

# CRYPTOCURRENCY PRICE PREDICTION USING SOCIAL MEDIA DATA MINING AND EPIDEMIC MODELING

by

Bushra Safa

17101179

Sanjida Noushin Eva

17101180

Sania Hossain

17101512

A.K.M Salauddin

17301201

Lubaba Fakruddin Upoma

17305003

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering  
Brac University  
October 2021

© 2021. Brac University  
All rights reserved.

## Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I/We have acknowledged all main sources of help.

### Students Full Name & Signature:



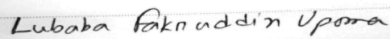
---

Sanjida Noushin Eva  
17101180



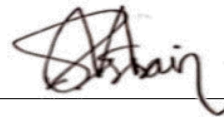
---

Bushra Safa  
17101179



---

Lubaba Fakruddin Upoma  
17305003



---

Sania Hossain  
17101512



---

A.K.M. Salauddin  
17301201

# Approval

The thesis/project titled Cryptocurrency Price Prediction using Social media Data mining and Epidemic modeling. submitted by

1. Bushra Safa (17101179)
2. Sanjida Noushin Eva (17101180)
3. Sania Hossain (17101512)
4. A.K.M Salauddin (17301201)
5. Lubaba Fakruddin Upoma (17305003)

of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 02, 2021.

## Examining Committee:

Supervisor:  
(Member)



---

Moin Mostakim  
Lecturer  
Department of Computer Science and Engineering  
BRAC University

Program Coordinator:  
(Member)

---

Md. Golam Rabiul Alam, PhD  
Associate Professor  
Department of Computer Science and Engineering  
BRAC University

Head of Department:  
(Chair)

---

Sadia Hamid Kazi  
Professor and Dean  
School of Data and Science  
Department of Computer Science and Engineering  
BRAC University

## **Abstract**

With the introduction of blockchain technology in recent years, there has been a massive increase in the use of Cryptocurrencies. In any event, due of the market's unpredictable behavior and excessive cost volatility, Cryptocurrencies are not viewed as a viable business prospect. Because of their deterministic character, the majority of the arrangements disclosed in the writing for Cryptocurrency value guaging may not be relevant for ongoing value prediction. The prior suggested models induce layer-wise haphazardness into the observed, which includes brain organization enactments to recreate market unpredictability. Our project will provide a method for grouping comparable coins based on their characteristics. The fluctuations in the value of the categorized cryptocurrency are then calculated. After examining some of the most frequently used deep learning algorithms in the presented articles, it is clear that neural network deep learning, as well as other forms of data mining, cannot handle the price prediction issue efficiently and effectively. As a result, it is critical to adopt and create new technologies in order to improve efficiency. Another approach that we may use is social media data mining and epidemic modeling. Using this, we should be able to make better predictions, given social media sites are masters at studying different people's opinions these days. In reality, it is currently being used by a significant number of organizations to forecast the value of the stock market, giving us the opportunity to improve time efficiency and provide better results.

**Keywords:** Social media data mining, Epidemic modeling, Neural network, prediction.

## **Dedication**

We dedicate this paper to those researchers priors who have made our concepts about this topic too clear for us to understand and have motivated us to work on this. Finally we dedicate our paper to our beloved parents, without their prayer we could do nothing.

## **Acknowledgement**

First of all, We want to thank Allah who helped us to continue our research without any major difficulties and helped us to get over our obstacles. Moreover, We are thankful to our respected faculty members for their continuous support who inspired us to work hard in spite of our confusions and mistakes. We are thankful to our parents and families for their enormous support. Lastly, We are grateful to our University.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Approval</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>Dedication</b>	<b>iv</b>
<b>Acknowledgment</b>	<b>v</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Acronyms</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem Statement . . . . .	3
1.3 Objective . . . . .	3
1.4 Thesis Structure . . . . .	5
<b>2 Related Work</b>	<b>7</b>
2.1 Literature Review . . . . .	7
<b>3 Model Used to Collect Data and Result Analysis</b>	<b>11</b>
3.1 Model Description . . . . .	11
3.1.1 Neural Network Models . . . . .	11
3.1.2 Gates of LSTM . . . . .	13
3.2 Algorithm of Used models . . . . .	14
3.3 Experimental Setup . . . . .	15
3.4 Model Setup . . . . .	16
<b>4 Data Collection and Reliability Analysis</b>	<b>18</b>
4.1 Dataset description . . . . .	18
<b>5 Result Analysis</b>	<b>25</b>
5.1 Basic terminology needed for result analysis: . . . . .	25
5.2 Model Result for the Model We Used: . . . . .	28
5.3 Accuracy Analysis for Each Models . . . . .	29

5.4 Future Plan and Conclusion : . . . . .	32
<b>Bibliography</b>	<b>34</b>



# List of Figures

1.1	Relationship cycle of bitcoin value and social media trend . . . . .	2
1.2	Thesis Structure . . . . .	5
3.1	A Long Short-Term Memory (LSTM) unit. . . . .	13
4.1	Dataset training process . . . . .	19
4.2	Dataset training process . . . . .	19
5.2	Details of LTC coins for 60 minutes . . . . .	26
5.3	Close and Volume chart for BTC, LTC and BCH . . . . .	26
5.4	Predicting price using target value . . . . .	27
5.5	Accuracy result using RNN model . . . . .	28
5.6	Accuracy result using RNN model . . . . .	29
5.7	Graph of ARIMA model . . . . .	30
5.8	Graph of ARIMA model . . . . .	31
5.9	On the left, we took data from previous 60 days, on the right we took from previous 90 days. . . . .	32

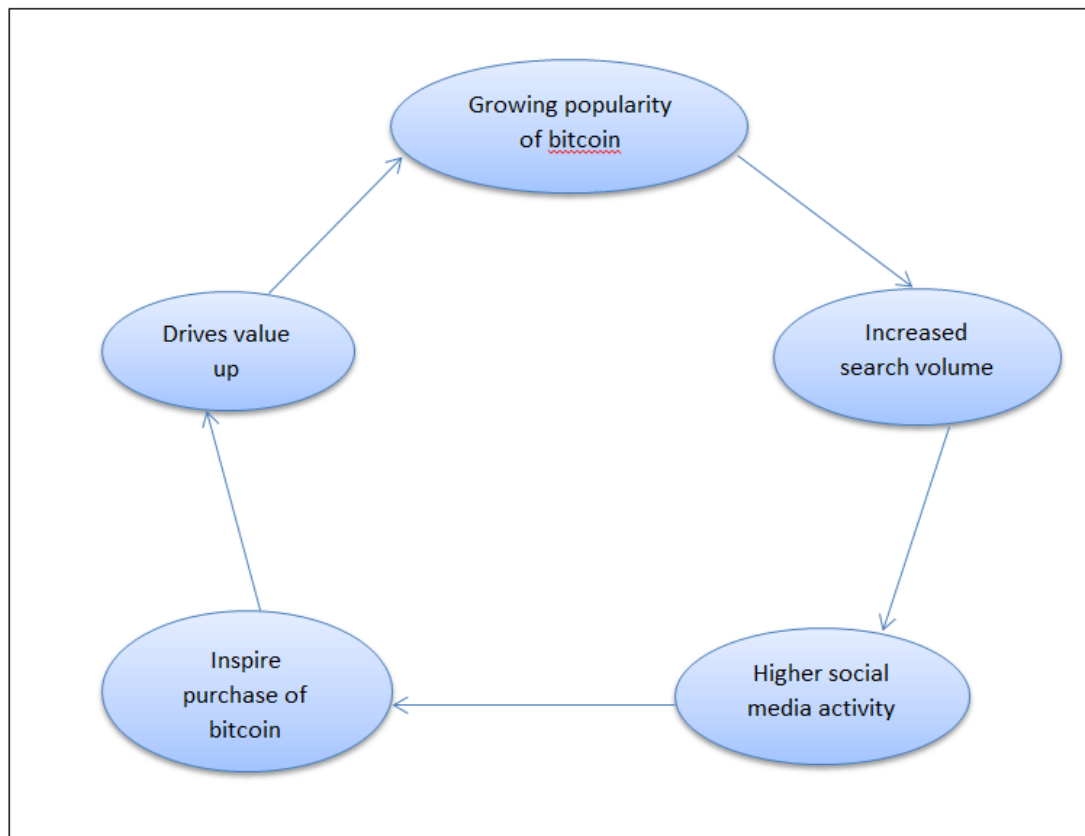
# Chapter 1

## Introduction

### 1.1 Motivation

We have now entered the digital era, and everything is becoming more tech-centric by the day. Digitalization entails progress, which necessitates the allocation of resources. So, when we speak about the evolution of this digital era, we have to talk about currency, which leads to the question, Shouldn't we have a digital form of cash as well? That, however, we already have. That's right, we're talking about cryptocurrency. Cryptocurrency is a type of digital or virtual currency that may be used to buy and sell goods and services. To obtain cryptocurrency, one must first spend actual money. Cryptocurrency comes in a variety of forms, the most well-known of which are bitcoins. Many purchasers believe that Bitcoin is the money of the future, and they are purchasing and storing it in anticipation of its increasing value. Another reason for this preference is the fact that it is decentralized. Blockchain is a method used in cryptocurrency. Unlike traditional currency management, which is controlled by a central bank, the blockchain system does not have a central controller. Many gadgets all across the world monitor and record transactions in the bitcoin realm. As a result, consumers feel more secure and receive a more customized experience with this technique, as no central controller keeps track of the records. Many people consider bitcoin to be a decent investment, despite the fact that it is quite volatile. The value grew once again before the end of 2020. Despite the fact that it is rather volatile, some users regard it as a decent type of investment, and they have good reason to believe so. The stock market is heavily influenced by cryptocurrency. It has been discovered that when the value of bitcoins rises, so does the price of gold and shares. There are a variety of techniques that people employ to make money with cryptocurrencies, and the majority of them are linked to the currency's value, or more precisely, its future worth. As a result, forecasting the future price of bitcoin is a crucial skill in the digital economy. There are several approaches for predicting the future price or worth of cryptocurrencies, one of the most well-known of which is the utilization of social media data. Another important aspect of our daily lives is social media. We have become more reliant on social media than ever before, especially in the wake of the coronavirus outbreak. We use social media for practically everything, including education, health care, entertainment, shopping, and running an internet company. As a result, it is apparent that social media is no longer just for fun. . More than we might imagine, the information and data we post on social media has a significant influence on global finance. New trends emerge

as a result of this knowledge, and trends result in changes in the overall financial condition. Considering and evaluating social media data may provide us with a big lead in predicting future financial growth, and in this case, the rate of cryptocurrencies. Many instances may be found if we look at the impact of social media on business or finance. Tweets promoting forthcoming movies, for example, are a basic example. When people are enthusiastic to see a forthcoming film and the topic of the film climbs on the trending list, it is only logical to assume that the film's box office sales will grow. In terms of the stock market, a Twitter trend analysis may accurately anticipate the stock market's future value with an 87 percent success rate. As a result, evaluating data or information obtained from social media or the web may provide the analyst with a clear picture and aid in the prediction of future value. In the case of cryptocurrencies, there is a cycle based on the link between the price of bitcoin and the popularity of social media. According to that cycle, when bitcoin or cryptocurrency becomes more popular, there will be more searches, tweets, comments, or other social media activity on the subject. People will naturally chat and discuss more about a topic that has grown in popularity, and the same is true for the issue of cryptocurrencies.



**Figure 1.1:** Relationship cycle of bitcoin value and social media trend

As the topic's popularity grew, so did the number of people who participated in social media discussions about it. When a topic gets popular, individuals are more likely to search for it, increasing the search volume for that particular topic. The more people that look into it, the more people get interested. When a result, as individuals learn more about bitcoins and cryptocurrencies, they begin to talk about it more

on social media sites. Furthermore, social media AI algorithms provide you with auto suggestions on things you've looked for or individuals with similar tastes have searched for. The trend comes first, followed by debate, which eventually leads to a real-life investment choice. This is when things start to get even more effective. We all know that when demand rises, so does the value. The value of bitcoin rises as more individuals begin to purchase it. When the value of bitcoins rises, so does the popularity of bitcoins, and the cycle begins all over again

## 1.2 Problem Statement

Cryptocurrency is a digital currency that can exchange for goods and services. Bitcoin, Litecoin (LTC) are some forms of cryptocurrency, and bitcoin is the most common one. In this era of digitalization, many people are using it as an attractive monetary and exchanging forte. Since cryptocurrency is thought to be a better investment and exchanging platform, many use it from the most straightforward form like paying for goods and services, tuitions to big things like donation, business transactions, investments, etc. But the problem is that the nature of cryptocurrency is volatile. Also, the market price of cryptocurrencies is based on demand and supply, so the swapping and trading rate swings a lot. So, this ongoing swings badly effects this sector since cryptocurrencies are gradually being used and are a famous medium for transactions now. Therefore, there is some reason behind this unpredictable fluctuation of the price of cryptocurrency.

**High Profile Losses (Negative News):** Any negative news changes peoples perception in any matter. Sometimes people change their interest in cryptocurrency if they notice that anyone is facing loss or facing dropping of value of their particular crypto coin.

**Speculation:** A particular cryptocurrencys value gets dropped if there is huge availability of that crypto coin. This situation may occur if some owner of a particular crypto coin sells his most coins in the market.

**Regulation:**When a particular type of cryptocurrency is banned by the government, trade or investment in that particular cryptocurrency gets interrupted . It creates disturbance in the price of that cryptocurrency.

**Investors React To security Issues:** People get distressed and feel uneasy to invest in cryptocurrency if they get to know about any breach of security in the system operation of cryptocurrency.It also creates disturbance in the price of cryptocurrency.

**Timing:**If any negative event occurs in the market of cryptocurrency overnight, it can lead to the change of the price of the cryptocurrencies. The change can happen due to any positive event also.

## 1.3 Objective

Attempting to predict future predict price movement of certain crypto currency based on the sequential and historical prices and volumes.(do with RNN)

The theory being that these crypto coins all have relationships with each other. We could possibly predict future movements of Litecoin, by analysing the last 60 minutes of prices and volumes for all 4 (Bitcoin, Litecoin, Ethereum, and Bitcoin) of these coins. I would guess that there exists some, at least better than random, relationship here that a Recurrent Neural Network (RNN) could discover. We can predict future price movement of certain crypto currency based on the sequential and historical prices and volumes.

Our data is not already in some beautiful format where we have sequences mapped to targets. In fact, there are no targets at all. It's just some data points every 60 seconds. So, we have got some work to do.

First, we need to combine price and volume for each coin into a single feature set, then we want to take these feature sets and combine them into sequences of 60 of these feature sets. This will be our input.

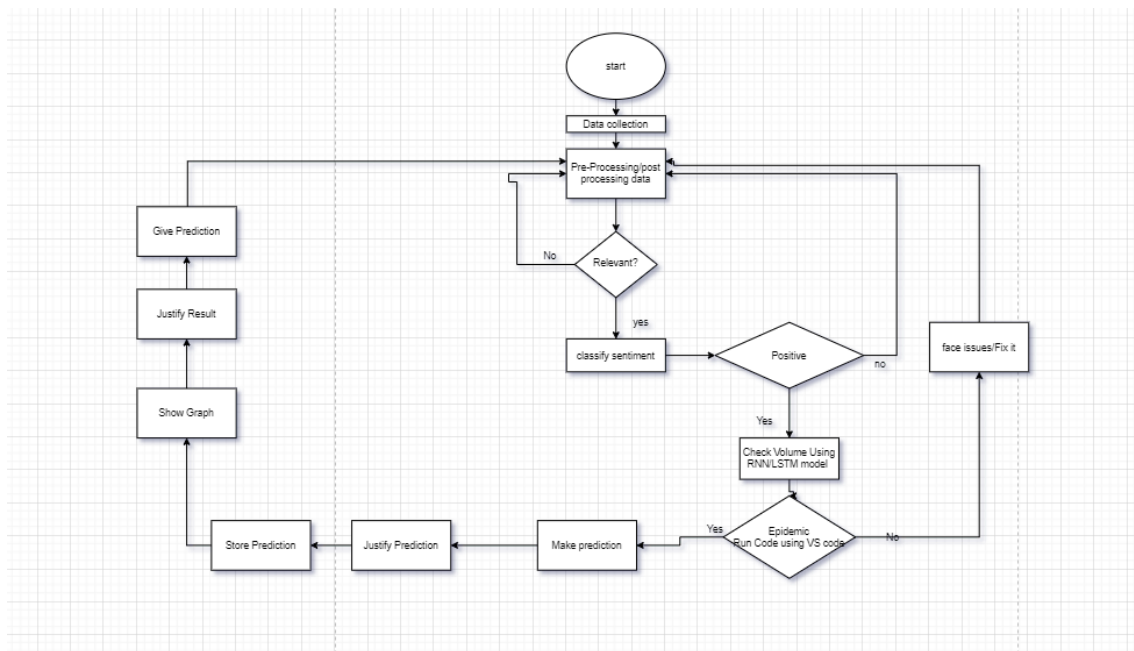
We are trying to predict if price will rise or fall. So, we need to take the "prices" of the item we are trying to predict. Let's stick with saying we are trying to predict the price of Litecoin. So, we need to grab the future price of Litecoin, then determine if it is higher or lower to the current price. We need to do this at every step.

We need to:

- Balance the dataset between buys and sells. We can also use class weights, but balance is superior.
- Scale/normalize the data in some way.
- Create reasonable out of sample data that works with the problem. Then output, which can be profit.

Our purpose is to use data processing to predict the charge of cryptographic money with the guide of utilizing gathering realities from social web, web locales like Facebook, LinkedIn, twitter and so on. We are going to have an outline with the guide of utilizing the employment of google shape to accumulate scrutinizes from the clients of Facebook, Instagram, LinkedIn. For twitter we will utilize hashtag assessment for our research. We will do opinion assessment from tweets or brief directives for get-together realities as tweets every so often explicitly evaluates roughly particular subjects. For this we will choose programmed feeling assessment method that's SentiStrength. SentiStrength can gauge the ability of overall quite helpless feelings in brief messages without issues. We are going to apply google style as appropriately for our examination.

## 1.4 Thesis Structure



**Figure 1.2:** Thesis Structure

There are stages of a Thesis Structure but we didnt go with the conventional routes since our main target goal was different.

Stages of a thesis

1. Abstract: Write the overview of whole thesis on 200-300 words
2. Introduction:
  - (a) Give a background
  - (b) Initial understanding
  - (c) Workflow planning
  - (d) brief review of current knowledge
  - (e) indicate the gap of knowledge and create a hypothesis.

Literature review:

Often part of the Introduction, but can have a separate section. It is an evaluation of previous founding on your topic, where you show that there is a gap in the knowledge that your research will attempt to fulfill. The key word here is evaluation.

Methods:

Often the easiest part of the thesis to write. Outlines which method you chose and why (your methodology); what, when, where, how and why you did what you did to get your results.

Results:

pointers that you found out in relation to your research hypotheses, presented in drawings and in written version with full details

Results contain the facts of your founding. Likely you will include a brief analysis on the significance of important results, with the expectation that more common comments about results will be made in the detailed area. Sometimes analysis and hypothesis are combined: Supervisor was the key

## Discussion

The Discussion section:

- analysis on results;
- explains what my results mean;
- interprets my results in a detailed context; shows which results were expected or unexpected;
- provides explanations for interruptive results.

The Discussion should be elaborated and full of information, the triumph of a successful result discussion helps us neutralize the points and counter points, and Data is the key for that detailed discussion

## Conclusions

Very important! This is where you emphasize that your research aims/objectives have been achieved.

We concluded by giving our closing statement. And how it affected us on a personal level.

# Chapter 2

## Related Work

### 2.1 Literature Review

Cryptocurrency price can be predicted using social media data, and for analyzing those data one needs to follow any model or a structural approach to analyze the data better. RNN can be a great way to predict any kind of data's future update. RNN means Recurrent Neural Network. RNN was first used for Apple's Siri and Google's voice assistance. RNN uses an internal memory to remember its values so that it's very easy to work with RNN with sequential datasets. RNN combined with LSTM which means Long Short-term Memory becomes a robust model which makes RNN a real front runner in the prediction accuracy model run.[15]

In one paper they said in RNN it is easy to use because they remember what the previous data was while working with the next data. This helps to make the calculation so far a lot easier. RNN may seem like an arbitrary long sequence but in real it just unfolds the layers and work with that. There isn't much difference between LSTM and RNN as they use the same concept, they just have a functional difference. LSTM uses cells like blank boxes and puts the data value inside the boxes as long as the process is running unlike RNN which uses data like a chain sequence.[2]

Sean McNally et al. used LSTM and RNN to predict the price of Bitcoin to USD. They also used ARIMA in that process. There were some limitations in neural networking while working with long sequences but RNN overcame that and became popular in deep learning. The in-depth training by replacing the hubs within the RNN with memory cells and an activation device, LSTM's neural systems address the challenges with RNN in reference to escape angles. For this, it's robust and deep learning brain architecture, since it simultaneously stores temporary information within the short term use technique which is efficient but can also be used in the future.[3]

There is a specific technique to deal with a data bubble. For analyzing data which shows bubble like feature, a pair of data is taken and compared, and the elements of that pair is exchanged if necessary. In one paper, they said that as the data showed bubble like approach, they went for the epidemic detection modeling. Epidemic detection modeling was first used by epidemiologists to detect an outbreak of epidemic. They followed the SIR model which can be split in three sub categories which are as following:



S = Susceptible, I = Infected, R = Recovery

Using this model they considered three categories among patients and they transmitted the data of patients from one category to another based on pre-defined rate formula.

HMM is a data model which was also used in the times of influenza outbreaks. HMM means Hidden Markov Model which they used to detect epidemic and non-epidemic states of social media data and trending volumes on their paper. Here, epidemic means trending and non-epidemic means non-trending, and they used the HMM approach to classify and find out whether the data was in the group of the trending or the non-trending topics on twitter. HMM has four components which are:

- a. Number of hidden states
- b. Observation probability distribution
- c. Transition probability
- d. Initial state distribution

In another paper, they mentioned the importance of sentiment analysis and decide on the positive or negative results. For example, lets say someone has a number of data about cryptocurrency. How is that person going to predict using that set of data? On what basis the decision will be taken? The answer to these questions is: Sentiment analysis. They analyzed their set of data and tried to understand the sentiment behind those responses. Sentiment can be classified in two ways: positive sentiments and negative sentiments. Positive sentiment means that the user or the person whose response is taken into consideration is interested in the topic mentioned, in their case, bitcoins. The more interested users will be found, the more increment will be added in the section of positive sentiment result. Negative sentiment, on the other hand means that the users are no longer interested in this topic or no longer talking about it. Negative sentiments can implement a downwards effect on the values of bitcoins, whereas positive sentiments can increase the value within few days. For example, lets say someone has a number of data about cryptocurrency. How is that person going to predict using that set of data? On what basis the decision will be taken? The answer to these questions is: Sentiment analysis. In a paper they have used ARIMA Model to predict the price of Cryptocurrency. ARIMA, short for Auto Regressive Integrated Moving Average is actually a class of models that explains a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values. Any non-seasonal time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models. An ARIMA model is characterized by 3 terms.,

where,  $p$  is the order of the AR term  $q$  is the order of the MA term  $d$  is the number of differencing required to make the time series stationary If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for Seasonal ARIMA. More on that once we finish ARIMA.

So, what does the order of AR term even mean? Before we go there, lets first look at the  $d$  term. Another group of researchers stated the importance of data cleaning in their paper. While collecting data in a big number, there can be a bunch of undesirable or

irrelevant data which will be of no use in the research purpose. A small yet important piece of information was stated by another group of researchers. In their paper, they mentioned the importance of sticking to only one type of cryptocurrency. We know, there are so many variations and types of cryptocurrencies available. For the research purpose they considered bitcoin as their chosen form of cryptocurrency as it is the most popular one as well. They chose bitcoins and they kept it constant the entire process of research. An informative idea was given A very interesting in another paper. The writers stated the importance of influence approach. Influence is a very important term when talking about cryptocurrency price detection using social media.[11]

But when someone decides to start analyzing sentiments, there are more things which need to be taken under consideration other than just the issue of positive and negative sentiments. In another paper, they stated some criteria which need to be taken under consideration when you want to go for sentiment analyzation for giving prediction. The first thing they followed is, in their set of data, they checked that how many of those data actually had sentiments. This is necessary because in social media, not always everyone posts about things driven by sentiments singularly. A piece of information can be fact, for example, and fact is of no use in sentiment analyzation. So choosing the right data for analyzing sentiment can save your time and spare the mess. Secondly, they talked about another criteria which is, there has to be a correlation between the sentiment and cryptocurrency. They collected the data and analyzed those to predict the price of cryptocurrency, so if there is no correlation between these two, or if the sentiments are irrelevant or illogical, then the worth of those data will be null. Also, while considering the response they mentioned the importance of finding out and differentiate responses of humans and bot and working with the data provided by humans only, as automated posts or response can corrupt the understanding of actual prediction. They also said, if the value is unchanged, the sentiment can be still positive. On their research they showed that the value of bitcoins got very low after 2017, yet people were interested about bitcoins and they were discussing about it. So just because there are positive sentiments that doesnt mean that the pricing or value will also go up every time. To solve this dilemma, they preferred the authenticity and success rate of trending volume more than sentiments, because according to their study, volume has 87% accuracy. So along with sentiments, volume should be a major aspect of consideration as well.[11]

Another group of researchers stated the importance of data cleaning in their paper. While collecting data in a big number, there can be a bunch of undesirable or irrelevant data which will be of no use in the research purpose. For example, while predicting price of bitcoins, you can get information like the history of bitcoins, or cryptocurrency classification which is not helpful in price detection so these types of data are considered as undesirable data. So, data cleaning needs to be performed to clean this undesirable information which could reduce the credibility of the necessary information.[7]

A small yet important piece of information was stated by another group of researchers. In their paper, they mentioned the importance of sticking to only one type of cryptocurrency. We know, there are so many variations and types of

cryptocurrencies available. For the research purpose they considered bitcoin as their chosen form of cryptocurrency as it is the most popular one as well. They chose bitcoins and they kept it constant during the entire process of research. They didn't switch to another form of cryptocurrency because mixing multiple types can cause corrupted results, wrong predictions and structural malfunction. So it is vital to stick with whatever form was selected first throughout the process.[7]

A very interesting and informative idea was given in another paper. The writers stated the importance of influence approach. Influence is a very important term when talking about cryptocurrency price detection using social media. Influence affects the social media platforms in so many ways, people follow influential persons, people get influenced by any content, or strong message can leave influence on the entire life view of any individual. So influence is very important and must be taken in consideration. According to the writers, influence is the strongest correlation between social media and cryptocurrency. They explained their claim in a very intelligent and resourceful way. Influence is basically the capacity of something or someone to have an effect over someone else's choices, decisions or preferences. Influence is normally controlled by users of central position in social media. Users of central position are those who are followed and admired by a large group of people. The effect of influence in social media and cryptocurrency can also be explained by social science, taking the user's behavior, psychological structure and the influencer's personality into consideration.

Influence can be classified and used in cryptocurrency in two ways, which are:

a. Influential users discovery:

They tried to find users who have effect or can bring difference to others. Finding these users would enable them to spread the idea or fame of cryptocurrency as those users already have influence over the others. These types of users were found by them following another two methods, which are:

1. Topology based search: This search was done by considering centrality of users in various aspects or areas.
2. Taking the history of users, contents shared by users, and their influence into consideration.

Also, in platforms like twitter or Instagram, hashtag can also be used as a key to find influential figures.

b. Influence diffusion information spreading:

This method was used by them by targeting general users to spread the information about cryptocurrency and ensuring maximum reach. Not everyone is interested about cryptocurrency or has the willingness to share a post about it. So, finding the right users who have interest and delivering the information in front of them is the key here. Also they made sure that the shares should reach maximum possible audiences.

So, influence plays a very important role in cryptocurrency price prediction using social media.[2]

# Chapter 3

## Model Used to Collect Data and Result Analysis

### 3.1 Model Description

#### 3.1.1 Neural Network Models

**RNN:** A feedback neural network with an internal memory is known as a recurrent neural network. RNN is recurrent in nature since it executes the same function for each data input, and the current input's outcome is dependent on the previous calculation. The output is duplicated and transmitted back into the recurrent network once it is created. It analyzes the current input as well as the output it has learnt from the prior input when making a decision. RNNs, unlike feedforward neural networks, may process sequences of inputs using their internal state (memory). As a result, activities like unsegmented, linked handwriting recognition or speech recognition are possible. All of the inputs in other neural networks are independent of one another. In an RNN, however, all of the inputs are linked to one another.

It first extracts  $X(0)$  from the series of inputs, then produces  $h(0)$ , which, together with  $X(1)$ , serves as the input for the following phase. As a result, the following step's inputs are  $h(0)$  and  $X(1)$ . Similarly, the following step's input is  $h(1)$ , and the next step's input is  $X(2)$ , and so on. As a result, it remembers the context while training.

The formula for the current state is:

$$h_t = f(h_{t-1}, x_t)$$

Applying Activation Function:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

Why is the weight at the previous hidden state, why is the weight at the current input state, and  $\tanh$  is the Activation function, which implements a non-linearity that squashes the activations to the range. [-1.1]

Output:

$$y_t = W_{hy}h_t$$

Here  $Y_t$  is the output state.

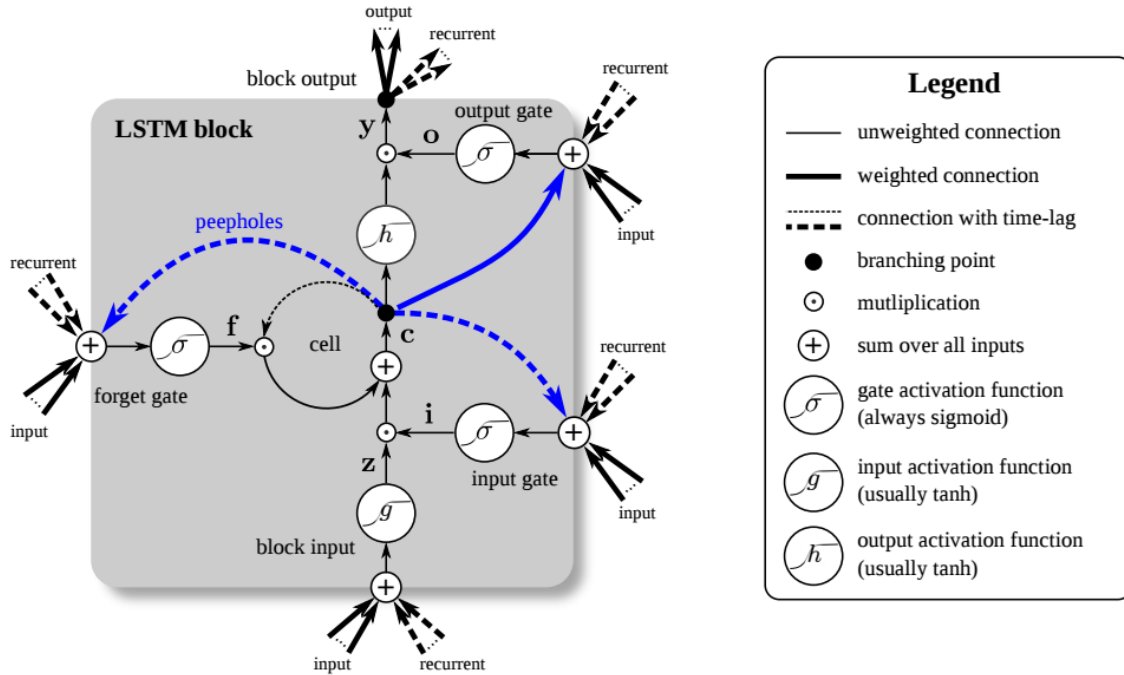
There is obviously some advantages and disadvantages exist in RNN.

**Advantages of RNN :** RNN can model data sequences so that each sample is considered to be reliant on the ones before it. Even convolutional layers are employed with recurrent neural networks to increase the effective pixel neighborhood. RNN can represent data sequences in such a way that each sample is dependent on the ones preceding it. With recurrent neural networks, even convolutional layers are used to improve the effective pixel neighborhood.

**Disadvantages of RNN:** The computation is slow because of its recurrent nature and training is very difficult. It becomes very difficult to process sequences which are too expanded.

Long short-term memory (LSTM) is a deep learning architecture based on an artificial recurrent neural network (RNN), In contrast to traditional feedforward neural networks having feedback connections. LSTM is a form of recurrent neural network that is specifically intended to keep the neural network output for a given input from fading or bursting as it cycles through the feedback loops. It can handle not only individual data points (such as photos), but also complete data streams (such as speech or video). Because there might be lags of undetermined duration between critical occurrences in a time series, LSTM networks are well-suited to categorizing, processing, and making predictions based on time series data. LSTMs were created to solve the problem of vanishing gradients that can occur when training traditional RNNs. In many cases, LSTM has an advantage over RNNs, hidden Markov models, and other sequence learning approaches due to its relative insensitivity to gap length.

**LSTM:** The Long Short-Term Memory Architecture is made up of linear units with a constant weight of 1.0 self-connection. This permits a value (forward pass) or gradient (backward pass) to be saved and retrieved at the necessary time step. The output the previous time step is the same as the output for the following time step when using the unit multiplier. The memory cell, a self-recurrent unit, is capable of retaining information from dozens of time steps before. This is quite useful for a variety of activities. For text data, an LSTM unit can, for example, retain information from the previous paragraph and apply it to a sentence in the current paragraph. The input sequence is trained on two LSTMs in bidirectional LSTMs: one on the standard input sequence and the other on the opposite input sequence. This can help LSTM networks function better by allowing future data to contextualize previous data in a time series. Simple feed-forward networks can't handle complicated sequence learning/machine learning problems as well as these LSTM networks.  
ng Short-Term Memory Architecture



**Figure 3.1:** A Long Short-Term Memory (LSTM) unit.

The LSTM unit has four input weights (from the data to the input and three gates) and four recurrent weights (from the output to the input and the three gates). Peepholes are extra connections between the memory cell and the gates, but they do not increase the performance by much and are often omitted for simplicity. Image by Klaus Greff and colleagues as published in [LSTM: A Search Space Odyssey](#). Image by Klaus Greff and colleagues as published in [LSTM: A Search Space Odyssey](#).

### 3.1.2 Gates of LSTM

1. **Input Gate:** It determines which input value should be used to modify the memory. The sigmoid function determines which values are allowed to pass through 0,1. and the tanh function assigns weight to the data provided, determining their relevance on a scale of -1 to 1.

$$i_t = \sigma (W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh (W_C \cdot [h_{t-1}, x_t] + b_C)$$

2. **Forget Gate:** It determines which details should be removed from the block. The sigmoid function determines this. For each number in the cell state  $C_t$ , it looks at the previous state ( $h_{t-1}$ ) and the content input ( $X_t$ ) and produces a number between 0 (omit this) and 1 (keep this).

$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$$

3. **Output Gate:** Here the output is determined by the block's input and memory. The sigmoid function determines which numbers are allowed to pass through 0,1. and the tanh function multiplies the output of Sigmoid by the weightage given to the values provided, determining their level of importance ranging from -1 to 1.

$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh (C_t)$$

We compared our RNN model with ARIMA model: The term "autoregressive integrated moving average" (ARIMA) is an acronym for "autoregressive integrated moving average." It's a time series model used in statistics and econometrics to track events across time. The model is used to decipher historical data or forecast future data in a series. ARIMA's key advantage is that it simply takes data from the time series in question. If we are forecasting a huge number of time series, this functionality comes in handy. ARIMA has a number of flaws, including difficulties generalizing results from a particular study. It can not remember the previous input like RNN.

## 3.2 Algorithm of Used models

Sequential data is simply ordered data in which related items appear one after the other. Financial data or the DNA sequence are two examples. Perhaps the most common form of sequential data is time series data, which is just a collection of data points in chronological order. In most cases, a conventional neural network analyzes the input and then continues on to the next without regard for the order. Sequential data, on the other hand, is processed in a certain sequence that is required to comprehend them clearly. An RNN model's design is comparable to that of a convolutional neural network (CNN) or other artificial neural networks. A recurrent neural network has three layers: an input layer, a hidden layer, and an output layer. These layers, on the other hand, function in a predictable order. The data is fetched by the input layer, which then performs data preprocessing before delivering the filtered data to the hidden layer. A hidden layer is made up of neural networks, algorithms, and activation functions that are used to extract valuable data from the data. Finally, the data is transmitted to the output layer, which produces the desired result. To comprehend RNN, you must first comprehend the concept of a feed-forward network. The distinction between a feed-forward and a recurrent neural network is seen in the diagram above. The information in a feed-forward neural network can only flow in one way. Because a feed-forward neural network only analyzes current input, it has no idea what has

happened in the past other than the training processes. The RNN method varies a great deal. The data that traveled through the architecture is looped back on itself. For decision-making, each input is dependent on the preceding one. For each layer in the network, RNN assigns the same and equal weight and bias. As a result, all of the independent variables become dependent variables. The RNN's loops ensure that the information in its memory is maintained. If you're curious in this algorithm's capacity to store data in memory, I'll reveal the secret. This is made feasible by the major component of the system, which is long-short term memory (LSTM). RNNs have been proven to be the most effective for machine translation, speech recognition, conversational AI (Chatbots), and other related technical advancements.

### 3.3 Experimental Setup

Basic libraries and functions needed for model setup:

We used Python v-3 for our work. Python is the vast and most used for Data Science, Machine Learning, Deep Learning etc., and it has most libraries for use.

1. **Pandas:** Pandas is a Python Library which is used for data analysis and we used it, Pandas allows importing data from various file formats such as comma-separated values, Microsoft Excel, SQL. And pandas allow us to manipulation operations such as merging, reshaping, selection, as well as data cleaning.
2. **NumPy:** NumPy is a Python Library, its mainly used for larger numbers of data set and facilitate advanced mathematical. We used NumPy in our data preprocessing part where we converted the feature set list into NumPy array
3. **TensorFlow:** TensorFlow is a Python library for fast numerical computing, it is a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries.
4. **Keras:** Keras is an open-source library which provides a Python interface for Neural Network, Keras acts as an interface for TensorFlow library, its train Neural Network models. We used Keras to train our model.
5. **Epoch:** An epoch is the number of times an algorithm visits the data set. We set epoch value more than one to optimize the process as the dataset have to pass through same neural network many times. And we can set the Epoch volume on our code.
6. **Batchsize:** The batch-size is the number of samples processed before the model is updated and will be transmitted through the network at one time. The size of a batch has to be more than one or equal to one. Generally, the larger the batch size, the faster the model can complete each epoch during training period.
7. **Adam Optimizer:** Adam optimizer is an algorithm that is used to update network weights recursively based on the training data rather than the traditional stochastic gradient descent method.
8. **Sequential():** We imported sequential library from Keras in order to get a sequential data list of the values received after data normalization.
9. **Model Checkpoint:** The Model Checkpoint library allows where to checkpoint the model weights. It also includes the file name process and the cir-



cumstances under which the model checkpoint should be created. Mainly, the value accuracy we got after compiling, the Model Checkpoint will provide a checkpoint for that.

10. **Preprocessing:**We imported preprocessing from Sklearn in order to balance and normalize the main data frame. We used preprocessing to process the data so that it can be trained afterwards.
11. **df.fillna():** fillna() method is used to fill Null values in the whole data frame. It can also be used to modify in place, columns.
12. **df.dropna():**dropna() method is used for cleanup or removing values. It will look for rows with missing values and will drop the entire row if there is any missing value in the row.
13. **deque():** After data normalization, we created a sequential data list. To control how many sequences, we will get in this sequential data list, we used deque().
14. **model.Compile():** model. Compile() is a method provided by Keras. It is used to compile a model. It sets the argument passed to it like optimizer, metrics or loss function.
15. **model.fit():** We used model.fit() to fit the train\_x data and train\_y data after compiling. It is mainly used to transfer training and validation data.
16. **model.add():** We used model.add() to add sequential data to the model.
17. **model.save():** model.save() is used here to save the trained model.

### 3.4 Model Setup

We used Python version 3 to implement our models. We have created our RNN model and used LSTM here, used NumPy library where we converted our feature set list into NumPy array. Then we used Pandas library for data manipulation and data analysis.

**For LSTM and RNN model:**In this model, we used train-test split method to divide the data into two segments that are train set and test set. We passed the test set to validation data when we used to fit our model. From Keras, we imported long short-term memory (LSTM) and Dense for creating the LSTM model. To get the data of last sixty minutes we set the SEQ\_LEN=60 and after training to get the validity of the data for three minutes we set the FUTURE\_PERIOD\_PREDICT= 3.If price goes up in three minutes then its a buy (1), otherwise its a sell (0). Then to predict the future price of our dataset we created future values of close(price) data in the ratio of the dataset. We fixed a time in the data frame to get the last 5% of data and predict future price of last 5% of data. After performing the validation of data, to balance and normalize the data we performed preprocessing. We imported Sequential from Keras to get a sequential data list of the normalized data set. After sequence, to rearrange the data we shuffled the data. After adding sequential data in the model, to get the data linearly we used to rely (rectified linear). Then using model. Compile() we compiled our model by defining loss function as sparse\_catagorical\_crossentropy , metrics as accuracy and optimizer as opt. Finally, to pass the training and validation data and display them in the tensor board we used model.fit() with batch-size as Batch-size and epochs as EPOCHS.

We preprocessed the test set in the same way as the train set. Then the value

accuracy we got after compiling, the ModelCheckpoint will provide a checkpoint for that. Everytime getting a better accuracy, the RNN-file in the filepath provides updated file and the best value is saved using save\_best\_only=True. Finally, we got our expected accuracy and loss percentage for prediction, we checked the score of the value accuracy of the dataset and saved the model.

# Chapter 4

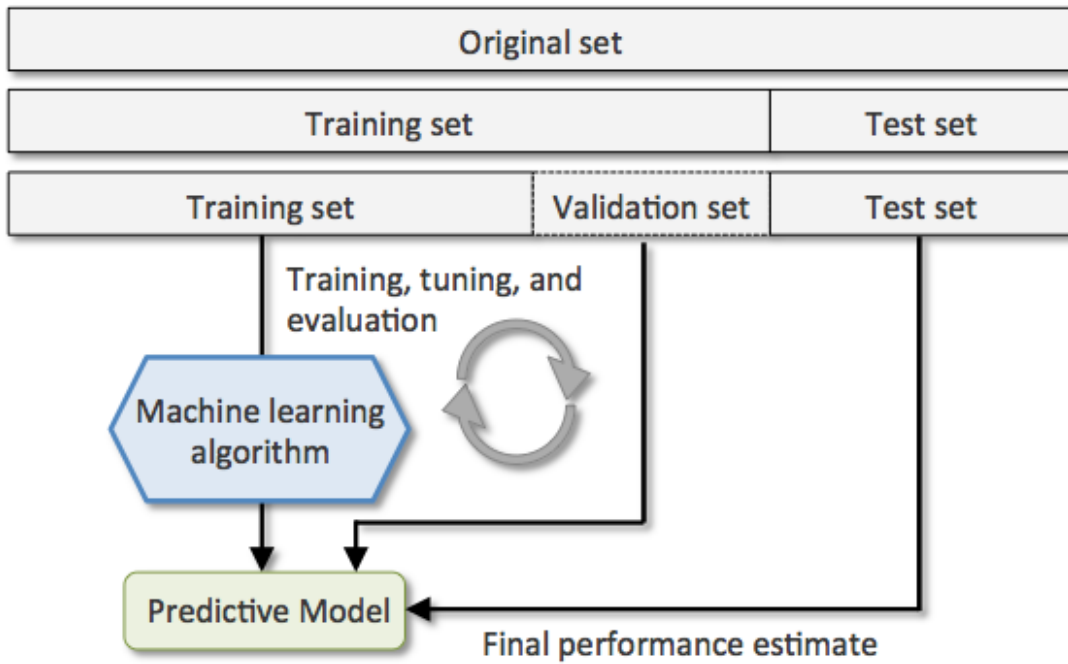
## Data Collection and Reliability Analysis

### 4.1 Dataset description

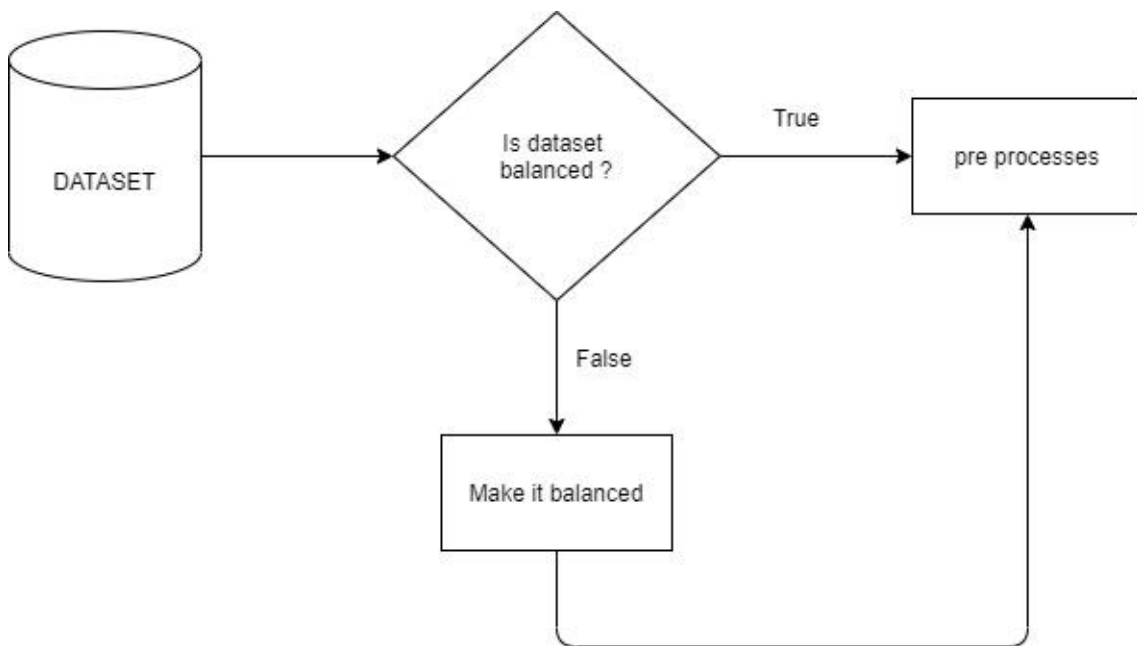
ML depends heavily on data, without data, it is impossible for an AI to learn. It is the most crucial aspect that makes algorithm training possible. No matter how great your AI team is or the size of your data set, if your data set is not good enough, your entire AI project will fail! I have seen fantastic projects fail because we didn't have a good data set despite having the perfect use case and very skilled data scientists.

I want to introduce you to the first two data sets we need: the training data set and test data set because they are used for different purposes during your AI project and the success of a project depends a lot on them (Alexandre, 1902-14).

1. The **training data set** is the one used to train an algorithm to understand how to apply concepts such as neural networks, to learn and produce results. It includes both input data and the expected output.
2. The **test data set** is used to evaluate how well your algorithm was trained with the training data set. In AI projects, we can't use the training data set in the testing stage because the algorithm will already know in advance the expected output which is not our goal.



**Figure 4.1:** Dataset training process



**Figure 4.2:** Dataset training process

In those diagrams above we found out about how dataset is trained and tested. It is vital for a dataset to get ready before post processing.

Testing sets represent 20% of the data. The test set is ensured to be the input data grouped together with verified correct outputs, generally by human verification. so

we can say data testing and training process aren't the most solid ones we have come across, it's like 20% of the work is done by the flow but 80% fixation had to be done manually. I have to go through the whole process together like seeing missing columns and missing values and where the recites then I have to fix them one by one. For our more complex and interesting ever changing data it wasn't easy. Cause for our price prediction program human errors are easy to come by but we fight for the best solutions and purity of our dataset so that we can get a good result above in those Diagram we found out about how dataset is train and tested. It is vital for a dataset to get ready before post processing.

All projects are somehow different but I'd say that you need 100 times as much data as the number of parameters in the model being Created. The more difficult the task, the more data set information needed. See when it was about our data we had to think really hard what we should do when it comes to our data. Because the more data we collect means the more precise prediction can be done by us. But it also means the more data we take the more processing we have to go through in order to get the best fitting result.

#### **How we get the dataset?**

1. Google Dataset Search.
2. Kaggle
3. Data.Gov
4. Datahub

These are the functional ways we collected our data. we went through the process and found out these data then went through pre processing and post processing. The data we will be using is- Open, High, Low, Close, Volume data for- Bitcoin, Ethereum, Litecoin and Bitcoin Cash. ( you can use any cryptocurrency in the code not just Bitcoin, Ethereum.... )

Open high Low volume data are really useful for the betterment of easy finding and gifting solution

For our purposes here, going to only be focused on the Close and Volume columns.

What are these?- The Close column measures the final price at the end of each interval (Creating a Cryptocurrency-Predicting Finance Recurrent Neural Network - Deep Learning Basics with Python, TensorFlow and Keras p.8, n.d.). In this case, these are 1 minute intervals. So, at the end of each minute, what was the price of the asset.

The Volume column is how much of the asset was traded per each interval, in this case, per 1 minute. (Close is the price of the thing. Volume is how much of thing.)

Now, we have a few of these "things/products" We are going to be tracking the Close and Volume every minute for-

Bitcoin, Litecoin, Ethereum, and Bitcoin Cash. The theory being that these cryptocurrencies all have relationships with each other. Could we possibly predict future movements of (say-) Litecoin, by analyzing the last 60 minutes of prices and volumes for all 4 (Bitcoin, Litecoin, Ethereum, and Bitcoin) of these coins?

I would guess that there exists some, at least better than random, relationship here that a Recurrent Neural Network(RNN) could discover. so how do we do this? Our data is not already in some beautiful format where we have sequences mapped to targets.

In fact, there are no targets at all. It's just some datapoints every 60 seconds. So, we have got some work to do. First, we need to combine price and volume for each coin into a single feature set, then we want to take these feature sets and combine them into sequences of 60 of these featuresets. This will be our input. what about our output? Our targets? we are trying to predict if price will rise or fall. So, we need to take the "prices" of the item we are trying to predict. Let's stick with saying we are trying to predict the price of Litecoin. So we need to grab the future price of Litecoin, then determine if it is higher or lower to the current price. We need to do this at every step.

Now what we need to do:

Balance the dataset between buys and sells. We can also use class weights, but balance is superior.

Scale/normalize the data in some way.

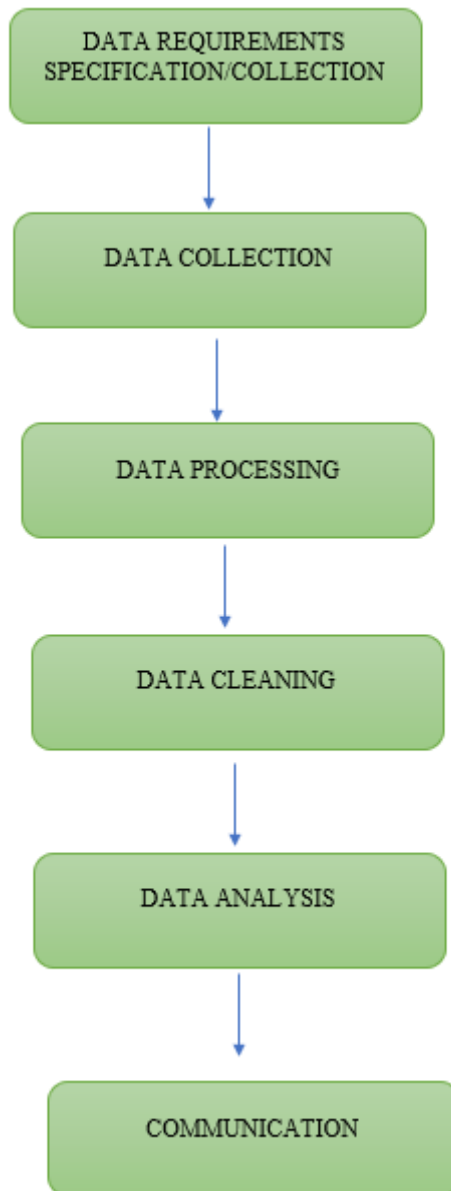
Create reasonable out of sample data that works with the problem.

then output what we did step by step:

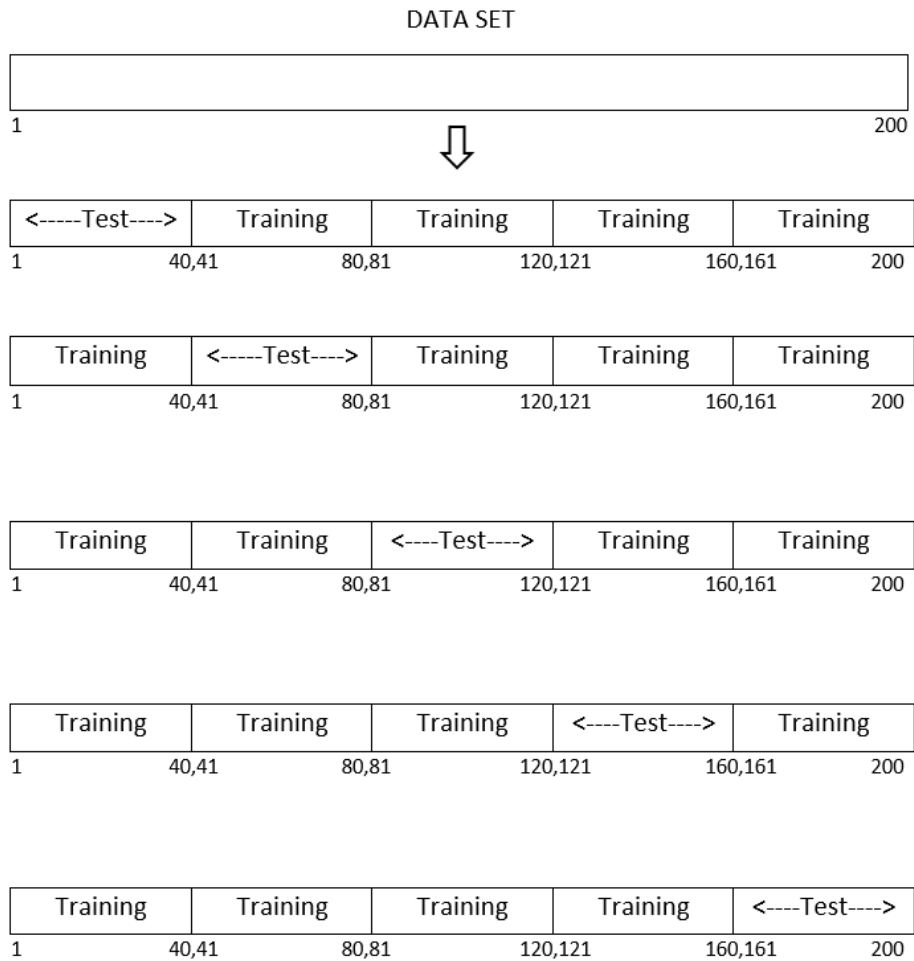
1. Find my suitable dataset above
2. I can learn more about your dataset searching online, but it is not something I have to do
3. Load the dataset as dataframe using pandas
4. Handle missing values if needed
5. Encode categorical features if needed
6. Scale all the values between 0-1 with proper scaling technique
7. Split the dataset into features and labels. Use your intuition to determine which column indicates the labels.
8. Perform classification and calculate accuracy using logistic regression (Use sklearn library). Perform necessary pre-processing on the dataset before classification. Use 8:2 train-test split.
9. Perform classification and calculate accuracy using decision tree (Use sklearn library). Perform necessary pre-processing on the dataset before classification. Use 8:2 train-test split.
10. Support Vector Machine (SVM), Neural Network (Multilayer Perceptron Classifier) and Random Forest are three very popular machine learning classifiers. Divide the dataset into 8:2 train-test split and perform Support Vector Machine, Neural Network (MLPClassifier) and Random Forest on it using sklearn library. In the previous assignment, we have already used Logistic Regression and decision tree classifiers from the sklearn library. Just change the imports and the function calls to use other classifiers. Take a look at the sklearn documentation for further information (How to Predict Dataset Using NaiveBayes, n.d.). Regression Algorithms are used with continuous data. Classification Algorithms are used with discrete data. The regression Algorithm can be further divided into Linear and Non-linear Regression. The Classification algorithms can be divided

into Binary Classifier and Multi-class Classifier. Binary classifier refers to classifying instances into one of two classes and multi-class classifier refers to classifying instances into one of three or more classes. We used Binary classifier here. We could also make the prediction with a regression question, using linear activation with the output layer but instead we did our work using binary classification. Binary classification is used to categorize new probabilistic observations into given categories and we have two categories in here( price and volume).

11. Perform Perform dimensionality reduction using PCA. Reduce the number of feature vectors into half (e.g. if your dataset has 10 columns, after applying PCA it should have 5 columns). Principal Component Analysis(PCA) is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. PCA is unsupervised statistical technique that is used to reduce the dimensions of the dataset. ML models with many input variables or higher dimensionality tend to fail when operating on a higher input dataset. PCA helps in identifying relationships among different variables then coupling them. Though our topic shows regression related problem but we used dimensionality reduction using PCA here after splitting the dataset. Before using dimensionality reduction we used SVM to handle sparse data. Mainly, for feature selection dimensionality reduction is used. Also it helps to find the important features that is needed for the prediction by reducing the number of random variables or features under consideration in an ML algorithm. In our dataset, initially we had many columns like open, high, low, close, volume but we only needed close and volume column for our prediction. And using dimensionality reduction using PCA we dropped the others. Dimensionality reduction also helps to reduce computation time, to remove redundant features. So we used dimensionality reduction using PCA here before scaling and normalizing.
12. Apply Support Vector Machine, Neural Network (MLP Classifier) and Random Forest again on the reduced dataset.
13. Compare the accuracy of the pre-PCA and post-PCA results. Plot them against each other. These are the ways we used pre-processing. We learned a lot about pre-processing how it is really vital for us as a project and it is the sole reason that helps us to modify the dataset as a whole. Without a processed data we can't go far in this race. Data is an essential part of this ever running project to find the prediction that will help us to be a better version of ourselves. Below we are showing some use case diagram that magnifies the way we approach these dataset collection process.







This diagram shows the Testing and Training data each data once a test data and in other iteration this data represent as a training data this solve the problem of overlapping.

Then we went with post processing data factors. Which were also vital for the whole process.

Postprocessing procedures usually include various pruning routines, rule filtering, or even knowledge integration. All these procedures provide a kind of symbolic filter for noisy and imprecise knowledge derived by an inductive algorithm. Finite element model data post-processing, software that makes computer calculation output easier to understand. so in short we can say, Data preprocessing is the process of putting raw data into an understandable more easier format so that machine can get it as well as humans. It is also an important part in data mining and collection program as we cannot work with raw unprocessed data. The quality of the data should be ensured before applying machine learning AI on full effect or data mining aka collection algorithms like tensor flow, we all always use tensor flow cause its sacred and effective (Sadhvi, 2021).

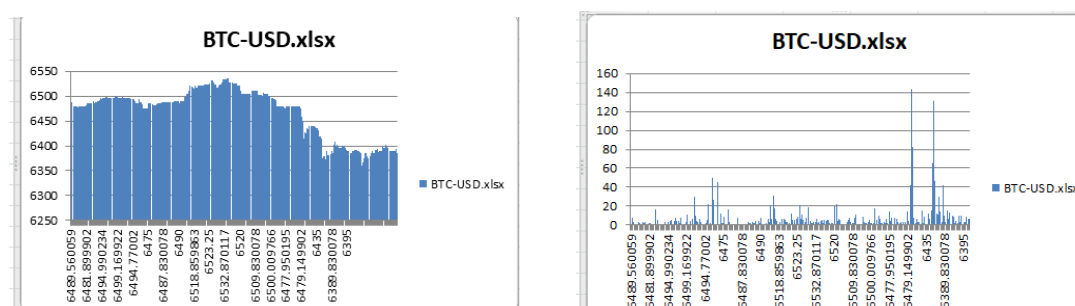
# Chapter 5

## Result Analysis

### 5.1 Basic terminology needed for result analysis:

First and foremost, we needed trained dataset to analyze the predictive model. In this case we are using RNN model with some functions of LSTM. Some of the basic terminologies for result analysis are:

1. **Training Dataset:** After we collected our dataset, it needed to go through a process of data training. Training dataset is the initial dataset used in neural networking models. Trained set helps the computer to process the information in an efficient and easier way. Trained datasets can be classified in so many ways. In our case, we took raw data and classified them according to the collection time, high, low and volume. All the data were sorted separately in an excel file and we used the data as a dataset as csv file in the model.
2. **Data Analysis:** We stored the data in an excel file and the analysis was done using those data. In our previous work history, we used ARIMA model which failed to give us better outcomes later. While using ARIMA model, our prediction accuracy percentage was very low whereas using RNN gave us a very good percentage for the prediction. The difference is shown below in two figures:



**Figure 5.1:** On left the result of RNN model and on right the end result of ARIMA model.

This figure shows the errors in the prediction we attempted using ARIMA

model whereas using RNN gave us a much better result.

- Running The Dataset in The Model:** We converted the dataset as csv format and it was ready to use in our model code. After passing the dataset through RNN model it will give us the time of collection, the highest and lowest price of the last 60 minutes duration and the volume of the desired coins. We mined data for mainly 4 coins which are BTC, ETH, BCH and LTC. We set the time limit as 60 minutes that means that the model will give us prediction based on the last 60 minutes state of the coins price update. We gave it a try for LTC and we found the table which shows us the detailed information of the updates of LTS for the last 60 minutes.

```

$ python cryptocurrency_price_prediction.py
      time      low      high      open      close      volume
0 1528968660 96.580002 96.589996 96.589996 96.580002  9.647200
1 1528968720 96.449997 96.669998 96.589996 96.660004 314.387024
2 1528968780 96.470001 96.570000 96.570000 96.570000  77.129799
3 1528968840 96.449997 96.570000 96.570000 96.500000  7.216067
4 1528968900 96.279999 96.540001 96.500000 96.389999 524.539978
  
```

**Figure 5.2:** Details of LTC coins for 60 minutes

As we know, for price prediction the only two main important variables are the close price and the volume. So we wanted to find out these two variables for multiple coins and we came up with the following chart:

```

time      BTC-USD_close  BTC-USD_volume
1528968660  6489.549805    0.587100
1528968720  6487.379883    7.706374
1528968780  6479.410156    3.088252
1528968840  6479.410156    1.404100
1528968900  6479.979980    0.753000
time      LTC-USD_close  LTC-USD_volume
1528968660  96.580002     9.647200
1528968720  96.660004    314.387024
1528968780  96.570000    77.129799
1528968840  96.500000    7.216067
1528968900  96.389999   524.539978
time      BCH-USD_close  BCH-USD_volume
1528968660  871.719971    5.675361
1528968720  870.859985    26.856577
1528968780  870.099976    1.124300
1528968840  870.789978    1.749862
  
```

**Figure 5.3:** Close and Volume chart for BTC, LTC and BCH

- Calculation of the prediction:** From fig 3 we can see that we have the time of data collection and the close and volume of the coins. Now question may raise that how does a computer calculate this and give us prediction? Here we will need

some mathematical knowledge. First we need to compare the close to the high and the low of the 60 minutes time interval. Then we will implement the following formula:

$$\text{Relative Closing Value} = \frac{(\text{Close} - \text{Low}) - (\text{High} - \text{Close})}{\text{High} - \text{Low}}$$

$$\text{AD Value} = (\text{Relative Closing Value} \times \text{Current Volume}) + \text{Previous AD Value}$$

AD value means Accumulation/Distribution Indicators. From this result we can predict the price. When the close is in the upper high position the value will be 1. This means the price will or may increase at some point in near future. If the value is < 1, that means there is high probability of the price getting reduced. We have done a demo prediction which showed us how the value was 0 when the price decreased and how it was 1 when the price increased. The demo chart is inserted below:

time	LTC-USD_close	future	target
1528968720	96.660004	96.389999	0
1528968780	96.570000	96.519997	0
1528968840	96.500000	96.440002	0
1528968900	96.389999	96.470001	1
1528968960	96.519997	96.400002	0
1528969020	96.440002	96.400002	0
1528969080	96.470001	96.400002	0
1528969140	96.400002	96.400002	0
1528969200	96.400002	96.400002	0
1528969260	96.400002	96.449997	1

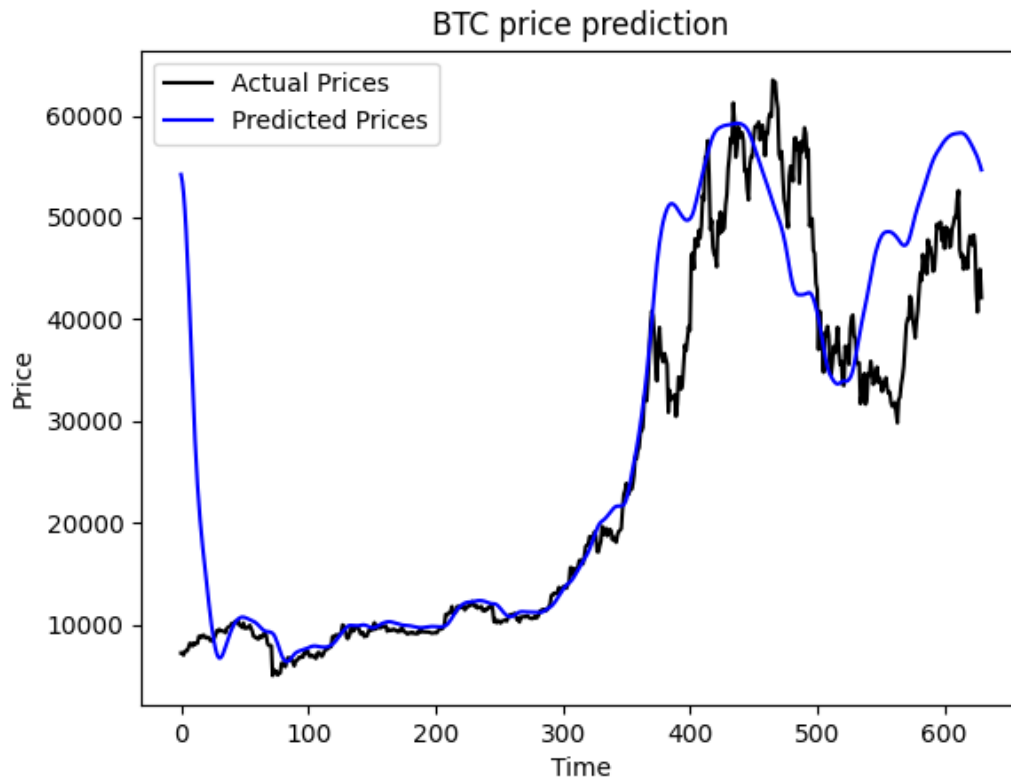
**Figure 5.4:** Predicting price using target value

Here target value is the value we will receive from the calculation. We can see from fig 4 that when the price decreased the target value was 0 and when it increased the target value was 1. So it can tell us a lot about where the price will go in the future.

- Setting Up the Environment:** We used Virtual Studio Code and Python 3.6 to do all our works. For the dataset training and graph modeling we used Excel and Power BI. We had to download libraries like Numpy, Panda etc to get our desired result. We had the help of python 3.9 to get our final graph and we used the inbuilt terminal of VS Code to get the result tables.

## 5.2 Model Result for the Model We Used:

The model we used is RNN model and it gave us 90.49% accuracy at best which is quite great for the market of cryptocurrency. We used only one model here which is RNN model so here we will talk about the result we got from RNN model.



**Figure 5.5:** Accuracy result using RNN model

From the graph in Fig 5 we can see that at 600th minute we got the close price or the final outcome. We drew a straight line from the peak of both graphical curved line to spectate the price range easily. From that we can see that the actual close price was almost 53000 and the predicted price was around 59000. While we run the code we noticed that through every epoch the accuracy level was nearly the same. For example in the first epoch the accuracy was 0.9216 and on the second epoch it was 0.8713 which is pretty close. So when we calculated the average of the accuracy rate from the epochs it was 90.49 percent.

	Time	Price
Actual	600	53000
Prediction	600	59000

**Figure 5.6:** Accuracy result using RNN model

In the above diagram we have shown the matrix for the RNN model graph. Between the ranges of 50000 to 60000 the RNN model could predict data accuracy of more than 200 minutes in average and the average high and low curve was predicted accurately with TP or True Positive. The FP or False Positive occurred between the time of 350 to 400 minutes and the ranges of 40000 to 50000 where the value was actually going down but the prediction was such that the value will go up. So here we witnessed FP in the RNN model. But in the end the close value was similar so we can say that in the matrix the left side has the time value which is same for both attributes and the right side has the price which is 89.83 or 90% accurate.

### 5.3 Accuracy Analysis for Each Models

For accuracy analysis we used two methods here. One is using True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The formula of finding accuracy this way is:

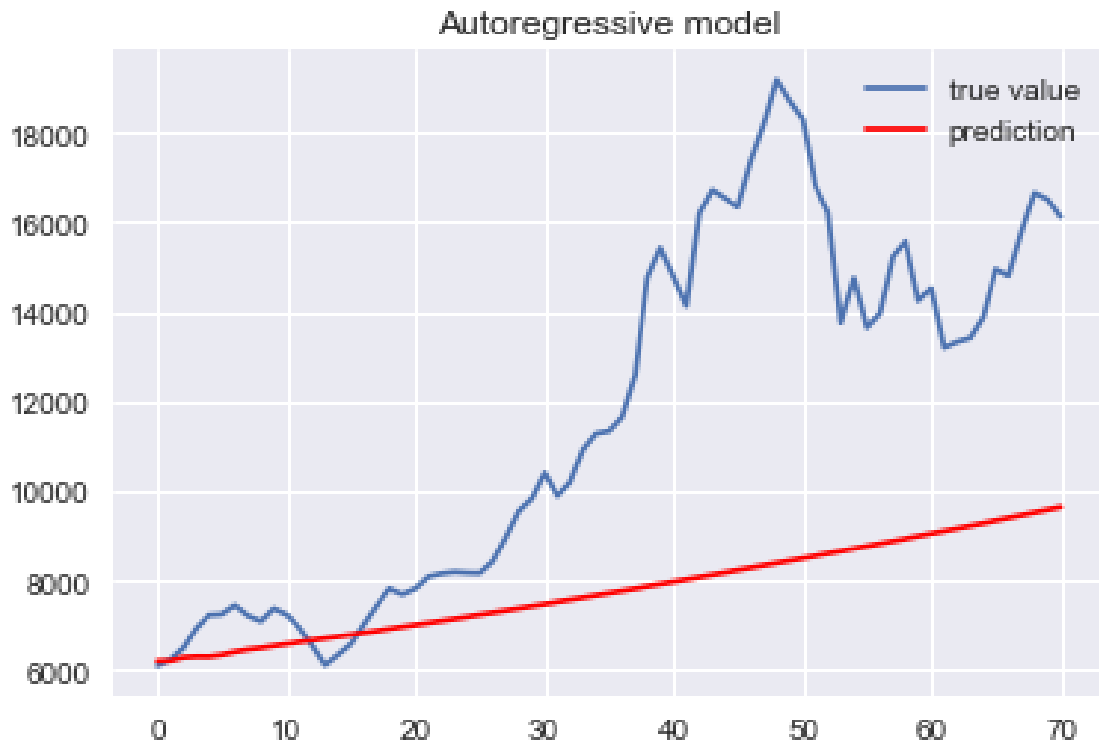
$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

After running our dataset we found the accuracy value = 0.9049

Multiplying this with 100 gives us =  $0.9049 \times 100 = 90.49$

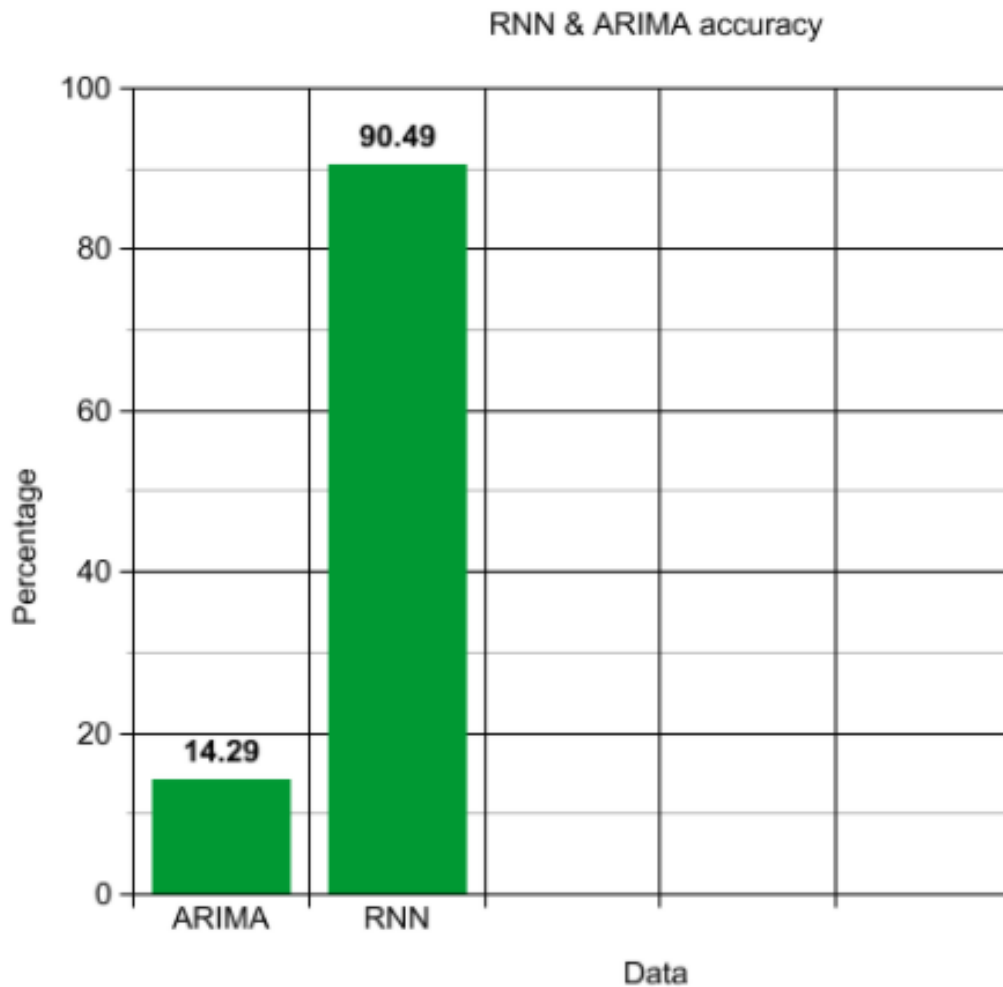
So after calculation we got 90.49% accuracy from the RNN model in the last 60 minutes.

Moving to our previously used ARIMA model, we know that if we divide the number of current prediction by the number of total prediction we will find the accuracy. We total 70 sets of data from random sources and only the range of 10-20 among all the 7 of them was where 2 predictions were correct. So the accuracy was only 14.29% accurate.



**Figure 5.7:** Graph of ARIMA model

So if we compare the accuracy and form it into a graph that will look like the following chart:

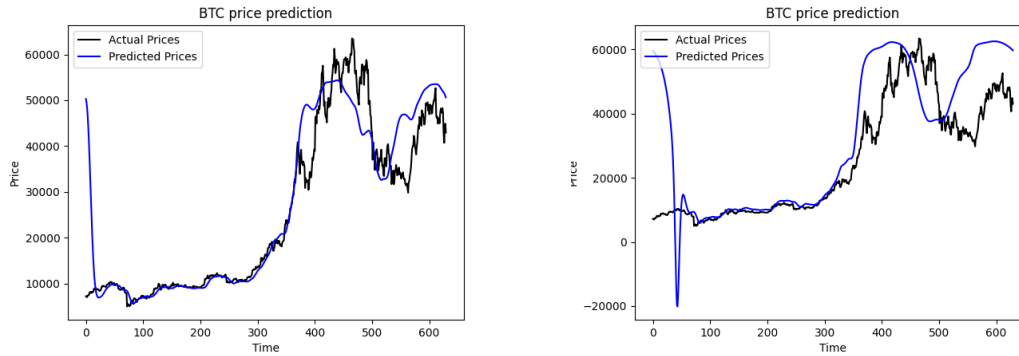


**Figure 5.8:** Graph of ARIMA model

Also, If we increase the time duration from 60 days to 90 days, we will see differences in data accuracy and prediction. We changed the time duration and when we keep the graphs side by side we can see clear difference between these two sets.

When we change duration of previously taken data, we also increased the prediction future range higher. First it was valid for 30 days. But after increasing the duration the prediction was valid for the next 60 days.





**Figure 5.9:** On the left, we took data from previous 60 days, on the right we took from previous 90 days.

## 5.4 Future Plan and Conclusion :

The model implemented in this paper have a lot of possibilities for the future to be used in different applications. Now RNN model uses one internal memory to save previous data .In this way it does not forget its previous works . If this model uses more than one internal memories to store data , the capacity of this model will be increased .So, we will be able to work for larger dataset in future with this model and we will get prediction for longer time .For example ,if we get predictions for seven days now using this model ,in future it will be capable of giving us predictable data of one month or more than that.Cryptocurrency has here, and this is the dawn of a new era. This is the domain and central location for everything coming forward, therefore we'd best get ready to use cryptocurrency in our daily lives. Although there were some reservations about cryptocurrencies and its use at first, the recent surge in public attention has made it the next big thing. Furthermore, the involvement of JP Morgan and Amazon will only add to the intrigue, and with great power comes great responsibility, so more in-depth studies should be conducted on it to help us understand the deep underlying methods of this enigma, and our attempt to predict its prices will only be a steroid. However, the volatility nature of this market, as well as such heinous scams as "Ponzi Schemes," make things impossible to comprehend, let alone anticipate. As a result, we have entered a new era in which Cryptocurrency is the ultimate game changer, and our efforts to forecast its unpredictability will continue. Cryptocurrency is a digital asset and a medium of exchange that is becoming increasingly popular in the financial world. People desire a secure financial system, or at the very least one that reduces the amount of money they lose after taking a risk. As a result, forecasting the price of cryptocurrencies will allay people's fears about investing or trading in various cryptocurrencies. We will be able to obtain more precise data from people using social media data mining and epidemic modeling, allowing us to complete our paper more correctly. As a result, this research will aid in the addition of another layer of prediction to the previously existing prediction process.

## References

- [1] M. Matta, I. Lunesu, and M. Marchesi, “Bitcoin spread prediction using social and web search media.,” in *UMAP workshops*, 2015, pp. 1–10.
- [2] R. C. Phillips and D. Gorse, “Predicting cryptocurrency price bubbles using social media data and epidemic modelling,” in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–7. DOI: 10.1109/SSCI.2017.8280809.
- [3] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, “Cryptocurrency price prediction using tweet volumes and sentiment analysis,” *SMU Data Science Review*, vol. 1, no. 3, p. 1, 2018.
- [4] J. M. Dias, K. Wegrzyn-Wolska, I. Rached, H. González-Vélez, R. Senkerik, C. Pop, T. Cioara, I. Salomie, and A. Bracciali, “Forecasting cryptocurrency value by sentiment analysis: An hpc-oriented survey of the state-of-the-art in the cloud era,” *High-Performance Modelling and Simulation for Big Data Applications*, p. 325, 2019.
- [5] A. Gonfalonieri, “How to build a data set for your machine learning project,” *Tech. Rep.*, 2019.
- [6] J. Koodziej and H. González-Vélez, *High-performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 CHiPSet*. Springer Nature, 2019.
- [7] S. Lahmiri and S. Bekiros, “Cryptocurrency forecasting with deep learning chaotic neural networks,” *Chaos, Solitons & Fractals*, vol. 118, pp. 35–40, 2019.
- [8] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, “Price movement prediction of cryptocurrencies using sentiment analysis and machine learning,” *Entropy*, vol. 21, no. 6, p. 589, 2019.
- [9] S. Anunaya, *Data preprocessing in data mining -a hands on guide*, Aug. 2021. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/>.
- [10] L. Y. Data, *What is a dataset in machine learning*, Apr. 2021. [Online]. Available: <https://labeyourdata.com/articles/what-is-dataset-in-machine-learning>.
- [11] N. Donges, *A guide to rnn: Understanding recurrent neural networks and lstm*, 2021.
- [12] Frankenfield, *Cryptocurrency*, 2021. [Online]. Available: <https://www.investopedia.com/terms/c/cryptocurrency.asp>.
- [13] N. Reiff, *Were there cryptocurrencies before bitcoin?* Sep. 2021. [Online]. Available: <https://www.investopedia.com/tech/were-there-cryptocurrencies-bitcoin/>.
- [14] dataaspirant, *How-recurrent-neural-network-(rnn)-works-dataaspirant*. [Online]. Available: <https://dataaspirant.com/how-recurrent-neural-network-rnn-works/>.
- [15] A. Sala, “Forecasting of the cryptocurrency market through social media sentiment analysis,”

- [16] Simeon-kostadinov, *How-recurrent-neural-networks-work-by-simeon-kostadinov*. [Online]. Available: [https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaf7?source=post\\_internal\\_links](https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaf7?source=post_internal_links).

bibliography/references

12%

SIMILARITY INDEX

7%

INTERNET SOURCES

2%

PUBLICATIONS

10%

STUDENT PAPERS

## PRIMARY SOURCES

1	Submitted to BRAC University Student Paper	1%
2	Submitted to University of Greenwich Student Paper	1%
3	www.machinelearningplus.com Internet Source	1%
4	www.mastersthesiswriting.com Internet Source	1%
5	pythonprogramming.net Internet Source	1%
6	student.unsw.edu.au Internet Source	1%
7	www.upgrad.com Internet Source	1%
8	Submitted to Charotar University of Science And Technology Student Paper	1%
9	labelyourdata.com Internet Source	1%

10	towardsdatascience.com Internet Source	<1%
11	Submitted to Camarines Sur Polytechnic Colleges Student Paper	<1%
12	Submitted to Coventry University Student Paper	<1%
13	Submitted to University of York Student Paper	<1%
14	www.javatpoint.com Internet Source	<1%
15	Razan Abdulhammed, Hassan Musafer, Ali Alessa, Miad Faezipour, Abdelshakour Abuzneid. "Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection", Electronics, 2019 Publication	<1%
16	Submitted to Anadolu University Student Paper	<1%

17 Fair Ibrahim Alarsan, Mamoon Younes. "Best" <1%

12 Submitted to Coventry University <1 %  
Student Paper

13 Submitted to University of York <1 %  
Student Paper

14 www.javatpoint.com <1 %  
Internet Source

15 Razan Abdulhammed, Hassan Musafer, Ali Alessa, Miad Faezipour, Abdelshakour Abuzneid. "Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection", Electronics, 2019  
Publication

16 Submitted to Anadolu University <1 %  
Student Paper

17 Fajr Ibrahim Alarsan, Mamoon Younes. "Best Selection of Generative Adversarial Networks Hyper-Parameters Using Genetic Algorithm", SN Computer Science, 2021  
Publication

18 www.nature.com

Internet Source <1 %

19 Submitted to Gazi University <1 %  
Student Paper

20 Submitted to University of Leicester <1 %  
Student Paper

21 ojs.kmutnb.ac.th <1 %  
Internet Source

22 Submitted to NIT Imphal <1 %  
Student Paper

23 dokumen.pub <1 %  
Internet Source

24 files.eric.ed.gov <1 %  
Internet Source

25 "Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2018  
Publication

26 "Progress in Artificial Intelligence", Springer Science and Business Media LLC, 2019  
Publication

Exclude quotes On

Exclude matches Off

Exclude bibliography On