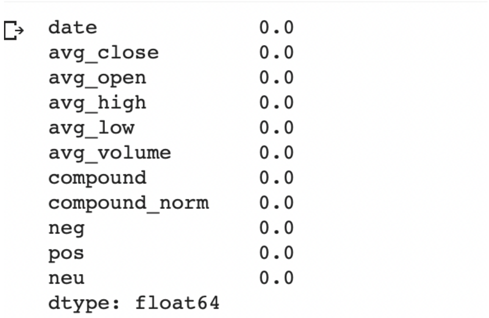


Figure 9: Workflow Framework

and later re-sampled the text data per hour and per minute. Thereafter, we aggregated and merged the dataset, we perform an EDA on the data to gather insights. Finally, we run the two models on the data and look for a future trend of the price using statsmodels.tsa library and using rmse library on the LSTM model, we then measured the execution time of each model from different environment based CPU and GPU resource allocation.

This phase of the research plays a significant part in achieving the research outcomes, (Bhardwaj et al.; 2012), poor implementation leads to poor outcomes, the process involves systematic running of the models. before implementing our models a check on the data is carried out to confirm that our data does not contain any missing values.



```
date             0.0
avg_close        0.0
avg_open         0.0
avg_high         0.0
avg_low          0.0
avg_volume       0.0
compound         0.0
compound_norm    0.0
neg              0.0
pos              0.0
neu              0.0
dtype: float64
```

Figure 10: Checking NA values

5.1 Data Preparation

- The tweet data is loaded using the pandas library and the date column is passed as the index column. The data is consisted of 800,165 rows with 13 columns, during the preprocessing phase columns that are not required for our analysis were dropped except for the date and text columns. The datatype of the date column which is in object format is converted to date using the datetime library, regular expression was used to clean the tweets. VADER(Valence Aware Dictionary for Sentiment

Reasoning) - sensitive to both polarity (positive/negative) and intensity(strength) of emotion is used to analyse the text, VADER package gives a score of negative, positive, neutral sentiment and compound score which ranges from -1 to 1, since our bitcoin data is in every one hour and one minutes, the tweets data was re-sample and aggregated into one minute and one hour data which led to data reduction.

- Bitcoin market operates 24/7, the hourly and per minutes historical price of the data are collected from three exchanges, price data from the exchanges gives a price index called Bitcoin Price Index which represents an average of bitcoin prices from the major global exchanges,the data from these exchanges contained 8 columns; UnixTimeStamp, Date, Symbol, Open, High, Low and Close and Volume. The data collected from the exchanges were merged using the pandas dataframe. The average value of the columns were taken using the mean function, we converted the date column to datetime format. Then we extracted the average closing price but dropping other columns.

- Tweet and Bitcoin data were merged to form a single data frame which served as our final dataset that the models runs upon . The merged is possible using the outer join based on the date column. The new dataframe formed consists of three column namely; date, average closing price, polarity (positive, negative, neutral and compound).

- Time series analysis is performed on the merged data to check whether there is a correlation between the dataset and check for lag between the variables, the bitcoin prices data is transformed to a stationary to run the correlation, change on per minutes and hourly basis is used to observe the change in price.

5.2 Data Mining Techniques

Due to the stochastic, high complexity and non linearity nature of the tweets and bitcoin historical price its prediction has remained a challenging task hence it required complex technologies. The ARIMA and Deep Learning are trending, ARIMA has good prediction accuracy and its flexibility for different types of time series data but cannot simulate nonlinear structure adequately. While Deep learning has the advantage of non-linear approximation however its treats both the inputs and the outputs data independently which is seen as shortcomings of the model, LSTM has been chosen as the deep learning model used for this research and the reason is because LSTM can be used to solve nonlinear problems (Wu et al.; 2021).

The diagram below represent the implementation workflow used for the research.

5.3 Sentiment Analysis: VADER

We applied sentiment analysis in order to find the opinions and subjectivity in the texts and to achieve high accuracy in predicting trends of the tweets based on the polarity classification. For the individual tweet analysis, VADER provides compound score together with the polarity score for each tweet which is between -1 to 1. Therefore, in order to classify each tweets as positive or negative its scores is compared to the compound sentiment score, any tweets that does not falls to either

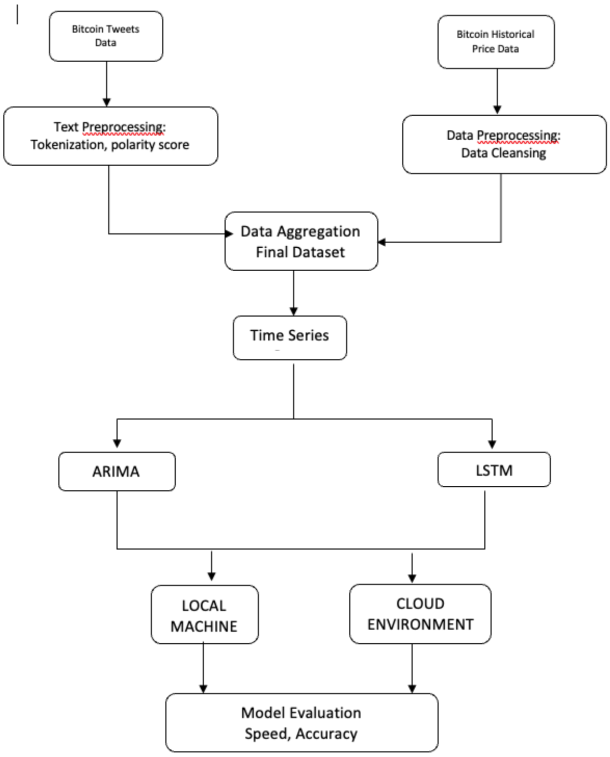

Bitcoin Tweets Data
Bitcoin Historical Price Data
Text Preprocessing: Tokenization, polarity score
Data Preprocessing: Data Cleansing
Data Aggregation Final Dataset
Time Series
ARIMA
LSTM
LOCAL MACHINE
CLOUD ENVIRONMENT
Model Evaluation Speed, Accuracy

Figure 11: Implementation Flow

of this class is left unclassified. Lastly, the individual sentiment scores are grouped at a minute and one hour interval, then the mean of the score is taken as the tweets score for the minute and for the hour.

5.4 ARIMA

ARIMA model is stochastic and is confirmed to be flexible and strong its consists three main step: identification, estimation parameters, and forecasting , it has been used in many fields, and it is a sequential model trained to forecast future data points used to capture complex patterns like the historical price, it captures the white noise and used in recording observations of lagged terms. ARIMA combines autoregressive (AR), Inteegration (I) and Moving Average (MA) functionalities which corresponds to p, d, q parameters (Wu et al.; 2021).

5.5 LSTM

This is an artificial recurrent neural network (RNN)commonly used in deep learning, the traditional RNN suffers gradient disappearance making it difficult to process long-term correlations. To resolve this issue gradient disappearance Long Short Term Memory (LSTM) which is a type of time cycle network specially used to solve RNN issues used for this research. LSTM unit is made up of input, output and a forget gate, the three gate is used to control the data the goes in and out of the unit furthermore these gates are used to remember values over arbitrary time intervals. LSTM is typically used for classifying and making predictions based on time series data in order to handle the possible lags in time series. For this project we add 2 minutes and 24 hours lags hereby creating a number of observations. (Wu et al.; 2021).

5.6 System Configuration

This is the phase of the system which focuses on the allocation of resources available to the containers, a good system configuration is important to the performance of the machine learning system, for this research we have set the configuration in order to increase the efficiency based on the number of records that is processed. The research is carried out using different processes on machine CPU local machine, Cloud CPU and GPU using Google Colab based on the following configuration:

- Local CPU Processor Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz 2.59 GHz Installed RAM 16.0 GB 64-bit operating system, x64-based processor.
- Cloud CPU - 1xsingle core hyper threaded Xeon Processors @2.3Ghz i.e(1 core, 2 threads) 16.0 GB RAM.
- Cloud GPU – 1xTesla K80 , compute 3.7, having 2496 CUDA cores , 12GB GDDR5 VRAM.

6 Evaluation

This section explained the result achieved from the two models used for the research. The loss and accuracy of the training and validation data is calculated per epoch for the LSTM model. The root mean square error (RMSE) and mean absolute error (MAE) which represent the standard deviation and the average difference between the predicted and actual value respectively is used to measure the performance of our models, it step is necessary to determine the models reliability and efficiency. The formula are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - xi)^2}$$

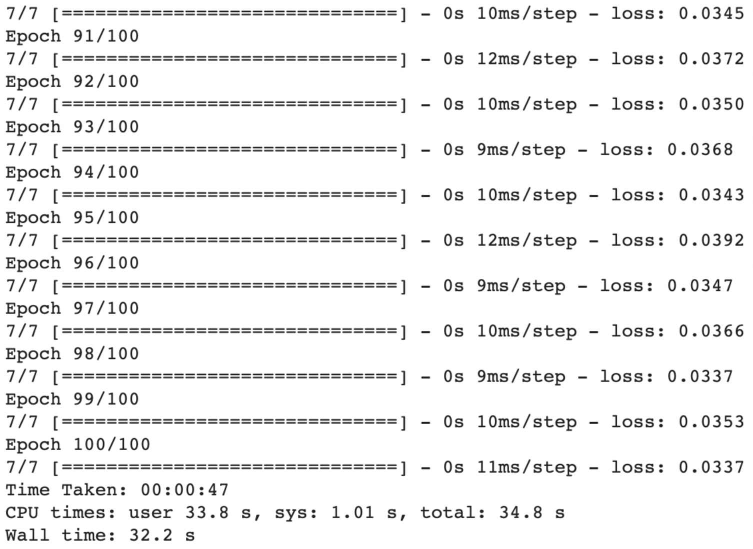
i = variable

Xi = Actual Observation Time Series

yi = Estimated Time Series

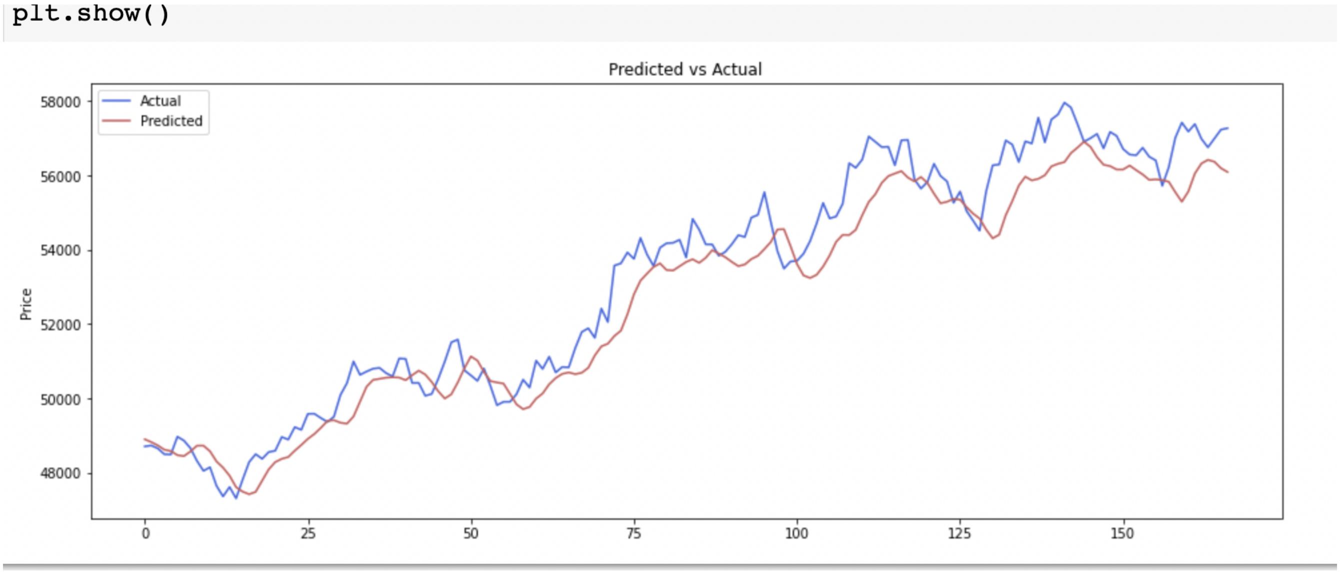
N = Number of data points

Due to the fact that the data being analysed is high frequency data ARIMA model is not suitable for the price prediction. Hence LSTM is best choice and we compared the RMSE and the execution of this model on local and the cloud using Google Colab.



```
7/7 [==============================] - 0s 10ms/step - loss: 0.0345
Epoch 91/100
7/7 [==============================] - 0s 12ms/step - loss: 0.0372
Epoch 92/100
7/7 [==============================] - 0s 10ms/step - loss: 0.0350
Epoch 93/100
7/7 [==============================] - 0s 9ms/step - loss: 0.0368
Epoch 94/100
7/7 [==============================] - 0s 10ms/step - loss: 0.0343
Epoch 95/100
7/7 [==============================] - 0s 12ms/step - loss: 0.0392
Epoch 96/100
7/7 [==============================] - 0s 9ms/step - loss: 0.0347
Epoch 97/100
7/7 [==============================] - 0s 10ms/step - loss: 0.0366
Epoch 98/100
7/7 [==============================] - 0s 9ms/step - loss: 0.0337
Epoch 99/100
7/7 [==============================] - 0s 10ms/step - loss: 0.0353
Epoch 100/100
7/7 [==============================] - 0s 11ms/step - loss: 0.0337
Time Taken: 00:00:47
CPU times: user 33.8 s, sys: 1.01 s, total: 34.8 s
Wall time: 32.2 s
```

Figure 12: earch for Minimal AIC



```
plt.show()
```

Figure 13: LSTM Price Prediction

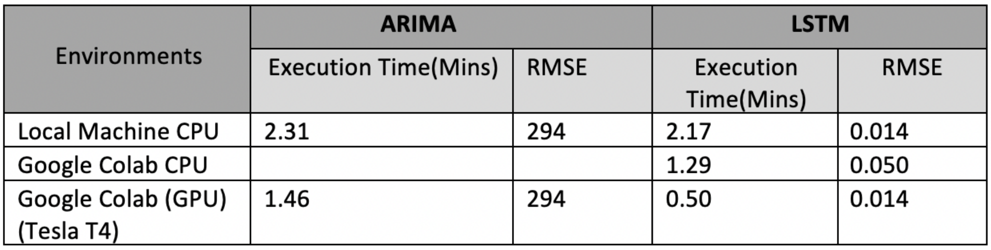
6.1 Experimental Setup

We run the first experiment on Google Colab on a CPU instances with xsingle core hyper threaded Xeon Processors @2.3Ghz i.e(1 core, 2threads) 16.0 GB RAM.

Minute Data: For the data of every minute, the RMSE is 0.014 and the execution time is 2.31 minutes and 2.17 minutes for the SARIMASX and LSTM model respectively on the CPU.

Hourly Data: For the hourly data, the RMSE is 0.022 and the execution time is 2.25 minutes and 1.09 minutes for the SARIMAX and LSTM model respectively on the GPU.

The result of the experiment is shown below.



Environments	ARIMA		LSTM	
	Execution Time(Mins)	RMSE	Execution Time(Mins)	RMSE
Local Machine CPU	2.31	294	2.17	0.014
Google Colab CPU			1.29	0.050
Google Colab (GPU) (Tesla T4)	1.46	294	0.50	0.014

Figure 14: Experiment Result

7 Conclusion and Future Work

This research objective is to study how accurate ARIMA and LSTM model can predict the price of Bitcoin with its related tweets using sentiment analysis, also the research is aimed to measure the execution time of the the models on the local machine and cloud environment using google colab. VADER sentiment analyser was used to measure the polarity and intensity of the the tweets. Alongside the polarity the VADER also generates a compound score which correlates with polarity scores. We merged the compound score with the closing price to predict future price at the same time measure the execution period of each model. However, because the data is high frequency data ARIMA model is not suited for the analysis, we used SARIMAX to forecast and the result of RMSE was high, However the execution time of running the model on GPU was faster than CPU on both local and cloud environment. As expected the LSTM model outperformed the SARIMAX with RMSE score of 0.014 at significant execution time on GPU resource 1.46 minutes on CPU and 50 seconds on GPU. The hyperparameter optimization used for the GPU is almost 5 times faster than CPU, the configuration of the platform in analysing and predicting time series data is significant. One major limitation is that the tweets data that is analysed is a sample data taken from a population of tweets which does not not represent the true value of our data hence its resulting into inaccurate price prediction. Another challenge is the forecasting of the data at granularity, it makes it difficult to identify a recognisable pattern or trend because of the frequency of the data. To further improved upon this research, two or more deep learning models can be used to analysed the data on GPU cloud environment to report for an improve execution time. Also this research can be improved upon by collecting tweets from other languages apart from English language for analysis.

7.1 Acknowledgement

My profound gratitude to my supervisor Jorge Basilio for his guidance and knowledge transfer during the meeting. I would also like to thank my wife for her support throughout the time of this research.

References

Ahuja, R., Rastogi, H., Choudhuri, A. and Garg, B. (2015). Stock market forecast using sentiment analysis, *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1008–1010.

Al-Jabery, K., Obafemi-Ajayi, T., Olbricht, G. and Wunsch, D. (2019). *Computational Learning Approaches to Data Analytics in Biomedical Applications*, Academic Press.

Alzahrani, S. and Daim, T. U. (2019). Analysis of the cryptocurrency adoption decision: Literature review, *2019 Portland International Conference on Management of Engineering and Technology (PICMET)*, pp. 1–11.

Angela, O. and Sun, Y. (2020). Factors affecting cryptocurrency prices: Evidence from ethereum, *2020 International Conference on Information Management and Technology (ICIMTech)*, pp. 318–323.

Ballings, M., Van den Poel, D., Hespeels, N. and Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction, *Expert systems with Applications* **42**(20): 7046–7056.

Bhardwaj, A., Sharma, A. and Shrivastava, V. (2012). Data mining techniques and their implementation in blood bank sector–a review, *International Journal of Engineering Research and Applications (IJERA)* **2**(4): 1303–1309.

Derakhshan, A. and Beigy, H. (2019). Sentiment analysis on stock social media for stock price movement prediction, *Engineering Applications of Artificial Intelligence* **85**: 569–578.
URL: *https://www.sciencedirect.com/science/article/pii/S0952197619301666*

Elbagir, S. and Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and vader sentiment, *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 122, p. 16.

Hutter, F., Dong, Y., Ifrim, G., Mladenic, D., Saunders, C. and Van Hoecke, S. (2021). *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part V.. Lecture Notes in Artificial Intelligence*, Vol. 12457, Springer Nature.

Kalra, S. and Prasad, J. S. (2019). Efficacy of news sentiment for stock market prediction, *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 491–496.

Kaminski, J. (2014). Nowcasting the bitcoin market with twitter signals, *arXiv preprint arXiv:1406.7577* .

Li, Q., Shah, S., Fang, R., Nourbakhsh, A. and Liu, X. (2016). Tweet sentiment analysis by incorporating sentiment-specific word embedding and weighted text features, *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 568–571.

Liang, J., Li, L., Chen, W. and Zeng, D. (2019). Towards an understanding of cryptocurrency: A comparative analysis of cryptocurrency, foreign exchange, and stock, *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 137–139.

Lim, M. and Yeo, C. K. (2020). Harvesting social media sentiments for stock index prediction, *2020 IEEE 17th Annual Consumer Communications Networking Conference (CCNC)*, pp. 1–4.

Madan, I., Saluja, S. and Zhao, A. (2015). Automated bitcoin trading via machine learning algorithms, 2015, *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep* .

Maimon, O. and Rokach, L. (2005). Introduction to knowledge discovery in databases, *Data mining and knowledge discovery handbook*, Springer, pp. 1–17.

Markowska-Kaczmar, U. and Dziedzic, M. (2008). Discovery of technical analysis patterns, *2008 International Multiconference on Computer Science and Information Technology*, pp. 195–200.

McNally, S., Roche, J. and Caton, S. (2018). Predicting the price of bitcoin using machine learning, *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*, IEEE, pp. 339–343.

Mittal, A., Dhiman, V., Singh, A. and Prakash, C. (2019). Short-term bitcoin price fluctuation prediction using social media and web search data, *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pp. 1–6.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system, *Decentralized Business Review* p. 21260.

Pahwa, K. and Agarwal, N. (2019). Stock market analysis using supervised machine learning, *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 197–200.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior, *Science* **346**(6213): 1063–1064.

Serafini, G., Yi, P., Zhang, Q., Brambilla, M., Wang, J., Hu, Y. and Li, B. (2020). Sentiment-driven price prediction of the bitcoin based on statistical and deep learning approaches, *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Shah, D., Isah, H. and Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques, *International Journal of Financial Studies* **7**(2).
URL: *https://www.mdpi.com/2227-7072/7/2/26*

Vats, P. and Samdani, K. (2019). Study on machine learning techniques in financial markets, *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–5.

Wang, J.-J., Wang, J.-Z., Zhang, Z.-G. and Guo, S.-P. (2012). Stock index forecasting based on a hybrid model, *Omega* **40**(6): 758–766. Special Issue on Forecasting in Management Science.
URL: *https://www.sciencedirect.com/science/article/pii/S0305048311001435*

Wu, X., Zhou, J., Yu, H., Liu, D., Xie, K., Chen, Y., Hu, J., Sun, H. and Xing, F. (2021). The development of a hybrid wavelet-arima-lstm model for precipitation amounts and drought analysis, *Atmosphere* **12**(1): 74.

Yang, F. and Zhang, J. (2015). Bullish-bearish-based neural network stock trading decision supportano its application in hong kong stock market, *2015 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, pp. 179–184.

Zengin, K., Esgi, N., Erginer, E. and Aksoy, M. E. (2011). A sample study on applying data mining research techniques in educational science: Developing a more meaning of data, *Procedia-Social and Behavioral Sciences* **15**: 4028–4032.

Zhong, L., Duan, X., Wang, Y., Chen, J., Liu, J. and Wang, X. (2019). eroc: A distributed blockchain system with fast consensus, *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pp. 205–214.

Zhong, X. and Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction, *Expert Systems with Applications* **67**: 126–139.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417416305115*

Zhu, X., Lu, X. and Wu, S. (2019). Research on pricing mechanism and security of digital currency, *2019 International Conference on Networking and Network Applications (NaNA)*, pp. 325–331.