REAL TIME CRYPTOCURRENCY PRICE PREDICTION USING SENTIMENT ANALYSIS AND MACHINE LEARNING

FATIN NABILAH BINTI ABDUL RAMAN

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2021

REAL TIME CRYPTOCURRENCY PRICE PREDICTION USING SENTIMENT ANALYSIS AND MACHINE LEARNING

FATIN NABILAH BINTI ABDUL RAMAN

RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF DATA SCIENCE

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2021

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Fatin Nabilah Bt Abdul Raman (I.C/Passport No: 930820045492)

Matric No: 17201487

Name of Degree: Master of Data Science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Real Time Cryptocurrency Price Prediction Using Sentiment Analysis And Machine Learning

Field of Study: Data Science

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 17/6/2021

Subscribed and solemnly declared before,

Witness's Signature Digitally signed by Erma Rahayu DN: cn=Erma Rahayu, o=Universiti Malaya, ou, email=erma@um.edu.my, c=MY Date: 2021.06.18 12:15:47 +08'00' Date: 18/6/2021

Name: ERMA RAHAYU BINTI MOHD FAIZAL ABDULLAH

Designation: SENIOR LECTURER

[REAL TIME CRYPTOCURRENCY PRICE PREDICTION USING SENTIMENT ANALYSIS AND MACHINE LEARNING]

ABSTRACT

Over the past few decades, cryptocurrency has arisen as a critical component of the business and capital market in the financial sector. It has been around for a long time and has now become very popular and widespread. In terms of capital market share, Bitcoin and Ethereum are the top two of the biggest cryptocurrencies. Thus, with the rising volatility in pricing and the growing opportunity for benefit of digital currencies, forecasting the price of cryptocurrencies has become a rather appealing research subject. Several experiments have also been performed to forecast crypto currency values using different machine-learning methods. Therefore, this study uses Twitter and Yahoo data to predict the price of the two cryptocurrency-Bitcoin and Ethereum by using "Autoregressive integrated moving average" (Auto-ARIMA) and "Long short-term memory" (LSTM) model. The performance of the obtained models is critically assessed using statistical indicators like "mean absolute percentage error" (MAPE) and "root mean squared error" (RMSE). The result shows that LSTM model gave better result than Auto-ARIMA model. LSTM gives lower RMSE and MAPE result for Bitcoins and Ethereum data than Auto-ARIMA.

Keywords: Cryptocurrency, Bitcoin, BTC, Ethereum, ETH, Auto-ARIMA, LSTM.

[RAMALAN HARGA MATA WANG KRIPTO MASA NYATA MENGGUNAKAN ANALISIS SENTIMEN DAN PEMBELAJARAN MESIN]

ABSTRAK

Sejak beberapa dekad yang lalu, mata wang kripto telah timbul sebagai komponen kritikal dalam pasaran perniagaan dan modal dalam sektor kewangan.Ia telah lama wujud dan kini telah menjadi sangat popular dan meluas.Dari segi bahagian pasaran modal, Bitcoin dan Ethereum adalah dua mata wang kripto terbesar.Oleh itu, dengan ketidaktentuan yang meningkat dalam harga dan peluang yang semakin meningkat untuk manfaat mata wang digital, meramalkan harga mata wang kripto telah menjadi subjek penyelidikan yang agak menarik.Beberapa eksperimen juga telah dilakukan untuk meramalkan nilai mata wang kripto menggunakan kaedah pembelajaran mesin yang berbeza.Oleh itu, kajian ini menggunakan data Twitter dan Yahoo untuk meramalkan harga dua model wang kripto-Bitcoin dan Ethereum dengan menggunakan model purata bergerak bersepadu Autoregresif (Auto-ARIMA) dan memori jangka pendek (LSTM). Prestasi model yang diperolehi dinilai secara kritikal menggunakan petunjuk statistik seperti min kesilapan peratusan mutlak (MAPE) dan akar bermakna kesilapan persegi (RMSE).Hasilnya menunjukkan model LSTM memberikan hasil yang lebih baik berbanding model Auto-ARIMA. LSTM memberikan hasil RMSE dan MAPE yang lebih rendah untuk data Bitcoins dan Ethereum daripada Auto-ARIMA.

Keywords: wang kripto, Bitcoin, BTC, Ethereum, ETH, Auto-ARIMA, LSTM

ACKNOWLEDGEMENTS

There are many individuals who contributed to this project which contribute through technical support, moral support, advice, or participation.

I am indebted and would like to express my appreciation to my supervisor **Dr Erma Rahayu Binti Mohd Faizal Abdulla**h for her patience, careful supervision, and encouragement throughout the completion of my project. It been both privilege and pleasure to have experience the opportunity to be student under her guidance. I sincerely thank you for being a helpful, supportive, and enthusiastic sort of supervisor that every student needs.

I would like to express my appreciation and sincerely thank you to **En Md Nor Ridzuan Bin Daud**, for the comment and help in providing the feedback during the presentation.

I would like to express my appreciation and sincerely thank you to **Mr Tan Chun Wei** for the comment and help in providing the feedback throughout the completion of my project.

Finally, I would like to thank my family and friends for their understanding especially my sister **Khairunnisa Binti Abdul Raman**, for the encouragement, and support for the completion of my project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

BTC	:	Bitcoin
ETH	:	Ethereum
LSTM	:	Long short-term memory
Auto-Arima	:	Autoregressive integrated moving average
$	:	USD Dollar
RM	:	Ringgit Malaysia
NFT	:	non-fungible token

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Background

The internet, and hence the “Internet of Things” (IOT), literally have brought the world closer to our fingertips and altered our lives during the last 20 years. The nature of the transition is like that seen in the transportation industry. It was introduced with the introduction of diesel engines. Travel has evolved from a carriage to a luxury vehicle because of this. Trade and the economy have changed from isolated local markets to borderless global marketplaces as aircraft and ships transfer people and products across continents and seas. The current transformation brought about by the web, or the web of Things is in the sphere of information and communication. Today's financial transactions and trade are conducted entirely online. With the introduction of cryptocurrencies, a new era has begun in which the monopoly of trade has been phased away, posing an instant challenge to fiat currencies. Interest in bitcoin, a decentralized digital asset created with blockchain technology, has grown dramatically since its inception. This digital currency is gaining popularity due to its volatility, and it is also offering a high-yielding digital trading opportunity (Huang et al., 2021). The market capitalization of cryptocurrencies has increased from $1 billion to $400 billion or RM4.13 billion to RM1651.20 billion over the past decade, and therefore the number continues to grow (Huang et al., 2021). Cryptocurrency has emerged as an important element within the corporate and capital industry within the financial field over the previous couple of decades. it's been around for several years and has now become quite common, widespread. Cryptocurrencies are a digital currency where, unlike the quality cash, cryptocurrency is structured supported cryptography, where transactions are meted out through online transactions (Ferdiansyah et al., 2019).

Bitcoin, the first or primary decentralized cryptocurrency, was published and released on January 3, 2009, as open-source software. After this release, approximately

4000 altcoins (other cryptocurrencies) are released, like Ethereum (ETH), Litecoin, and etc. (Mohapatra et al., 2019). Bitcoin is the most well-known and established digital money. Unlike "normal" currencies, the value of Bitcoin is determined by the difficulty of its computations rather than a tangible item (Stenqvist & Lonno, 2017). In its most basic form, Bitcoin is an open-source software application that runs on a networked computer (node). These nodes work together to form a distributed database blockchain. It functions because the only trusted source of all transactions on the blockchain network. It allows Bitcoin to work in keeping with its original design, touching cryptographic software engineering and economic topics (Judmayer et al., 2017).

The second-largest cryptocurrency, valued at $138 billion or RM569.66 billion at its peak in 2018, could be a decentralized Turing-complete computing platform. Ethereum's live blockchain was initially launched on July 30, 2015 (Harm et al., 2016). Although Ethereum is usually said to be a competitor to Bitcoin's cryptocurrency, it's actually quite a currency. The Ethereum Foundation is explaining that the token, Ethereum, isn't intended to be a currency. it's a by-product of a way larger vision and fuel for manipulating the 'world's computer' (Gerring, 2017).

Ethereum could be a program that permits other Ethereum addresses/actors to be audited and took part in blockchain-based by providing a platform of smart contracts. It enables the decentralized execution of "world computers." Bitcoin promotes the concept of smart contracts, but its use is proscribed to currency trading. The difference between Ethereum and Bitcoin is "the built-in Turing complete programing language that enables anyone to form all-purpose contracts" (Buterin, 2013). those that aren't accustomed to blockchain are likely to own learned of Bitcoin, the technology-using cryptocurrency and payment method. Some analysts expect the opposite cryptocurrency Ethereum, which also uses blockchain, to surpass Bitcoin within the following years.

Bitcoin and Ethereum were a decentralized cryptocurrency that's not regulated anywhere. Bitcoin is exclusive in this its price fluctuates daily and changes daily while Ethereum maintains its price day by day. As of May 2018, the value of those two largest cryptocurrencies, measured by market cap, totaled $160.9 billion or RM664.20 billion. Bitcoin alone accounts for about $115 billion or RM474.72 billion of this value. Given the numerous values of those currencies, some use them as real currencies to seek out value, while others see them as an investment opportunity. As a result, the values of both currencies fluctuated significantly during a short period of your time.

In 2017, the worth of 1 bitcoin increased by 2000%, reaching a high of $17,550 or RM72446.40 on December 11, 2017, from $863 or RM3562.46 on January 9, 2017. Eight weeks later, by February 5, 2018, the value of 1 Bitcoin had halved to $7,9643 or RM328766.30 100 thousand ringgit for its value. On June 28, 2019, the Yahoo Financial securities market rate of exchange for Bitcoin to (USD) was $12354.73 or RM50901.49. from time to time, it continued to rise and fell suddenly in March, with a price of $3900 or RM16068.00. The stock exchange can affect by many uncertain factors, like political issues, affected local or global economic problems. However, it hard to spot or prove these factors.

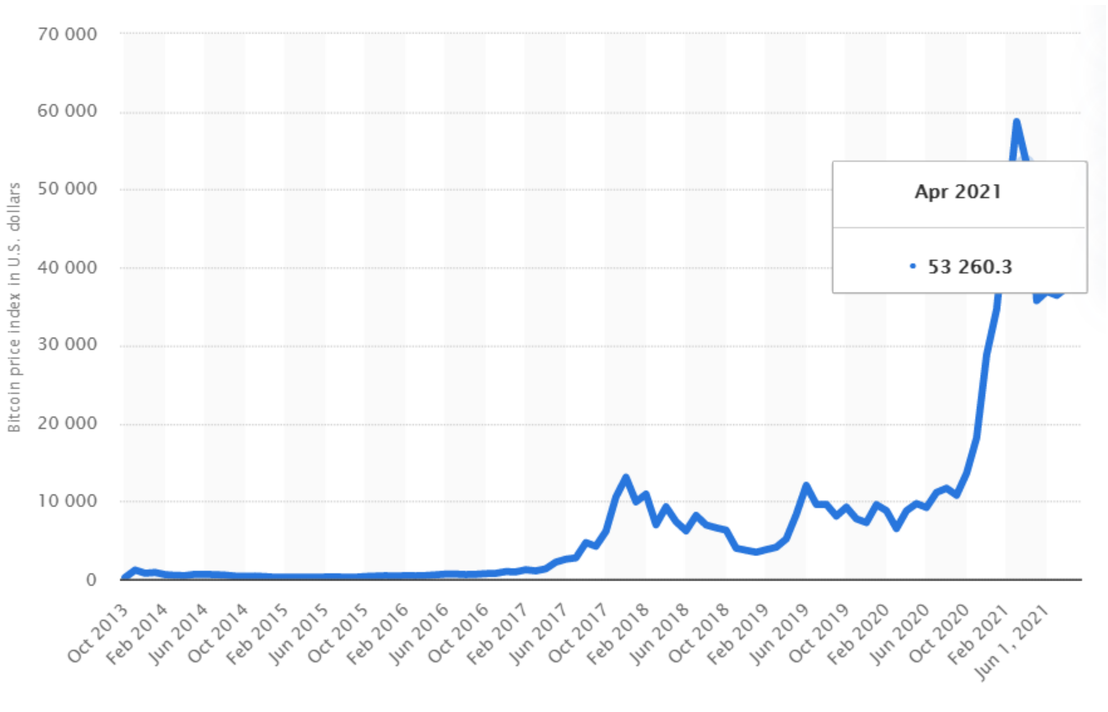

Figure 1.1: Bitcoin Price from Oct-2013 to 1-Jun-2021[1]

Based on Figure 1.1, Bitcoin's (BTC) price rose during April 2021 and peaked at $53260.30 or RM219858.52. In February 2021 and April 2021, Bitcoin (BTC) was valued at over $60,000 each thanks to Tesla and coin-based events. The announcement that Tesla has acquired $1.5 billion worth of digital coins and therefore the nation's largest cryptocurrency exchange IPO sparked public interest. The recent tweet from CEO Elon Musk's on Twitter to own raised cryptocurrency prices like Bitcoin by posting positive messages encouraging more people to shop for digital currencies.

[1] https://www.statista.com/statistics/326707/bitcoin-price-index/

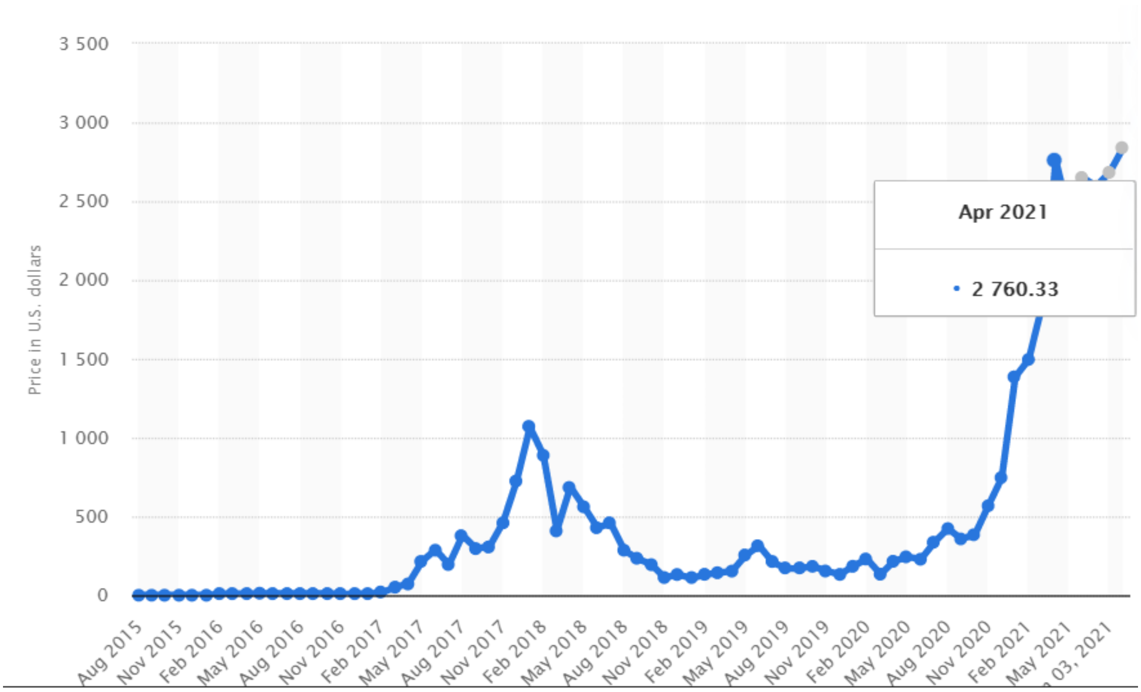

Figure 1.2: Ethereum Price from Aug-2015 to 3-Jun-2021[2]

Meanwhile, the price of Ethereum (ETH) continued to grow in April 2021, reaching a high of $2760.33 (RM11394.64). ETH's value increased in 2021 for a variety of factors, similar to that of Bitcoin (BTC). Ethereum, for example, made headlines when digital art pieces sold for more than 38,000 ETH, or $ 69.3 million or RM286.07 million, as the world's most expensive non-fungible token (NFT). In April 2021, the so-called "Berlin Update" was issued on the Ethernet Rium Network. This change opened the ground for cheaper ETH gas costs and transaction fees. Furthermore, the release of Uniswap V3, a novel contract protocol, in May 2021 was projected to substantially enhance Ethernet trading. These two releases have caused lots of pleasure among traders recently.

On the other hand, the rise of social media platforms such as Twitter, Reddit, and Facebook have made the most recent financial market news and social media posts

[2] https://www.statista.com/statistics/806453/price-of-ethereum/

publicly available. As a result, investors are utilizing a variety of digital tools to make business judgments. Previous study has discovered evidence of a link between fluctuations in share price and social media (Bollen et al., 2011). Negative emotions (such as fear and sadness), neutral emotions (such as calm and uncertainty), or positive emotions (such as confidence and happiness) from cryptocurrency social media content are often wont to predict fluctuations of the value of cryptocurrencies and help in making investment decisions.

Therefore, accurate prediction can help investors in cryptocurrencies make the proper investment decision and result in a possible increase in profit in investing either on Bitcoin or Ethereum (ETH). additionally, this might also help the study of cryptocurrency market behaviors by the present policymakers and financial analysts (Pintelas et al., 2020). Meanwhile, many researchers' attention has been drawn to Twitter to research popular opinion about cryptocurrencies. They are able to also examine the association between price fluctuations and tweet sentiments with the extraction of tweets from Twitter posts (Raju & Tarif, 2020). However, the matter is to search out a technique to predict the price of cryptocurrency that included social media factors. during this project, we are going to specialize in the two cryptocurrencies, namely Bitcoin (BTC) and Ethereum (ETH).

This study presents a method for predicting the value of cryptocurrencies, specifically Bitcoin and Ethereum, using two machine learning models: “Autoregressive integrated moving average” (Auto-ARIMA) and “Long short-term memory” (LSTM). The Auto-ARIMA prediction method was chosen because it is best suited to statistical study of a single variable at several periods in time. LSTM is an efficient sequence prediction approach that uses the value of previous stock prices to reliably forecast future values. To capture and store tweets that mention Bitcoin or Ethereum for sentiment analysis, the Twitter API and a Python package named "Tweepy" are used. These tweets are then examined daily to generate sentiment scores, which are then compared to the day's price fluctuations to see whether a correlation between Twitter sentiment and bitcoin price changes can be established. TextBlob is used to extract sentiment from tweets and the concurrent price data will be extracted or obtained from Yahoo Finance and examines important characteristics which are mapped with the simultaneous price of cryptocurrencies in order to create a forecasting curve that will in the near future predict cryptocurrencies prices.

1.2 Problem Statement

The prediction of cryptocurrency has been proven to be difficult and challenging due to its complexity. Thus, accurate prediction can help investors in cryptocurrencies make the right investment decision and lead to a potential increase in profit. In addition, this may also help the study of cryptocurrency market behaviors by the current policy makers and financial analysts. However, the problem is to find a method to predict the price of cryptocurrency based on social media factors. While analyzing the sentiment data of Twitter and comparing with Yahoo finance data price, we come out with two basic research question to be answered:

1. Is there price pattern between Bitcoin and Ethereum (ETH) price and tweets data?
2. Among these two models, Auto-ARIMA and LSTM model, which model predict the best or accurate result?

1.3 Research Objective

This research is aimed to identify the best model between the two proposed model that would provide the highest possible accuracy for the price prediction. The objective of this research was as the following:

1. to determine the predictable price of Bitcoin and Ethereum (ETH) by sentiment analysis and machine learning.
2. to identify the pattern between Bitcoin and Ethereum (ETH) price and sentiments in tweets.
3. To investigate which existing machine learning (Auto-ARIMA and LSTM) that can predict Bitcoin and Ethereum (ETH) price better.

The rest of this paper is organized as follows. Chapter II gives a background study associated with cryptocurrency and discusses the previous work. Chapter III describes the proposed workflow of this study. Chapter IV for result and discussion of this project. Finally, Chapter V concludes the paper and therefore the limitation of this project.

CHAPTER 2: LITERATURE REVIEW

2.1 Cryptocurrency, Twitter, and Sentiment Analysis Background

Cryptocurrency may be a digital financial asset, and its ownership and transfer registration are secured through encryption technology instead of banks or other trusted third parties. in step with Guidici (2020), it is often considered financial assets because they need a particular value to cryptocurrency holders (discussed below), whether they are doing not represent equivalent liabilities of the other party, nor do they need valuable assets (such as gold).

The term cryptocurrency, yet because the use from "currency" and other terms to "wallets" within the original white book (proposed support technology for Bitcoin) (Nakamoto, 2008), all, implied that the first developers consciously tried to develop a digital device. The transfer mechanism corresponds to the direct transfer of physical cash used for payments or other financial assets (such as precious metals and "bonds"); like cash, it also changes hands through physical transfers (Guidici, 2020).

According to a piece by Nathan Reiff (Reiff, 2019), one amongst the primary attempts to form a cryptocurrency came from Kingdom of The Netherlands within the late 1980s. The filling station suffered theft, and a bunch of individuals tried to link the money with the new open-end credit rather than using cash. At the identical time, American cryptographer David Chaum, who in his article proposed a completely unique encryption scheme that blinded the content of a message before signing it so the signer could not make sure of its content. These blind signatures are publicly verified as normal digital signatures. Chaum proposed a digital cash method that cannot be traced elsewhere (Judmayer et al., 2017). He developed a "blinding formula" to encrypt information which will be transmitted between people. The tool can safely transfer money between people just by verifying the authenticity of the signature. This innovation will play a crucial role within the way forward for cryptocurrency.

Some companies applied these basic principles within the 1990s, and the company with the foremost lasting impact was PayPal. This world-renowned company has modernized individual-to-person online payments. People can transfer money quickly and securely via the web. one in every of the foremost successful applications is egold, which provides users the chance to exchange currency for physical gold or other metals. The cryptocurrency boom began in B-money in 1998. Dai Wei (1998) proposed b-money, an anonymous distributed electronic cash system. during this method, two methods supported an untraceable network are described. A protocol, within which the sender and receiver are only identified by numbers (such as their public keys), and every message are signed by its sender to the receiver.

Bit Gold In 1998, Nick Szabo (2008) proposed a brand new model of digital currency, which is predicated on a puzzle of the cryptographic system, which is solved and sent to the Byzantine fault-tolerant public registry and mapped to the general public key of the solver. by Adam Back, Haschash may be a system that relies on cryptographic hash functions to derive and calculate proof of labor probability as a Pow biometric identification system (Proof of Work) (Back, 1997) and at last RPOW Halfini proposed supported Proof Reusable Workforce 2004 (RPOW) (Finney, 2004). Between 2008 and 2009, Satoshi Nakamoto built Bitcoin because the first decentralized cryptocurrency. Satoshi Nakamoto published the Bitcoin white book in 2008 (Nakamoto, 2008). After January 3, 2009, the genesis block of the Bitcoin protocol was created. Since the launch of Bitcoin, additionally to the over 700 Bitcoin-based altcoins (such as Litcoin, Etherum) circulating within the world, it's the foremost successful cryptocurrency in capitalization today.

In this project, we analyzed the data of the two largest cryptocurrencies within the world by capitalization, the most important is Bitcoin, followed by Ethereum.

2.1.1 Bitcoin

Bitcoin is that the most well liked and established encrypted digital currency. it absolutely was 1st introduced by Satoshi Nakamoto in 2008. Bitcoin was designed to act as a suburbanized payment system that's secure peer to look. Since everything is shown on the blockchain, the general public ledger, you'll be assured that the dealing is legitimate and also the alternative parties have to be compelled to trust is negated (Harm et al., 2016). in contrast to the "normal" currency, Bitcoin's worth isn't a physical product, however it's the quality of the calculation (Judmayer et al., 2017). within the most elementary sense, Bitcoin is associate degree ASCII text file software system program that runs on a network pc (node). These nodes share distributed databases that act as distinctive truth sources for all network transactions and permit Bitcoins to control in step with their original design: encoding, software system engineering, and might work in step with the economic topic (Judmayer et al., 2017). According to Judmayer et al. (2017), Bitcoin's currency is that the commonest application of blockchains, however the blockchains themselves will be utilized in systems that replace the values to not enable the duplication of the assets.

The majority of the world's currencies are issued and governed by governments, either directly or indirectly (ie by a central bank). In all circumstances, the government's aims and rules must be led and adapted to its currency (Franco, 2014). In the case of a central bank, the aforementioned is true despite the central bank's continued direct authority - the bank's independence from the government. The function of central banks is to meet the management organizations' objectives in areas such as economic growth, economic stability, and currency value stability (Franco, 2014).

The value of a currency depends on a number of factors, a remarkable creature; Public trust, acceptance and social expectations (over value) (Franco, 2014). While Fiat, the actual species of real money are dominated by a concentrated goods and an institute, possibly started with the actual value guarantees, which rarely occur in the financial environment (Franco, 2014). Since the Fiat currency is controlled, there are holes in the way the central agency decides to affect a currency. The irrational monetary policy can lead to an artificial long-term reduction through the use of short-term methods (one of which prints money, ie increasing currency supply, but a reduced value) to solve the decision on questions or crises (Franco, 2014). Bitcoin, on the other hand, there is no central agency, and there is no direct way to affect bitcoin values or bitcoin supply (Franco, 2014). According to the design, this will remove the middle-aged man that most money systems are produced around, d. H. Central Bank and Bank System (Franco, 2014). The only way to increase bitcoin offer is to participate in transaction calculations, resulting in predictable growth, which provides bitcoin (Franco, 2014) and is paid for the infrastructure.

At the same time, the currency's currency is impacted by the same factors that affect a Fiat currency (Franco, 2014). The Bitcoin network's design exemplifies this non-centralized approach. Bitcoin is designed to be a decentralized peer network (Franco, 2014), which means that any modifications to the architecture or particular implementation elements must be approved by at least half of the colleagues (Franco, 2014). The distributed database - commonly referred to as a ledger and previously referred to as a blockchain - is a part of the non-centralized architecture. This book contains all of the previous bitcoin transactions as well as all of the current bitcoin owners (Nakamoto, 2008). Temporal transaction blocks are used to build the database. A new block is made by gathering current transactions and then encrypting them with previous blocks, resulting in a block string - block N. This approach allows you to alter or modify a prior block in the string while keeping it secure and transparent (Franco, 2014). In 2008, Nakamoto revealed his Bitcoin design and theoretical work for the first time (Nakamoto, 2008).

2.1.2 Ethereum

In the meanwhile, Ethereum was created to be much more than a payment mechanism. Vitalik Buterin was the first to propose it in 2013. It is a “decentralized platform that executes smart contracts: apps that execute exactly as planned with no downtime, censorship, fraud, or third-party interference.” Ethereum Foundation (Ethereum Foundation, 2016). In general, Ethereum functions similarly to other blockchain-based systems. Because it is built on transactions and uses blockchain as its structure, it is a platform with many features similar to Bitcoin (Buterin, 2014).

Ethereum is frequently referred to be Bitcoin's cryptocurrency competition, but it is much more than that. The Ethereum Foundation said that their token, Ether, is not a money; rather, it is a byproduct of a larger worldview and serves as the fuel for the "global computer" (Ethereum Foundation, 2017). By offering a framework for blockchain-based smart contracts, Ethereum enables the decentralized execution of the "global computer." Smart contracts are programs that can be audited and participated in by other Ethereum addresses/participants. A contract, for example, might encode a fair lottery ticket, which anybody on the network may inspect. To ensure the right execution of the blockchain consensus, the contract's execution is validated by all updated blockchain participants Bitcoin supports the notion of smart contracts as well, although its use is confined to financial transactions. The distinction between Ethereum and Bitcoin is that Ethereum is "a fully integrated Turing programming language that allows anybody to construct contracts for any purpose" (Buterin, 2014).

The Ethereum protocol generates Ether (ETH), a cryptocurrency. It is used to add blocks to the blockchain as a reward for miners in the proof-of-work method (Buterin, 2014). It is the sole currency accepted for payment of transaction fees and may also be used by miners. Ordinary users, for example, exchange value (money) with one another in a transaction. Special users' group and verify various transactions (also called miners). These transaction groupings, known as blocks, are linked in a chain to form a blockchain (Buterin, 2014). According to Buterin, the Ethereum blockchain is a one-of-a-kind shared state transaction machine architecture (2014). As a result, Ethereum is a generalization of this model (Wood, 2014). Ethereum is the first blockchain-based platform to use smart contract technology to create a full distributed consensus Turing machine (Wood, 2014). The Ethereum platform is comprised of a decentralized virtual machine (referred to as the Ethereum Virtual Machine (EVM)) that executes smart contracts (Wood, 2014).

2.1.3 Twitter

Twitter began in July 2006 as two applications: a social media area (which includes other applications/websites such as Facebook, Instagram, LinkedIn, and others) and a microblog. Microblogs are smaller than blogs and have more frequent updates (Abraham et al., 2018). Twitter, like the characteristics of the microblogging service, allows users to publish brief messages (Pak & Paroubek, 2010). Twitter users can send up to 140 text messages (called "tweets") every day (Abraham et al., 2018). In November 2017, the character limit was increased to 280 characters. It also lets users to insert metadata inline in the text of tweets using # ('hashtag') or @ ('at') (Lundmark et al., 2017; Pak & Paroubek, 2010). This is a '#' sign followed by a continuous string in the same manner as the '@' sign. The two operators have distinct purposes, with the former (#) serving as a sign to signify a specific context and the latter (@) referring to other Twitter users (Kouloumpis et al., 2011). Hashtags place tweets in context, forming a web of comparable data points. It is used to identify and search for tweet topics. This will be used to collect tweets from the data section later.

It has expanded swiftly on Twitter since its inception in 2006. Twitter has 330 million monthly active users, 1.3 billion accounts, 83 percent of world leaders have Twitter accounts, 23 million active Twitter users are bots, not humans, and 500 million tweets are posted everyday as of the first quarter of 2019 (Abraham et al., 2018). Twitter now has 353 million monthly active users. Twitter also allows you to search for and use live streams of tweets based on certain hashtags (Lundmark et al., 2017). Twitter is a concentrated location for publishing (and consuming) material both inside and outside (Lundmark et al., 2017). Some businesses utilize it as a resource to provide other enterprises with extra avenues for connecting with the market (Lundmark et al., 2017). Twitter news, corporate announcements, government communications, personal viewpoints, worldviews, and daily living have all become popular media outlets (Pak & Paroubek, 2010).

Twitter users have created millions of brief messages, some of which have already been tagged with contextual data (Lundmark et al., 2017). Twitter has been gold-mined in a semi-structured form of random data due to the message length limit and hashtag categorization properties of tweets (Kouloumpis et al., 2011). Researchers and other organizations scour Twitter for value, information, and insight on a wide range of topics and fields (Kouloumpis et al., 2011). As a result, Twitter is frequently employed as a source when searching for emotional information (Kouloumpis et al., 2011).

With all these remarkable facts, Twitter can become a highly adaptable data source for what people think about practically any issue. The ability to observe when a tweet was made can provide insight into how that emotion evolves over time. This is thus an excellent resource for gathering text data on topics such as Twitter virtual currency and analyzing the link between that topic and pricing.

2.1.4 Sentiment Analysis

According to Vuleta (2021), around 2.5 gigabytes of data are produced every day. Many of the data are in the form of structured text data, whether they be Tweets, articles published on the Internet, text messaging, email, or another type. Because of the vast amount of unstructured data, "Natural Language Processing" (PNL) has emerged as a field of study and research. NLP refers to a collection of strategies that computers use to analyze and produce text (Algorithmia, 2016). In this study, we employ a collection of natural language processing procedures known colloquially as "sentiment analysis."

The use of text analysis tools to analyze and classify emotions (positive, negative, and neutral) within text data is known as sentiment analysis. For the analysis, a Textblob will be used. The purpose of this research is to use sentiment analysis to evaluate if the gathered tweets are typically favorable or negative in their opinions on cryptocurrency. Furthermore, we wish to distinguish between tweets that convey views (subjective tweets) and tweets that just deliver facts without a positive or negative view (objective tweets).

2.1.4.1 TextBlob

“TextBlob is a Python (2 and 3) library for processing text data. It provides a simple API to dive into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and translation” (Loria, 2018). Below are the available features (Loria, 2018):

- “Noun phrase extraction”
- “Part-of-speech tagging”
- “Sentiment analysis”
- “Classification (Naive Bayes, Decision Tree)”
- “Tokenization (dividing text into words and sentences)”
- “Word and phrase frequencies”

- "Parsing"
- "n-grams"
- "Word inflection (pluralization and singularization) and lemmatization"
- "Spelling correction"
- "Add additional models or languages through extensions"
- "WordNet integration"

2.2 Prediction Model

2.2.1 Autoregressive integrated moving average (Auto-ARIMA)

Auto-ARIMA is a prominent statistical tool for predicting time series. The model uses historical values to forecast future values. The Autoregressive Integral Moving Average (ARIMA) is a linear model that combines the auto-regressive (AR) process, the Moving Average (MA) process, and the integration component to distinguish time series for static process conversion. Equation (2.1)[3], AR model express time series x_t at time t as a linear regression of the previous p observations, ϵ_t is the residual white noise and ϕ_i was real parameter:

$$x_t = a + \sum_{i=1}^{p} \phi_i x_{t-1} + \epsilon_t \qquad (2.1)$$

Equation (2.2)[3], MA model depend on the residual error to predict the next period value. The model helps to adjust the unpredictable events. q^{th} order MA model was defined as below, ϕ_i and a were real parameter:

$$x_t = a - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots \theta_q \epsilon_{t-q} + \epsilon_t \qquad (2.2)$$

---

[3] Source: Chapter 4 (Pal & Prakash ,2017).

ARIMA is an extension of the ARMA model that includes built-in components, which is advantageous when the data is not fixed (Pal & Prakash, 2017). To remove non-stationary ones, ARIMA employs time-series differentiation. ARIMA (p, d, q) denotes the sequence of AR, MA, and different components (Pal & Prakash, 2017). The autocorrelation function (ACF) calculates how a series interacts with itself at various delays. It can aid in determining the moving average numbers of lags (q). The partial autocorrelation function (PACF) is a regression of a time series from its previous lag. In the same way, he can propose a possible autoregressive term order (p). It is also possible to evaluate the best ARIMA model using the Akaike Information Criteria (AIC). After checking the residuals, it is possible to move on to predictive calculation (Pal & Prakash, 2017). This model is renowned for its forecasting accuracy. However, since it is a linear model, ARIMA has some limitations in dealing with nonlinear problems, as it should perform better over shorter forecast periods (Pal & Prakash, 2017). Below are the description of ARIMA:

- **AR** (Autoregression): The dependent relationship between observation the number of lagged observations.
- **I** (Integrated): The use of differencing of raw observations to obtains the time series stationary.
- **MA** (Moving Average): The dependency between observation to obtain the time series stationary.

There are three important components in Auto-ARIMA:

- p: associated with the auto-regressive (AR) aspect of the Model, which incorporates past values to forecast the next value.
- d: associated with the integrated (I) part of the Model, which is related to the order of differencing to apply to a time series.

- q: associated with the moving average (MA) part of the Model, which uses past forecast errors to predict the future value.

This p, d, and q will use during initialization of the Auto-ARIMA model and set to 1 for the started value and the max value to 3.

2.2.2 Long short-term memory (LSTM)

"Long short-term memory", or LSTM, is an advanced version of Recurrent Neural Network (RNN) that is capable of learning long-term dependencies by storing important past information and forgetting the information that is not important. LSTM has been widely used for sequence prediction problems is are proven to be extremely effective. LSTM (Long Short-Term Memory) is another module provided for RNN. LSTM was created by Hochreiter & Schmidhuber (1997), and later developed and popularized by many researchers. Like RNN, LSTM network (LSTM network) is also composed of modules with repetitive consistency. LSTM is an updated version of RNN, the difference lies in the connection between the hidden layers of RNN.

Architecturally, LSTM consists of cells with three gates that regulate the "movement" of information in the network. The LSTM unit has one entry door and two exit doors, namely the forgetting door and the memory door (Brownlee, 2017). The cell evaluates the information and can transmit or block the acquired information depending on the strength of the signal. LSTM is programmed to assess and understand the time period in which early information is stored, while considering what to remember and what to forget (Brownlee, 2017). The held data is also associated with the new entry, allowing it to inherit significant long-term dependencies. Unlike other modeling techniques, the weights are also spread back and forth through the layers. This unique feature ensures that weights are retained throughout processing (Brownlee, 2017). This is important in our research because the previous price of cryptocurrencies is crucial in predicting future prices.

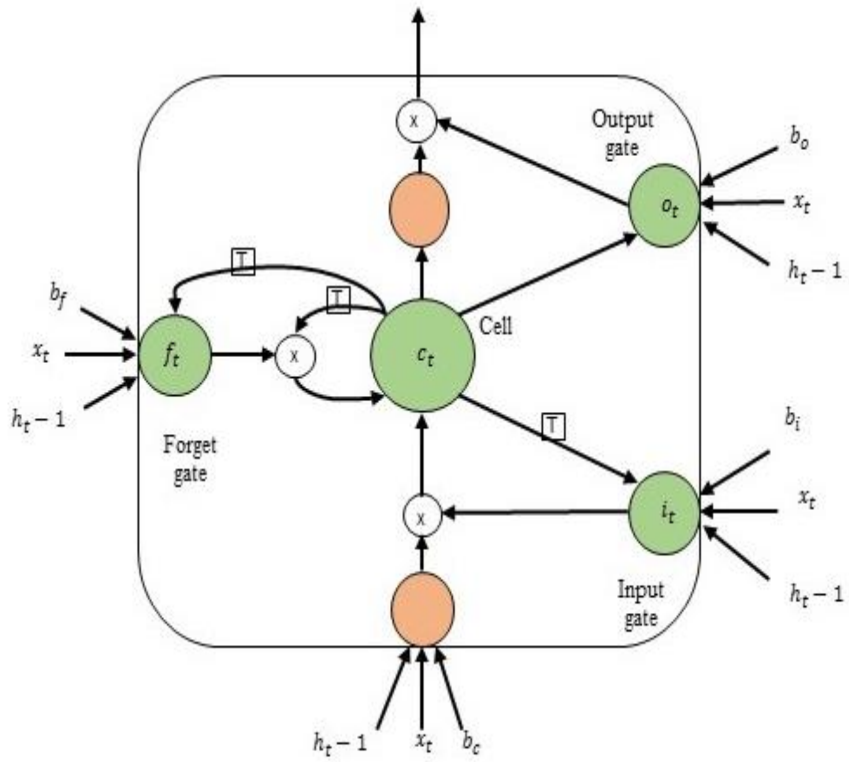

Figure 2.1: memory cell structure LSTM of hidden layer

On the simple way, LSTM is a special type of RNN with additional elements for storing sequence data. The important element of LSTM is the state of the cell that transmits information through the sequence chain. It acts as the memory of the network. The drive status can only carry the relevant information in sequence because the information can be removed or added through gates. The gate will learn what information should be saved or forgotten during training. Therefore, information from earlier stages will now have an impact on later stages in the sequence.

Refer to Figure 2.1 for diagram on the three LSTM gates or LSTM memory cell structure of the hidden layer. The three gates are as below:

- The **Input gate** adds information or data to the cell state

- The **Forget gate** removes or deletes information that is no longer required by the Model
- The **Output gate** selects or chooses the information to be shown as output

2.3 Related Work

Many additional capabilities are available in the bitcoin market that are not available in the regular stock market. Traditional stock trading has its own set of trading hours (for example, 09:00 am to 03:30 pm). In most situations, state statutory holidays or even weekends are not observed by businesses. In contrast, cryptocurrency exchanges are open 24 hours a day, seven days a week, and respond to any incident instantly. As a result, if a person is quick enough, he can earn from fast trading. Keep in mind, however, that this also implies that traders must be online the majority of the time in order to capitalize on possibilities.

In comparison to the typical stock market, Bitcoin and Ethereum, on the other hand, is extremely volatile. Unlike the stock market, the cryptocurrency market is readily manipulated, which means that the price of the cryptocurrency may quickly rise or fall. Finally, bitcoin trading necessitates investors storing the coins themselves, which are vulnerable because novice traders are unsure how to ensure their storage. Because of these qualities, predicting the price of cryptocurrencies is challenging, and there is a rising demand for new analytic methodologies fit for the cryptocurrency market. Aside from that, many academics are conducting research on cryptocurrency prediction, particularly on the Bitcoin price, and other cryptocurrency-related work is difficult to come by.

Jang and Lee (2017) introduced a Bayesian neural network model that predicts bitcoin prices based on blockchain information concerning bitcoin supply and demand and validated that the most recent bitcoin prices had high predictive performance. Their technique, however, is confined to Bitcoin since other cryptocurrencies find it difficult to retrieve information from the blockchain. Ciaian et al. (2016) used a linear model to examine Bitcoin's price formation, which was categorised into numerous market characteristics, attractive to investors, and global macro-financial aspects. They think that the first and second variables described above will have a substantial impact on Bitcoin's price, but that they will fluctuate with time. To assist the study of linear models, the same researchers reduced the number of regressors. For bitcoin price prediction, Karakoyun et al. (2018) compared the ARIMA time series model to the LSTM deep learning algorithm. The conclusion reveals that ARIMA results for approximately 11.86% of MAPE and LSTM results for about 1.40% of MAPE. Hashish et al. (2019) “proposed a new model that aims at predicting Bitcoin prices, based on Hidden Markov Models and Genetic Algorithm-optimized LSTM (HMM-LSTM). The goal is to allow optimized M2M payments in the context of the Internet of Things domain.” It shows that HMM-LSTM has the better result with the lowest MSE, RMSE and MAE between three models for Bitcoin prices prediction.

Kim et al. (2016) used simulated investments to analyze user comments on online social media to forecast bitcoin price and transaction volume and shown that the recommended single dependency estimator (AODE) technique is appropriate for cryptocurrency transactions. Keep in mind, however, that this methodology just analyzes social media and does not use historical price data, which is the most reliable data for forecasting cryptocurrency values in the future. Jiang et al. (2017) proposed a deep reinforcement learning method that directly uses historical cryptocurrency price data collected from the Poloniex API as input to directly generate portfolio vectors to solve the problem of portfolio management. A backtest experiment was conducted in the cryptocurrency market, and the performance was compared with three benchmarks and three other portfolio management algorithms, and positive results were obtained. However, since they proposed a portfolio management method to determine which cryptocurrency requires a large amount of investment, their research is not a single cryptocurrency trend forecast.

According to Chen and Lazer (2013), different machine learning approaches have been developed over the last several decades to anticipate stock market price changes utilizing social media such as Twitter feed comments and data. Bollen et al. (2011) predicted daily rising and decreasing fluctuations in the “Dow Jones Industrial Average” closing price using neural networks and daily Twitter feeds as supplementary predictors. In stock forecasting, Pimprikar et al. (2017) discovered that “Long Short-Term Memory” (LSTM) paired with Twitter sentiment analysis beats existing machine learning models such as Support Vector Machine in stock forecasting.

Greaves and Au (2015) suggested a strategy for predicting bitcoin price that used SVM and Logistic Regression was examined using Graph. They investigated the possibility to forecast Bitcoin’s future price based on the properties of the blockchain network. For transaction analysis, they employ logistic regression algorithms, neural networks, and SVM. Their investigation yielded the following comparative results: linear regression, MSE of 1.94, SVM of 1.98, and baseline of 2.02. The greatest accuracy attained in the classification model is 55.1 % for neural network, 53.4 % for baseline model, 53.4 % for logistic regression, and 53.7 % for SVM.

Karasu et al. (2018) examine and forecast the Bitcoin price. The authors used existing models such as support vector machines (SVMs) with linear regression and kernel functions to collect data for the study on a daily basis. After computing several combinations and analyzing them using measurement metrics such as MSE, MAE, and RMSE Pearson correlations, the authors finalized the model with the least error. The authors came to the conclusion that a polynomial SVM (support vector machine) combined with a 2-day weighted moving average filter produced the best accurate findings.

Abraham et al. (2018) investigated Bitcoin price forecasts using social media data. The author discusses the influence of social network trends like Google and Twitter on decision-making. A vast amount of news on social networks i.e.; Twitter and other news sources can influence people's thinking processes and, as a result, their purchase decisions. Only transaction volume was used by the author. Although the message's sentiment, whether negative or positive, was mapped and determined to be positive, the author did not employ sentiment-based data, but merely transactions based on transaction volume. Following data filtering, a linear noise regression model with a correlation matrix is used for analysis. According to the author, among cryptocurrencies, Bitcoin and Ethereum are associated with social media travel. Prashanth and Vineetha (2018) concentrate on price forecasts for six major cryptocurrencies: Ethereum, Dash, Monera, Litecoin, Ripple, and Bitcoin. For analysis, the author use the LSTM model. According to the study, Bitcoin's prognosis is accurate when compared to other cryptocurrencies, with an MAE of 0.038. Because no additional parameters or models were used in this study, it is hard to say if LSTM produces the best outcomes.

Next, Sin and Wang (2017) investigated the relationship between the closing price of bitcoin one day and the price change the next day. The authors utilized the data obtained and processed using the deep learning algorithms listed below. GASEN is a model of selective neural networks based on artificial neural networks and genetic algorithms. According to the study, the lucrative rate of return is 85%, and the model's accuracy varies from 58% to 63%. Meanwhile, Balcilar (2017) investigated Bitcoin price fluctuations and operation volume. The researchers concluded that although transaction volume and returns are causally related, they did not show the same results as volatility.

The data was collected from the coinmarketcap.com database, starting in January 2015, and lasted for three years and nine months. The high and low opening and closing prices of Bitcoin on the day are plotted. The study reached a precision level of 60% to 70%, with an error of less than 6%. The author (Rane & Dhage, 2019) emphasized the fact that due to the increase in cryptocurrency trading volume, there is a large amount of data available. To complete this large amount of data, the author used various models to find the best method for an accurate price prediction. The author used the following models: binomial generative linear model, support vector machine model, nonlinear autoregressive model, long and short-term memory model, autoregressive moving average model, regression model and multilayer perceptual neural network model for comparative research. They concluded that nonlinear autoregressive model was the best among all the other models with 53% of accuracy.

Garcia & Schweitzer (2015) propose to demonstrate the effect of social cues and trading strategies on Bitcoin using economic signals of trading volume and price in exchange for USD. They analyze social signals considering search, negative and positive aspects of tweets, as well as Bitcoin trading volume. They stated that the increase in sentiment analysis and trading volume takes precedence over the increase in Bitcoin price. They also verify that they are performing well in terms of profitability using robust statistical methods that consider account costa trading and risk analysis.

Galeshchuk et al. (2018) pay attention to behavioral signals and predict Bitcoin's volatility by analyzing users' Twitter comments. They were evaluated both the numerical value-based Bitcoin exchange rate data set and the text-based Twitter data set. Random Walk (RW) and Integrated Moving Average (ARIMA) models are used to predict the exchange rate of Bitcoin, while Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) are used to analyze Bitcoin's sentiment. Twitter users. The author concludes from the research that compared with other methods; the use of CNN has more advantages. The author (Badiola Ramos, 2019) proposes to evaluate the impact of user comments on Twitter to predict the direction of Bitcoin by considering only the sentiment score of each user comment. In Aggarwal et al. (2019), the influence of Twitter sentiment on Bitcoin price prediction was studied by evaluating convolutional neural network (CNN), “long short-term memory” (LSTM), and “Gate Recurrent Units” (GRU). It shows that LSTM resulted with least RMSE with 47.91 for predicting the Bitcoin Price.

Yamak et al. (2019) proposed ARIMA, LSTM and Gated Recurrent Unit (GRU) for time series forecasting for Bicoin price. The researcher shows that the ARIMA model gave better results than the other two models with 2.76% and 302.53 for MAPE and RMSE respectively followed by GRU with 3.97% and 381.34 of MAPE and RMSE respectively. Chandrasekaran (2019) proposed two models Auto-Regressive Integrated Moving Average (ARIMA) and Long short-term Memory (LSTM) to predict Litecoin. The data was collected for five-and-a-half-year period from 2014 to 2019 and will be evaluated with MAPE, ME, MAE, and RMSE performance parameters. It shows that LSTM model predict a better result with a MAPE of 5.759%.

Ahmed et al. (2010) conducted a large-scale comparative study on various machine learning models used for time series forecasting. The prediction results are compared with the prediction results of the ANN, and ARIMA, Zhang’s hybrid models. These models achieve relatively the best prediction precision. (Khashei et al., 2010) also proposed a hybrid artificial neural network model, which uses the “autoregressive integrated moving average” (ARIMA) model to produce a more accurate prediction model than the artificial neural network. At the same time, Bakar et al. (2018) engaged in the study of the use of the weighted moving average method or MA to predict the exchange rate of Bitcoin. They use the calculation of the “mean absolute percentage error" (MAPE) to verify the validity of the prediction model. The result shows that the average absolute percentage error is 0.72%.

Mittal et al. (2018) were proposed an "Automated cryptocurrencies prices prediction using machine learning technique based on the historical trend (daily trend) data" where they will make use of LSTM to analyze the cryptocurrency price for their future work. Wu et al. (2018) aims to summarize the best ways to predict Bitcoin prices on a daily basis. The author used two different models to map and analyze the data collected in seven months by accumulating 208 data sets. There are two models chosen for research, Traditional LSTM and autoregressive LSTM. Both models are measured based on MSE, RMSE, MAE and MAPE. The author concluded that the LSTM autoregressive results are more accurate than the traditional LSTM model. At the same time, Qian & Chen (2019) used the LSTM method to produce a common or standard stock market prediction model which based on a different factor that impacts the market using three stock data with similar trends. It is shown that the LSTM model was performed well for the prediction.

Roy et al. (2018) aims to predict the price of Bitcoin using a time series analysis model. Four years of data was collected for the research. Three models are ARIMA Autoregressive Moving Average, Autoregressive Moving Average (AR) and Moving Average (MA) was used for prediction. The author concludes that the "autoregressive integrated moving average" model gives the best results with an accuracy rate of 90.3%, while the accuracy rates of other models are 89% and 87.58%, respectively. The study forecast price data for ten consecutive days. Saad et al. (2018) focuses on deriving the most accurate model for predicting the price of Bitcoin. The authors investigated the data evaluated within 20 months. The research was use deep learning models and regression: gradient descent, linear regression (LR), and random forest. Different attributes and characteristic estimates of cryptocurrencies have been analyzed. In contrast, the research concluded that the best prediction model was the linear regression model which gave the best results, with an accuracy rate of 99.44%, a MAE of 0.0060 and an RMSE rate of 0.0113.

Meanwhile, McNally et al. (2018) proposed recommended to use deep learning techniques models such as the Recurrent Neural Network (RNN), “Auto-Regressive Integrated Moving Average” (Auto-ARIMA), and “Long Short-Term Memory” (LSTM) to predict Bitcoin prices which originate or obtained from the Bitcoin Price Index (BPI). Among them, LSTM, RNN, and Auto-ARIMA are 6.87%, 5.45%, and 53.4%, respectively. To this end, optimized Bayesian cyclic neural networks and short-term memory networks are used. The highest classification accuracy of LSTM is 52%, and the RMSE is 8%. In addition, ARIMA is used as a time series model to compare the performance of deep learning methods. Experimental results show that the performance of the deep learning model is better than that of ARIMA, which is worse.

In Stenqvist and Lönnö ‘s article “Using Twitter Sentiment Analysis to Predict Bitcoin Price Fluctuations,” the authors describe their process of collecting Bitcoin-related tweets and the period from May 11 and June 11, 2017 Bitcoin price. Non-alphanumeric symbols have been removed from tweets (using “#” and “@” as examples of removal symbols). Tweets that are irrelevant or found to be too influential are removed from the analysis. The author then “uses VADER (Valence Aware Dictionary for Sentiment Reasoning) to analyze the sentiment of each tweet and classify it as negative, neutral, or positive”. In the final analysis, only tweets that can be considered positive or negative are retained (Stenqvist & Lönnö, 2017).

Lamon et al. (2017) use the sentiment of the tweets and news headlines to predict changes in Bitcoin, Ethereum, and Litecoin (one of the many alternative cryptocurrencies currently available on the market). The research found that logistic regression works best to classify these tweets, and they can correctly predict 61.9% of price falls and 43.9% of price increases (Lamon et al., 2017). Colianni et al. (2015) collected tweets from November 15, 2015, to December 3, 2015, and used Support Vector Machine and Naive Bayes to classify tweets, achieving an accuracy improvement of 255. Finally, Shah et al. (2014), using historical prices and Bayesian regression analysis, successfully established a trading strategy.

The research above proposed various methods for the prediction of cryptocurrency which we can see mostly related to Bitcoin price and lesser research for the other cryptocurrency. However, in this project, we will analyze and constructing two model to predict the price of Bitcoin and Ethereum using LSTM and Auto-ARIMA which was the one of the popular models. Table 2.1 Literature Review Summary will provide a brief summary of the selected paper.

Table 2.1: Literature Review Summaries Related to LSTM and ARIMA

No.	Title	Model	Contribution	Limitation
1.	Comparison of ARIMA time series model and LSTM deep learning algorithm for bitcoin price forecasting.	ARIMA and LSTM deep learning algorithm.	Compared the ARIMA time series model to the LSTM deep learning algorithm. The conclusion reveals that ARIMA results for approximately 11.86% of MAPE and LSTM results for about 1.40%	-

			of MAPE. LSTM perform the best.	
2.	Use of machine learning algorithms and twitter sentiment analysis for stock market prediction.	LSTM, Linear Regression, SVM and Neural Networks	implemented LSTM, Linear Regression, SVM and Neural Networks. It concludes that the result 82% for Linear Regression, 60% for SVM, under 0.3% for Neural Network and predicted result of LSTM were closest to the actual values.	-
3.	Cryptocurrency price prediction using long-short term memory model.	LSTM	For analysis, the author uses the LSTM model to predict on price forecasts for six major cryptocurrencies: Ethereum, Dash, Monera, Litecoin, Ripple, and Bitcoin. According to the study, Bitcoin’s prognosis is accurate when compared to other cryptocurrencies,	it is hard to say if LSTM produces the best outcomes

			with an MAE of 0.038.	
4.	Systematic erudition of bitcoin price prediction using machine learning techniques.	Binomial generative linear model, support vector machine model, nonlinear autoregressive model, long and short-term memory model, autoregressive moving average model, regression model and multilayer perceptual neural network model	nonlinear autoregressive resulted as the highest accuracy analysis with 53% accuracy compared to the others model and followed by ARIMA with 52%, LSTM 51.7%.	-
	Bitcoin response to twitter sentiments.	Random Walk (RW), CNN, ARIMA	CNN predict more accurate result which nearest to the actual data with accuracy rate reaches around 95 % on the training data and 68.6% with the test data followed by ARIMA.	-
5.	Deep learning approach to determine the	CNN, LSTM, and "Gate	LSTM resulted with least RMSE with 47.91 for	Require the live dataset input streams of

	impact of socio economic factors on bitcoin price prediction.	Recurrent Units" (GRU)	predicting the Bitcoin Price.	various parameters to improve the bitcoin price prediction.
6.	An artificial neural network (p, d, q) model for timeseries forecasting.	ARIMA, ANN, Zhang hybrid model and author proposed model.	The author proposed model yielded more accurate than the other three model and followed by ARIMA.	-
7.	A new forecasting framework for bitcoin price with LSTM.	Traditional LSTM and autoregressive LSTM	The author concluded that the LSTM autoregressive results are more accurate than the traditional LSTM model.	The result shows Traditional LSTM has some limitations for prediction where it only suitable to predict for price of financial product(rate return).
8.	Stock prediction based on lstm under different stability.	LSTM and ARIMA	The author proposed to prediction using the same dataset with LSTM model and ARIMA model. The result was shown RMSE of LSTM model in this experiment is	LSTM model takes a lot of time to train the model and requires large sample of data.

			0.0374, and the RMSE of ARIMA model in this experiment is 0.10158.The LSTM model performed better with least RMSE result.	
9.	Bitcoin price forecasting using time series analysis.	ARIMA, AR, and MA model	Proposed three models ARIMA Autoregressive Moving Average, Autoregressive Moving Average (AR) and Moving Average (MA) was used for prediction. It shows that ARIMA performs better with 90.3%.	-
10.	Predicting the price of bitcoin using machine learning.	Recurrent Neural Network (RNN), Auto-ARIMA, and LSTM	Proposed these 3 models to predict the Bitcoin price. LSTM resulted least error than RNN, and Auto-ARIMA with 6.87%, 5.45%, and 53.4% respectively.	-
11.	A comparison between arima, lstm, and gru for	ARIMA, LSTM and Gated	proposed ARIMA, LSTM and Gated	-

	time series forecasting.	Recurrent Unit (GRU)	Recurrent Unit (GRU) for time series forecasting for Bicoin price. The researcher shows that the ARIMA model gave better results than the other two models with 2.76% and 302.53 for MAPE and RMSE.	
12	Prediction of Litecoin Prices using ARIMA and LSTM	ARIMA and LSTM	proposed two models Auto-Regressive Integrated Moving Average (ARIMA) and Long short-term Memory (LSTM) to predict Litecoin. It shows that LSTM model predict a better result with a MAPE of 5.759%.	-
13	A hybrid model for bitcoin prices prediction using hidden Markov models and	Hidden Markov Models and Genetic Algorithm-optimized	proposed a new model that aims at predicting Bitcoin prices the HMM-LSTM model which later	-

	optimized LSTM networks.	LSTM (HMM-LSTM),	compared with LSTM, and ARIMA. It shows that HMM-LSTM has the better result with the lowest MSE, RMSE and MAE between three models for Bitcoin prices prediction.	

CHAPTER 3: METHODOLOGY

This research will be based on the CRISP-DM Model or "Cross-Industry Process for Data Mining." Figure 3.1 shows the adapted CRISP-DM Model.

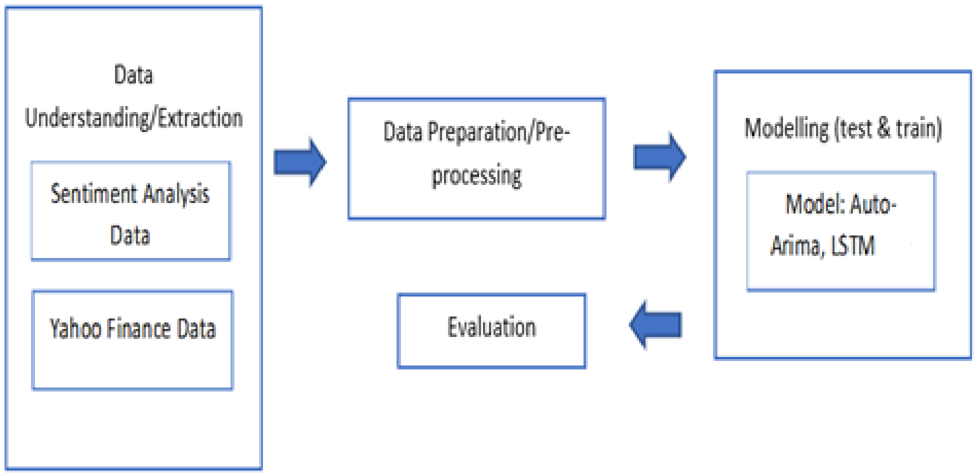

Figure 3.1: Workflow of this study (CRISP-DM)

Figure 3.1 shows the four phases of the CRISP-DM Model: data understanding/extraction, data preparation/preprocessing, modeling, and lastly evaluation. In this research, the Twitter data and yahoo finance data will be used or extracted. Section 3.1 to 3.4 describes in more detail the data understanding/data extraction, data preparation/preprocessing, sentiment analysis, and modeling. The evaluation of the Model will be presented in Section 4.

3.1 Data Understanding/Extraction

Data is a piece of information obtained from real-world measurements. Data retrieval, or data gathering or data extraction in a more sophisticated and technical sense, is the method of obtaining information from different sources. It is a crucial step in any project because, with proper data, the outcome and result can be analyzed. Therefore, several distinct data sources are explored as viable inputs to the model to tackle the challenge of forecasting both cryptocurrencies price fluctuations. The sentiment analysis of collected tweets on Bitcoin or Ethereum is the first input analyzed. The second source of information was Yahoo financial data. This section describes how each of these data sources was collected, cleaned, and changed as needed. Thus, the data for this project is primarily derived from two sources below:

- Twitter data
- yahoo finance data (https://finance.yahoo.com/)
    - For Bitcoin USD price - https://finance.yahoo.com/quote/BTC-USD/
    - For Ethereum USD price - https://finance.yahoo.com/quote/ETH-USD/

3.1.1 Twitter Data

Twitter data is the information collection of user details, the access point, what's in the tweet, and how users view or use the tweet. Twitter data may differ from most other social platform data in that it represents information that people chose to post openly. So, for this project, the Twitter data was collected or extracted from Twitter API using the Tweepy library. Tweepy library will be used and imported in Python. For using the Twitter API, a developer access Twitter account was required. The keys and tokens or the login credentials were required to retrieve the data from Twitter.

Tweets will be extracted or filtered based on the search tag #Bitcoin and #Ethereum as this search tag was a common tag in Twitter, and only the most recent tweets will be extracted. However, Twitter only allows a maximum of 3200 tweets for each extraction, and it also depends on the plan or product package. For this project for 21238 rows for Bitcoin tweets and 5155 for Ethereum search tag tweets with five columns was a success for the extraction during the process. Only those tweets in English language will be filtered. The data was consisting of the 3-month worth of data from February 2021 to April 2021 for this research. The data was later stored into a CSV file named Tweets-BTC.csv and Tweets-ETH.csv.

The Twitter data consists of the following variables with the details Table 3.1 and Table 3.2 sample of 6 rows of Twitter data.

Table 3.1: Twitter Data Description

Variables	Descriptions
user_name	The name of the user, as what the user defined it.
user_location	The user-defined location for the account profile.
Content	The actual UTF-8 text of the tweet.
Date	UTC time and date when the tweet was created.
search_tag	The search tag or search word

Table 3.2: Sample 6 rows for Twitter data for Bitcoin tag

user_name	user_location	Content	date	search_tag
Rahul Chahal		#Bitcoin and #ETH both have bullish setups for a move higher ... #BTC it would just be great if daily close (in aboâ€¦ https://t.co/e5jTbaw43h	5/2/2021 10:53	#Bitcoin
Lion Period with MR.Emre	Los Angeles, CA	$PERL 0.06. I have insisted that since 0.02 it will be 0.071. It increased 300% in about 2 months.	5/2/2021 10:54	#Bitcoin

		#bitcoin #btcâ€¦ https://t.co/cNNkg0SpFV		
TOP AIM STOCKS	United Kingdom	#Bitcoin braces for $48,000 as inverse head-and-shoulders favor #BTC bulls $BTC/USD fades bounce off $36,192 whileâ€¦ https://t.co/8uU0A1rn2w	5/2/2021 10:58	#Bitcoin
Crypto Trader		Bitcoin: $37,412.78 -0.46% (-$174.48) High: $37,916.21 Low: $36,200.10 Volume: 400 $BTC #BTC #bitcoin	5/2/2021 11:00	#Bitcoin
coinOK		1 BTC Price: Bitstamp 37475.23 USD Coinbase USD #btc #bitcoin 2021-02-05 06:01 https://t.co/yY9Q0mIVwW	5/2/2021 11:01	#Bitcoin
EM_CryPT0	Nederland	To-do or not To-do. #crypto #btc #Bitcoin #Ethereum https://t.co/i5qZvYmBEZ	5/2/2021 11:08	#Bitcoin

3.1.2 Yahoo Finance Data

Yahoo finance was a media property that part of Yahoo! Network since 2017 and was owned by Verizon Media. Yahoo Finance offers financial news, data, and commentary, including stock quotes, press releases, financial reports, and original content. For this project, the yahoo finance data was collected or extracted using the finance library. The library will be used and imported in Python. This finance library is a popular open-source library. It allows users to import or download the available data using Python and has some excellent features that make it ideal for stock data analysis.

The Bitcoin data (BTC-USD) and Ethereum data (ETH-USD) will be extracted. For this project, the data will be extracted into two types, one data with a similar date as Twitter data will be extracted, and another data with five years of data from January 2015 to December 2020 will be extracted. The first dataset with a similar date as the Twitter dataset with 31 rows of data for Bitcoin and Ethereum with eight columns was a success for the extraction during the process. The data was consisting of the 3-month worth of data from February 2021 to April 2021 for this research. The data was later stored in a CSV. A similar step will be done to extract the five years of data for Bitcoin and Ethereum, which are later stored into a CSV.

The Yahoo data consists of the following variables with the details in Table 3.3 and Table 3.4 show a sample 6 rows of data.

Table 3.3: Yahoo Data Description

Variables	Description
Date	The date at which the stock is traded
Open	The starting price which stock is traded at that day
High	The maximum price of the stock for that day

Low	The minimum price of the stock for that day
Close	The final price of stock traded for that day
Volume	The number of stock or shares bought or sold on that day
Dividend	The final price that includes the dividend
Stock Splits	The final price that includes the stock splits

Table 3.4: Sample of 6 rows of Bitcoin Yahoo finance data

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
1/2/2021	33114.58	34638.21	32384.23	33537.18	61400400660	0	0
2/2/2021	33533.2	35896.88	33489.22	35510.29	63088585433	0	0
3/2/2021	35510.82	37480.19	35443.98	37472.09	61166818159	0	0
4/2/2021	37475.11	38592.18	36317.5	36926.07	68838074392	0	0
5/2/2021	36931.55	38225.91	36658.76	38144.31	58598066402	0	0
6/2/2021	38138.39	40846.55	38138.39	39266.01	71326033653	0	0

3.2 Data Preparation/Preprocessing

Data processing is important in any project because the effectiveness and cleanliness of the data directly impact the outcome or results. In this part, each collected data (Tweet data and Yahoo data) was preprocessed to ensure that the finding will not impact by some errors such as missing value or duplicate value. All the records with missing values and duplicates will be removed from the dataset. Then, the Tweet data and Yahoo data will be combined and merged into one CSV file after the cleaning.

3.2.1 Tweets Preprocessing

Tweets can be consisting of many words like acronyms, emoticons, and unnecessary records like images and URLs. So, tweets are preprocessed to symbolize accurate feelings of the public. For preprocessing of tweets, there are involved three steps:

1) Removing Twitter Handles (@user)

   Due to privacy issues, the Twitter addresses are now masked as @user. As a result, these Twitter accounts have no detail about the essence of the tweet.
2) Removing the special character (#tag)
3) Remove any hyperlinks
4) Remove emoji

3.2.2 Yahoo data Preprocessing

The Yahoo finance dataset is a historical dataset of the successfully traded stock of the live data. Therefore, it hard to find any null or duplicate data. However, the data will still undergo some preprocessing or cleanup for the preliminary action. Below are some of the preprocessing data that has been done:

- Remove duplicate data.
- Remove null data.
- Remove any missing date, especially for a public holiday.

  There is a very small percentage of missing data points in the data. As for the information, the basis of trading days is conducted during business days only.

3.3 Sentiment Analysis

3.3.1 Tweets Sentiment Analysis

Tweets are graded as Positive (polarity >0), Negative (polarity < 0), or Neutral (polarity =0) depending on their sentiment. The Textblob will automatically transfer the tweet text for sentiment analysis and polarity score for each person's tweet sentiment score.

3.3.2 The Merging of Datasets

After the sentiment value has been extracted, the two different datasets will be merged into a single dataset by using the merge data program, which used Pandas library. The additional features from the sentiment analysis will be merged with the Yahoo price data for prediction. The merged dataset will only consist of price, polarity, date features etc. The sample of 5 rows of the merged dataset can refer in Table 3.5.

Table 3.5: Sample of 5 rows merged dataset

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Polarity
5/2/2021	1594.793	1756.511	1594.793	1718.651	40108628454	0	0	0.059511
6/2/2021	1717.797	1738.314	1649.069	1677.847	39873420648	0	0	0.056411
7/2/2021	1677.606	1690.037	1501.75	1614.228	39889440151	0	0	0.047816
8/2/2021	1613.642	1770.591	1571.58	1746.617	48012285956	0	0	0.0865
9/2/2021	1746.926	1815.964	1711.621	1768.035	44180727529	0	0	0.071601

3.4 Model Implementation

There will be two algorithms that will be used in this research include “Long short-term memory” (LSTM) and “Autoregressive integrated moving average” (Auto-

ARIMA). These two algorithms were very popular algorithms or methods for the forecasting application.

3.4.1 Autoregressive integrated moving average (Auto-ARIMA)

As stated previously in Chapter 2, the Auto-ARIMA (p, d, q) defines the lag observations as the auto-regressive part (*p*), the number of time differencing (*d*) and the moving average numbers of lags (*q*). To select the best suitable Auto-ARIMA model, AIC score for each combination of *p*, *d*, *q* will be observed. Each combination details can be referred in the Figure 3.2 and Figure 3.3. It shows that the least AIC result with Auto-ARIMA (2,1,2)(0,0,0) and Auto-ARIMA (1,1,1)(0,0,0) were the best suitable model for Bitcoin and Ethereum data.

```
Performing stepwise search to minimize aic
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=24787.530, Time=0.51 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=24795.333, Time=0.04 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=24794.840, Time=0.09 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=24794.699, Time=0.13 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=24793.750, Time=0.02 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=24788.890, Time=0.77 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=24788.920, Time=0.97 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=24795.182, Time=0.17 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=24795.475, Time=0.13 sec
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=24780.919, Time=2.28 sec
 ARIMA(3,1,2)(0,0,0)[0] intercept   : AIC=24759.497, Time=4.88 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=24789.106, Time=1.08 sec
 ARIMA(3,1,3)(0,0,0)[0] intercept   : AIC=inf, Time=4.62 sec
 ARIMA(2,1,3)(0,0,0)[0] intercept   : AIC=24759.339, Time=3.45 sec
 ARIMA(1,1,3)(0,0,0)[0] intercept   : AIC=24789.799, Time=0.93 sec
 ARIMA(2,1,3)(0,0,0)[0]             : AIC=24757.501, Time=1.91 sec
 ARIMA(1,1,3)(0,0,0)[0]             : AIC=24788.121, Time=0.54 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=24755.814, Time=1.77 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=24787.234, Time=0.40 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=24787.223, Time=0.27 sec
 ARIMA(3,1,2)(0,0,0)[0]             : AIC=24791.144, Time=0.45 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=24785.850, Time=0.24 sec
 ARIMA(3,1,1)(0,0,0)[0]             : AIC=24787.430, Time=0.57 sec
 ARIMA(3,1,3)(0,0,0)[0]             : AIC=inf, Time=3.63 sec

Best model:  ARIMA(2,1,2)(0,0,0)[0]
Total fit time: 29.845 seconds
```

Figure 3.2: AIC score for Bitcoin data

```
Performing stepwise search to minimize aic
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=13931.539, Time=0.92 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=13934.792, Time=0.08 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=13933.899, Time=0.10 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=13933.866, Time=0.23 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=13932.828, Time=0.06 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=13933.024, Time=1.42 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=13933.044, Time=1.40 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=13935.807, Time=0.38 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=13935.846, Time=0.28 sec
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=13934.205, Time=2.05 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=13929.575, Time=0.35 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=13931.899, Time=0.08 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=13931.933, Time=0.05 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=13931.058, Time=0.75 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=13931.078, Time=0.64 sec
 ARIMA(0,1,2)(0,0,0)[0]             : AIC=13933.841, Time=0.18 sec
 ARIMA(2,1,0)(0,0,0)[0]             : AIC=13933.879, Time=0.11 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=13932.231, Time=0.78 sec
Best model:  ARIMA(1,1,1)(0,0,0)[0]
Total fit time: 9.857 seconds
```

Figure 3.3: AIC score for Ethereum data

The Auto-ARIMA (2,1,2)(0,0,0) and (1,1,1)(0,0,0) model will then use as the baseline for the prediction model for Bitcoin and Ethereum price. The data will be split to 80% for training and 20% for testing. The result will be shown and discuss in the Chapter 4.

3.4.2 Long short-term memory (LSTM)

To be consistent with the Auto-ARIMA algorithm, the data will be split to 80% for training and 20% for testing. Generally, whenever neural network model was used, the data should be normalized or scaled first. To scale data to a specified range, Min-MaxScaler from the sklearn preprocessing library will be used. So, to train the LSTM model, the data need to be transformed into the acceptable shape.

After data transformed, it was a time to create the LSTM architecture. LSTM model should be a sequential model with multiple layers. The first layer was the input layer, where it will be defined a sequence length as similar to the training data 1691. The LSTM layer will be set to 50. Then the dense layer was set to 1. This model will be used for the training and 100 epochs. It is crucial to compile the model. The ADAM optimizer and mean_squared_error will be set during the compilation to reduce the loss on the next evaluation. Refer Figure 3.4 for the configuration of LSTM. The result will be shown and discuss in the Chapter 4.

```
# Build the LSTM model
model = Sequential()
model.add(LSTM(50, return_sequences=True, input_shape=(X_train.shape[1], 1)))
model.add(LSTM(50, return_sequences=False))
model.add(Dense(25))
model.add(Dense(1))

# Compile the model
model.compile(optimizer='adam', loss='mean_squared_error')

# Train the model
history = model.fit(X_train, y_train, batch_size=1, epochs=50)
```

Figure 3.4: LSTM model configuration in python

CHAPTER 4: RESULT & DISCUSSION

4.1 Sentiment Analysis

4.1.1 Bitcoin Tweets

Figure 4.1 below shows the scatter plot result for the Subjectivity and Polarity of the tweets data. It is shown that most tweets were positive polarity and subjective sentences where it more pointed to the center of the plot where the polarity value was between 0-0.25 and subjectivity value between 0.1-0.6. Subjective sentences usually refer to personal thought, emotion, or judgment, whereas objective sentences refer to factual information.

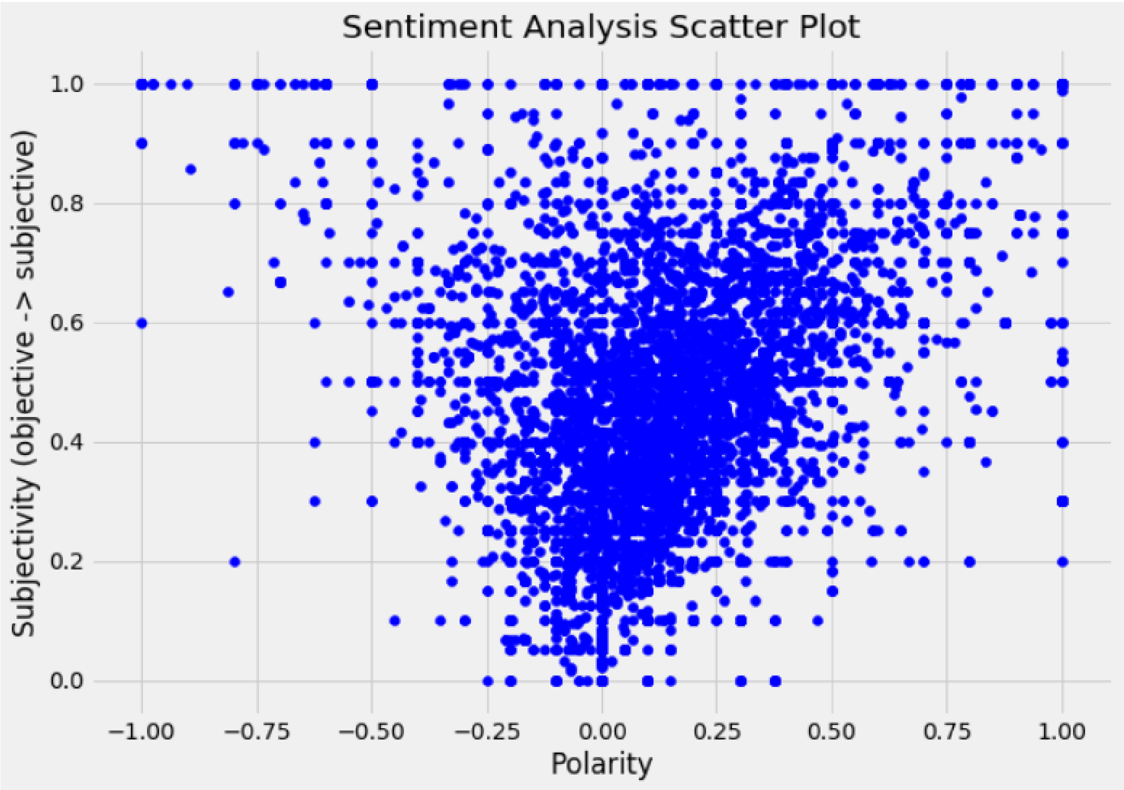

Figure 4.1: Scatter Plot graph for Polarity and Subjectivity for BTC data

Figure 4.2 shown the tweets have a more neutral sentence with 10879 and respectively 8020 positive sentences and 2339 negative sentences. The counts of sentiment can be found in Table 4.1.

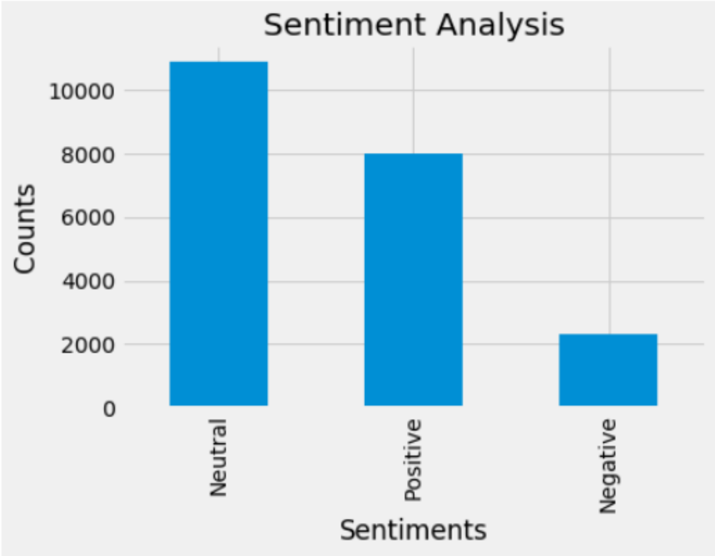

Figure 4.2: Bar graph for Sentiments for BTC data

Table 4.1: Sentiment Counts for BTC

Sentiments	Counts
Neutral	10879
Positive	8020
Negative	2339

4.1.2 Ethereum Tweets

Figure 4.3 below shows the scatter plot result for the Subjectivity and Polarity of the tweets data. The result was quite similar to Bitcoin tweets data. It is shown that most tweets were positive polarity and subjective sentences where it more pointed to the center of the plot where the polarity value was between 0-0.5 and subjectivity value between 0.2-0.6.

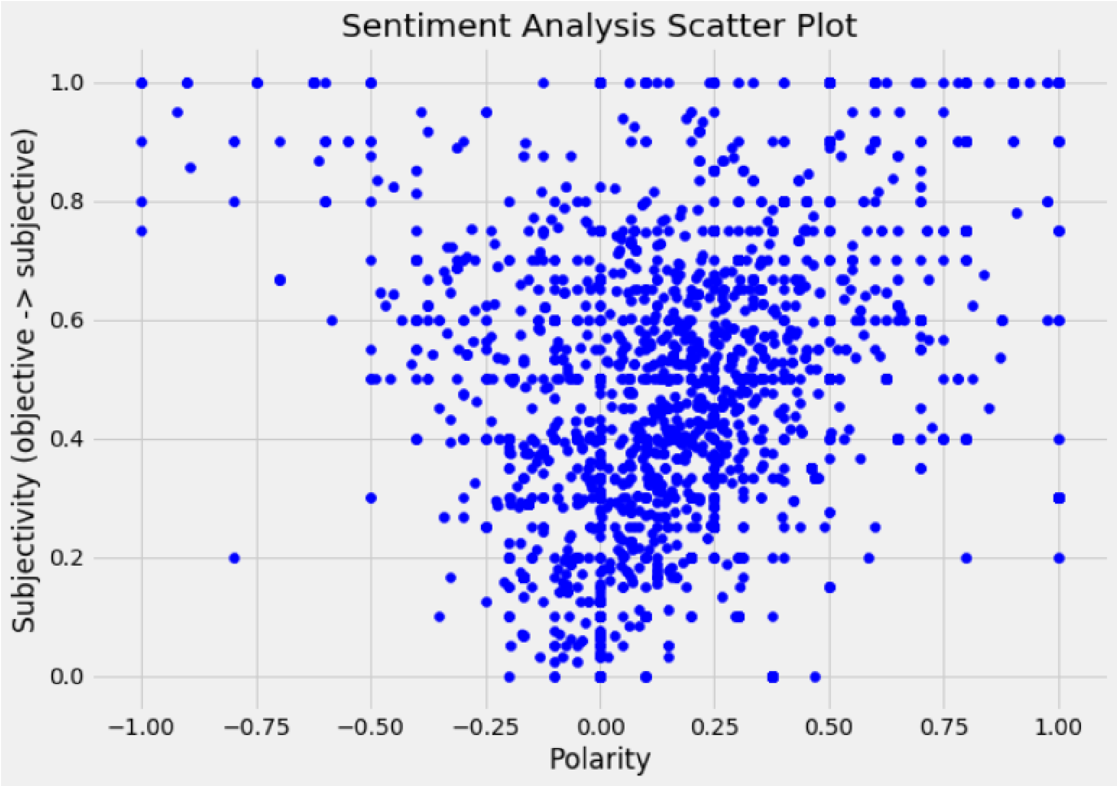

Figure 4.3: Scatter Plot graph for Polarity and Subjectivity for ETH data

Figure 4.4 shown the tweets have a more neutral sentence with 2643 and respectively 1968 positive sentences and 526 negative sentences. The counts of sentiment can be found in Table 4.2.

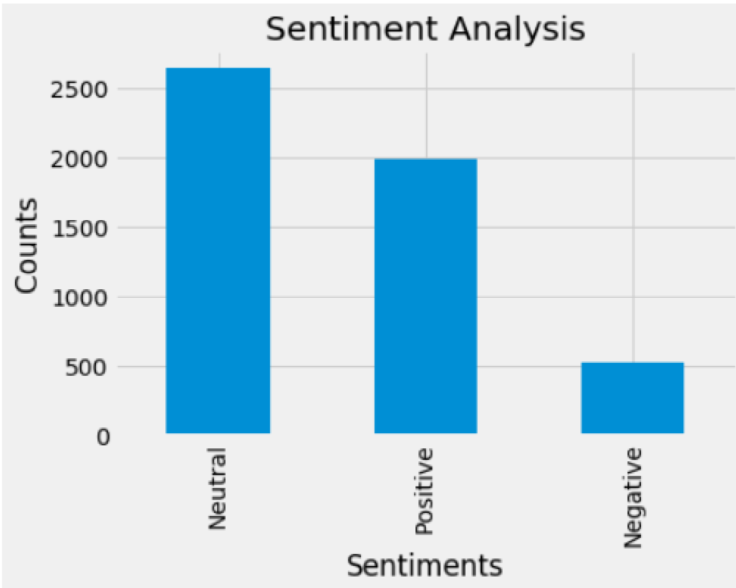

Figure 4.4: Bar graph for Sentiments for ETH data

Table 4.2: Sentiment Counts for ETH

Sentiments	Counts
Neutral	2643
Positive	1986
Negative	526

4.2 Bitcoin and Ethereum Price Prediction

Figure 4.5 depicts the preprocessing results used to load the dataset into the algorithm, followed by the last day's Closing price data for Bitcoin and Ethereum based on the 3-month data before train, test, and forecast the results. As we can see from the graph below shows the differences in daily Closing Price between Bitcoin and Ethereum/Ether (ETH). It shows that Bitcoin was mostly traded from \$30000 to \$60000 (RM123840.00 - RM247680.00) and Ethereum traded below \$5000 (RM 20640.00) between Feb - April 2021.

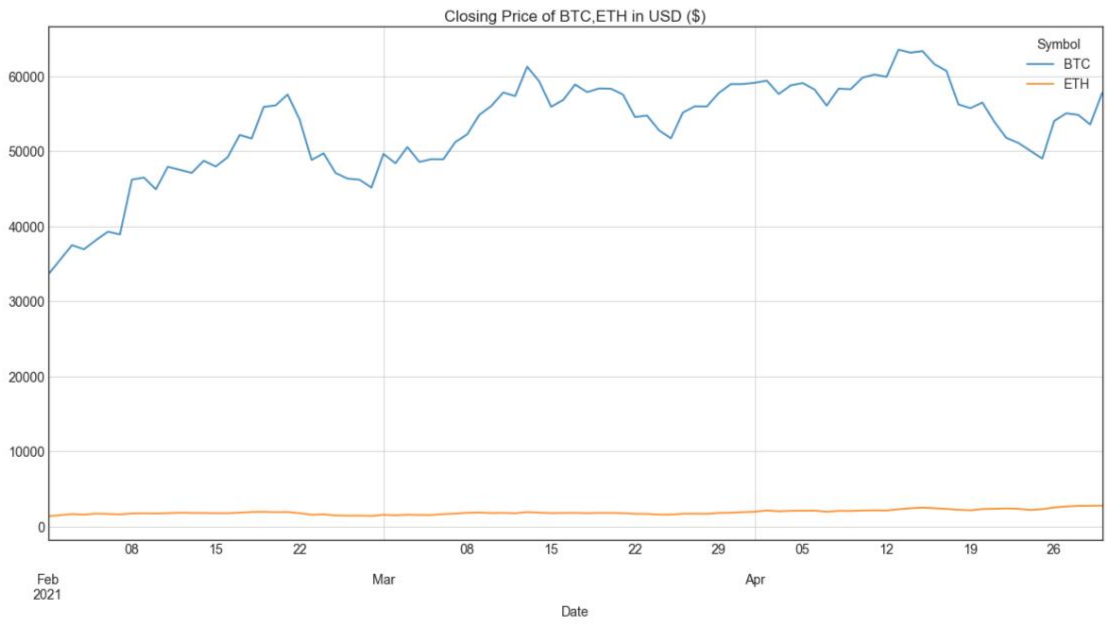

Figure 4.5: Closing Price of BTC and ETH (3-month data)

Meanwhile, for the 5 years data, it shows that Bitcoin was mostly traded from $0 to $30000 or RM123840.00 and Ethereum traded below $5000. Based on Figure 4.6, 2017-2018 was a year of prosperity for Bitcoin. While for Ethereum, in early 2018 Ethereum price was rose as the highest traded price with approximately US$1066.72 or RM4394.89. According to reports, at the beginning of 2016 and beyond, the price almost doubled to approximately US$950 or RM3914.

By mid-2017, the price had finally tripled, reaching near a peak of around US$20000 or RM82400. However, as of mid-2018, the price began to gradually fall for both Bitcoin and Ethereum, as indicated by the blue and orange line on the chart. However, it shows that Bitcoin price rose at the end of 2020 and Ethereum price maintain it price below $5000 or approximately RM 20640.00 and gradually increasing at the end of 2020.

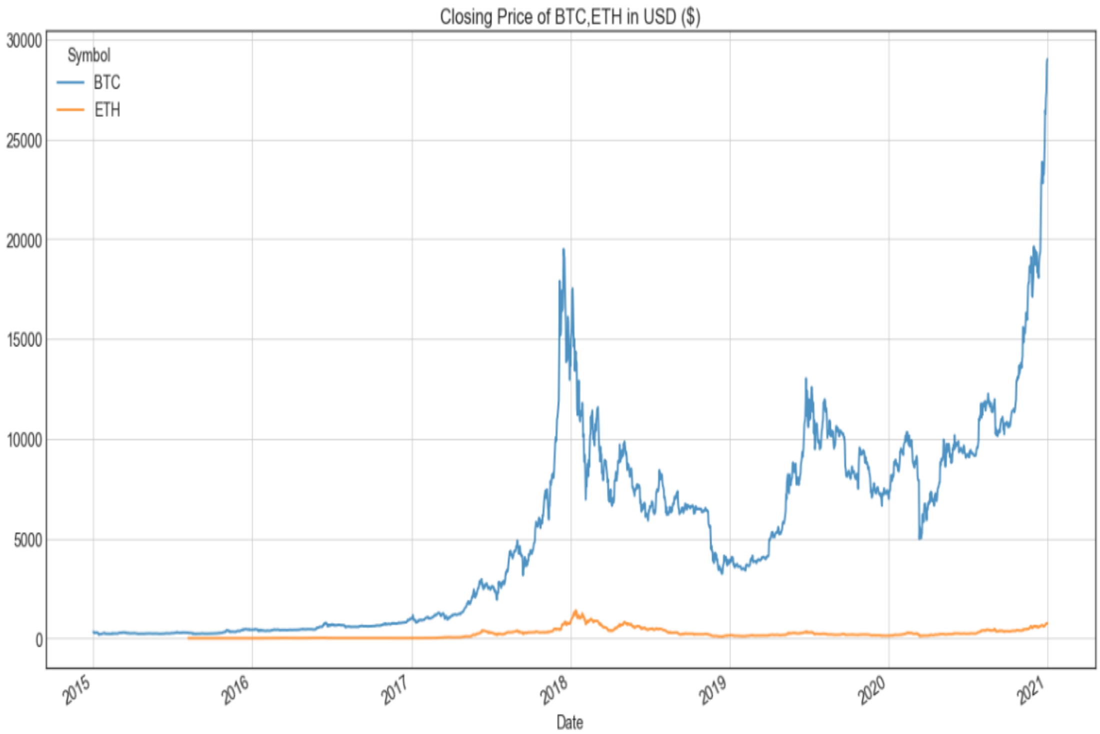

Figure 4.6: Closing Price of BTC and ETH (5 years data)

The five years of Bitcoin and Ethereum price data will be used and extracted for testing and training, and during the final prediction, the closing price of the Twitter data (Figure 4.5) or the merged data will be used.

As we all know, Auto-ARIMA is robust and efficient in time series forecasting, while the LSTM model is very powerful in series forecasting problems. This is due to the important features of using the "remember and forget" gate architecture and convey relevant past information for precise long-term dependencies. In this project, a compatible hardware core AMD Ryzen 5 2500U with Radeon Vega Mobile Gfx 2.00 GHz processor, 16 GB RAM is used. To run the models efficiently, many libraries were installed and run using Anaconda Python. The data was collected over a 5-year period from January 2015 to December 2020, and the 3-month daily dataset provides an effective data set for analysis. Data is preprocessed and cleaning is done with split to train and test to 80% and 20%. The results of the models are compared by different error calculations of MAPE and RMSE to evaluate the accuracies of the results to understand the impact of machine learning models on accurate price predictions. The best-fit model with the least error is also determined.

4.2.1 Analysis result using RMSE and MAPE

The model performance will be evaluated by using “Root Mean Square Error” (RMSE) and “Mean Absolute Percentage Error” (MAPE). RMSE is a square root of the average of squared differences between actual and predicted value. MAPE is an average of the absolute differences between actual and predicted value over the actual value. The n is the number of the time steps. The RMSE and MAPE can be defined as in equation (4.1) and (4.2).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Actual_i - Predicted_i)^2} \qquad (4.1)$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|(Actual_i - Predicted_i)|}{Actual_i} \qquad (4.2)$$

"The RMSE metric evaluates how well a model can predict a continuous value. The RMSE units are the same units as your data's dependent variable/target (so if that's dollars, this is in dollars), which is useful for understanding whether the size of the error is meaningful or not." (Ferdiansyah et al., 2019). Meanwhile, MAPE is useful when analyzing prediction models where only the magnitude of the variance between predicted and observable values is necessary to consider, and the direction of the difference is ignored (Myttenaere et al., 2016). MAPE evaluation overcomes the significant variance bias inherent in Root Mean Square Error (RMSE) and demonstrates robustness for datasets with long tails (Mohan, 2019). The Model's efficiency improves as both RMSE and MAPE result decreases or lesser (Squark, "Glossary of AI Terminology").

For Bitcoin, LSTM obtained RMSE result 165.961514 and MAPE result 3.317423, which was lower or less than Auto-ARIMA RMSE and MAPE results. Similar to Bitcoin, Ethereum LSTM obtained a lower RMSE result of 7.971901 and MAPE result 5.11625 compared to Auto-ARIMA. In Table 4.3 below, we can see that LSTM was the best performance model for Bitcoin and Ethereum which it clearly indicates that LSTM model improved Auto-ARIMA model on average of RMSE and MAPE with 93.37%,85.4% for Bitcoin price and 94.23%, 86.91% for Ethereum price.

Table 4.3: Auto-ARIMA and LSTM RMSE, MAPE results

	Bitcoin Price	Ethereum Price

	RMSE	MAPE	RMSE	MAPE
Auto-ARIMA	2503.411293	22.727595	138.122499	39.095144
LSTM	165.961514	3.317423	7.971901	5.116205
Reduction %	93.37%	85.4%	94.23%	86.915%

The LSTM price prediction graph can be observed in Figure 4.9 and Figure 4.10. Based on Figure 4.9, we can see that the line on the right side has been compared to the actual and the predicted price between the early years of 2020 to the end of the year 2020 or 60-days interval. It is shown that Prediction result for Closing Price of Bitcoin data was closer to Valid data, blue and green line is from Train data.

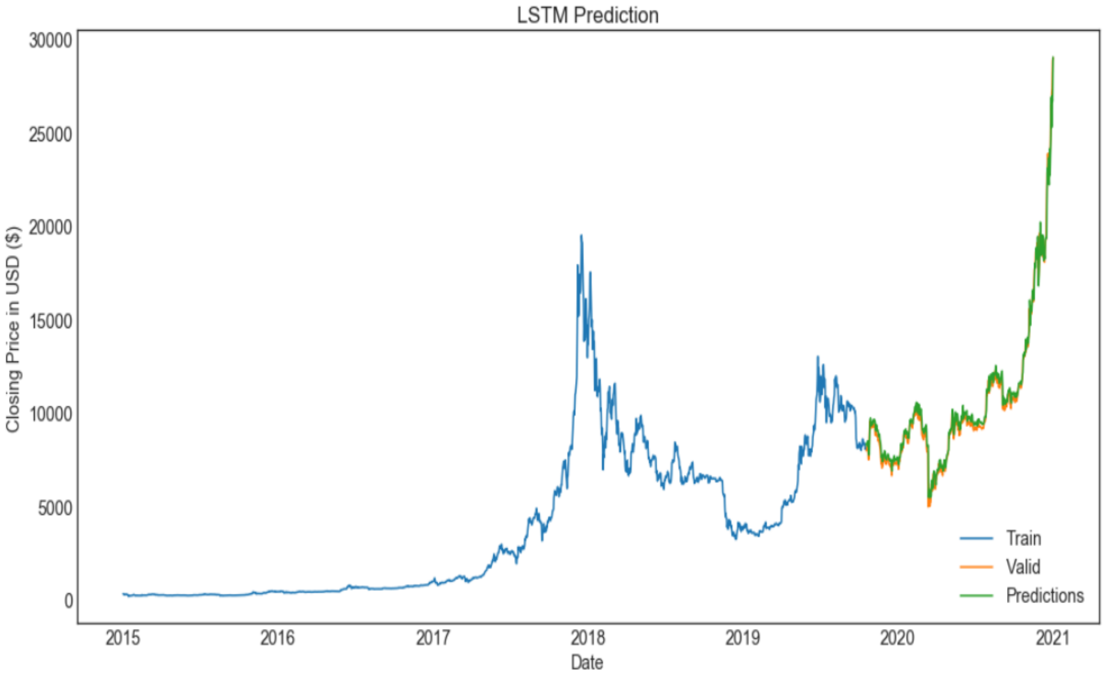

Figure 4.7: LSTM Prediction graph for Bitcoin data

In Figure 4.10, similar to Bitcoin, the line at the right side has been compared to the actual and the predicted price between the early years of 2020 to the end of the year 2020 or 60-days interval. It is shown that the Prediction result for the Closing Price of

Ethereum data was closer to Valid data. It is shown that LSTM predicts a quite accurate prediction result for both Bitcoin and Ethereum.

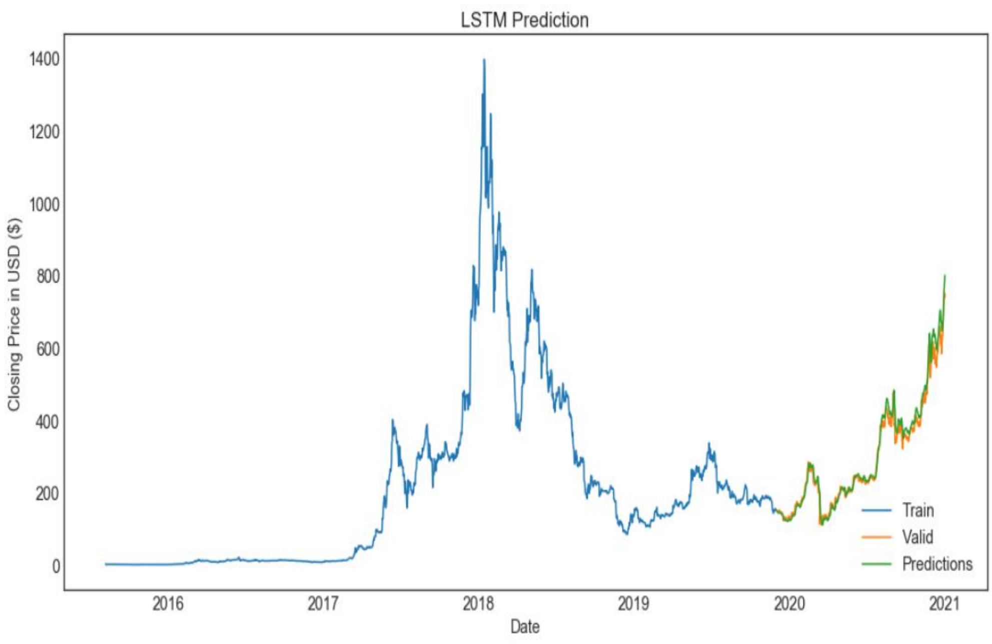

Figure 4.8: LSTM Prediction graph for Ethereum data

4.2.2 Price Prediction

Lastly, we will run the merged dataset to do the prediction. Table 4.4 was the LSTM result for those highest polarity values with more than 0. We will observe the differences and the price changes between the actual and predicted value for the 4 days to identify the prediction pattern. It shows that LSTM predicts 58966.93 which higher from the actual price for Bitcoin, where 2021-04-20 around 2493.89875 differences from the actual price. While for Ethereum that LSTM predicts 2214.7039 with 115.507038 differences than the actual price for 2021-04-20, refer Table 4.4 for the result. It shows that it was not much difference of predicted price for the next day for Ethereum, where the pattern of the price was increasing or decreasing (up or down) for those four days. However, for Bitcoin the next day price, we can see that the predicted price was decreasing and increasing for the next two days which was inconsistent from the actual price. Thus, it shows that even though this date has most positive tweets, the predicted

Bitcoins price still showing decreasing pattern for this 4-consecutive day. It shows the result with merged data did impact on the accuracy of LSTM prediction result.

Table 4.4: Bitcoin and Ethereum Close Price Prediction for future data

Date	Bitcoin Close Price (USD)		Ethereum Close Price (USD)	
	Actual	Predicted	Actual	Predicted
2021-04-20	56473.031250	58966.93	2330.210938	2214.7039
2021-04-21	53906.089844	47710.543	2364.751709	2358.67
2021-04-22	55762.2753438	48396.668	2403.535156	2439.1094
2021-04-23	50050.867188	49859.062	2363.586182	2387.6804

CHAPTER 5: CONCLUSION AND LIMITATION

In this paper, it can be concluded that the LSTM model predicts the Bitcoin and Ethereum prices more accurately compared to Auto-ARIMA. This research focuses on the closing price of Bitcoin, the closing price of Ethereum, and sentiments on the development of prediction models. The sentiments analysis was extracted and listed as one of the features for the price prediction. The Model developed using LSTM was predicted more accurately compared to Auto-ARIMA which it yields least error for both MAPE and RMSE result.

Besides that, we can foresee some pattern or differences of the predicted and actual closing price for the 4 days interval especially for Ethereum Closing Price. LSTM is obviously more effective at learning training data than Auto-ARIMA, and LSTM is better at identifying long-term dependencies. The result for LSTM gives a lower MAPE and RMSE result for Bitcoin; RMSE: 165.961514, MAPE: 3.317423 and Ethereum; RMSE:7.971901, MAPE: 5.116205. Meanwhile, during the prediction of the closing price, LSTM predicted a little lower than the actual value. However, we are still able to identify the pattern of the predicted price. This study uses the daily price fluctuations of Bitcoin and Ethereum to further study the predictability of the Model's hourly price fluctuations in the future. This study only includes the comparison result of MAPE and RMSE between ARIMA and LSTM.

However, we found several limitations when developing this project. This might affect the result of this project. The first limitation was tweets details. The results of our analysis show that in an environment where prices are falling, sentiment analysis is less effective for changes in cryptocurrency prices. This is because tweets about cryptocurrencies are usually subjective (without a clear vision, not fact information) or positive in nature and have nothing to do with price changes. Besides that, mostly the tweets are very short and some of it was meaningless. Thus, this very short and meaningless information has reduced the accuracy of the sentiment classification algorithm and the accuracy of cryptocurrency price prediction. So, there are several other social network option like those Facebook, Reddit that may have a long textual data.

The second limitation was the Twitter API. This Twitter API was used to crawl or extracts the tweets data. However, there is a rate limit per fetch (1500) per hour where it takes around 2-3 days just to extract 10000++ of Bitcoin tweets tag and 1-2 days to extract 4000 of Ethereum tweets tag. It would require a day-by-day extraction to get the more detailed data. Besides that, the extraction of tweet using Twitter API will only consider the most top result.

Lastly, the sentiment analysis model. For this project, we are using Textblob. TextBlob is a simple library that supports complex operations and analysis of text data. It supports more formal language usage instead of slang, emoji etc. Slang, emoji should be removed from dataset and cannot be use for the analysis. So, there are several other options of model can support this limitation like Vader, Flair or even any sentiment analysis tools can be use further analyze the more complex tweets data.

REFERENCES

Pintelas, E., Livieris, I. E., Stavroyiannis, S., Kotsilieris, T., & Pintelas, P. (2020, June). Investigating the Problem of Cryptocurrency Price Prediction: A Deep Learning Approach. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 99-110). Springer, Cham.

Raju, S. M., & Tarif, A. M. (2020). Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis. *arXiv preprint arXiv:2006.14473*.

Bakar, N. A., & Rosbi, S. (2018). Weighted Moving Average of Forecasting Method for Predicting Bitcoin Share Price using High Frequency Data: A Statistical Method in Financial Cryptocurrency Technology. International Journal of Advanced Engineering Research and Science, 5(1), 237361.

Pal, A., & Prakash, P. K. S. (2017). Practical time series analysis: master time series data processing, visualization, and modeling using python. Packt Publishing Ltd.

Hashish, I. A., Forni, F., Andreotti, G., Facchinetti, T., & Darjani, S. (2019, September). A hybrid model for bitcoin prices prediction using hidden Markov models and optimized LSTM networks. In 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (pp. 721-728). IEEE.

Karasu, S., Altan, A., Saraç, Z., & Hacioğlu, R. (2018, May). Prediction of Bitcoin prices with machine learning methods using time series data. In 2018 26th signal processing and communications applications conference (SIU) (pp. 1-4). IEEE.

Chandrasekaran, R. (2019). Prediction of Litecoin Prices using ARIMA and LSTM (Doctoral dissertation, Dublin, National College of Ireland).

Prashanth, J. R., & Vineetha, S. (2018). Cryptocurrency price prediction using long-short term memory model. Int. J. Res. Sci. Innov.(IJRSI) V.

Garcia, D., & Schweitzer, F. (2015). Social signals and algorithmic trading of Bitcoin. Royal Society open science, 2(9), 150288.

Aggarwal, A., Gupta, I., Garg, N., & Goel, A. (2019, August). Deep learning approach to determine the impact of socio economic factors on bitcoin price prediction. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-5). IEEE.

Judmayer, A., Zamyatin, A., Stifter, N., Voyiatzis, A., & Weippl, E. (2017). understanding bitcoin cryptography engineering and economics.

Galeshchuk, S., Vasylchyshyn, O., & Krysovatyy, A. (2018). Bitcoin response to twitter sentiments. In CEUR Workshop Proceedings (pp. 160-168).

Balcilar, M., Bouri, E., Gupta, R., & Roubaud, D. (2017). Can volume predict Bitcoin returns and volatility? A quantiles-based approach. Economic Modelling, 64, 74 -81.

Rane, P. V., & Dhage, S. N. (2019, March). Systematic erudition of bitcoin price prediction using machine learning techniques. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 594-598). IEEE.

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 29(5-6), 594-621.

Brownlee, J. (2017). Long short-term memory networks with python: develop sequence prediction models with deep learning. Machine Learning Mastery.

Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for

Colianni, S., Rosales, S., & Signorotti, M. (2015). Algorithmic trading of cryptocurrency based on Twitter sentiment analysis. CS229 Project, 1-5.

Karakoyun, E. S., & Cibikdiken, A. O. (2018, May). Comparison of ARIMA time series model and LSTM deep learning algorithm for bitcoin price forecasting. In The 13th Multidisciplinary Academic Conference in Prague (Vol. 2018, pp. 171-180).

Ciaian, P., Rajcaniova, M., & Kancs, D. A. (2016). The economics of BitCoin price formation. Applied Economics, 48(19), 1799-1815.

Shah, D., & Zhang, K. (2014, September). Bayesian regression and Bitcoin. In 2014 52nd annual Allerton conference on communication, control, and computing (Allerton) (pp. 409-414). IEEE.

Heo, J. S., Kwon, D. H., Kim, J. B., Han, Y. H., & An, C. H. (2018). Prediction of cryptocurrency price trend using gradient boosting. KIPS Transactions on Software and Data Engineering, 7(10), 387-396.

Madan, I., Saluja, S., & Zhao, A. (2015). Automated bitcoin trading via machine learning algorithms. URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan, 20.

Lamon, C., Nielsen, E., & Redondo, E. (2017). Cryptocurrency price prediction using news and social media sentiment. SMU Data Sci. Rev, 1(3), 1-22.

Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., & Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PloS one, 11(8), e0161197.

Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. arXiv preprint arXiv:1706.10059.

Radityo, A., Munajat, Q., & Budi, I. (2017, October). Prediction of Bitcoin exchange rate to American dollar using artificial neural network methods. In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 433-438). IEEE.

Ferdiansyah, F., Othman, S. H., Radzi, R. Z. R. M., Stiawan, D., Sazaki, Y., & Ependi, U. (2019, October). A LSTM-Method for Bitcoin Price Prediction: A Case Study Yahoo Finance Stock Market. In 2019 International Conference on Electrical Engineering and Computer Science (ICECOS) (pp. 206-210). IEEE.

Mohapatra, S., Ahmed, N., & Alencar, P. (2019, December). KryptoOracle: A Real-Time Cryptocurrency Price Prediction Platform Using Twitter Sentiments. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 5544-5551). IEEE.

McNally, S., Roche, J., & Caton, S. (2018, March). Predicting the price of bitcoin using machine learning. In 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP) (pp. 339-343). IEEE.

Greaves, A., & Au, B. (2015). Using the Bitcoin transaction graph to predict the price of Bitcoin. 2015. Quoted, 3, 22.

Harm, J., Obregon, J., & Stubbendick, J. (2016). Ethereum vs. bitcoin. Economist Case Study, 3-9.

Ethereum Foundation. (2016). Ethereum Homestead Release. Retrieved December 26, 2020 from ethereum.org: https://www.ethereum.org/

Jang, H., & Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. Ieee Access, 6, 5427-5437.

Nakamoto, S., & Bitcoin, A. (2008). A peer-to-peer electronic cash system. Bitcoin. URL: https://bitcoin. org/bitcoin. pdf, 4.

Loria, S. (2018). textblob Documentation. Release 0.15, 2.

Sin, E., & Wang, L. (2017, July). Bitcoin price prediction using ensembles of neural networks. In 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (pp. 666-671). IEEE.

Christidis, K., & Devetsikiotis, M. (2016). Blockchains and smart contracts for the internet of things. Ieee Access, 4, 2292-2303.

Mittal, R., Arora, S., & Bhatia, M. P. S. (2018). Automated cryptocurrencies prices prediction using machine learning. Division of Computer Engineering, Netaji Subhas Institute of Technology, India, 8, 2229-6956.

Wu, C. H., Lu, C. C., Ma, Y. F., & Lu, R. S. (2018, November). A new forecasting framework for bitcoin price with LSTM. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 168-175). IEEE.

Qian, F., & Chen, X. (2019, April). Stock prediction based on lstm under different stability. In 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 483-486). IEEE.

Uras, N., Marchesi, L., Marchesi, M., & Tonelli, R. (2020). Forecasting Bitcoin closing price series using linear regression and neural networks models. arXiv preprint arXiv:2001.01127.

Caamaño-Tubío, R. I., Pérez, J., Ferreiro, S., and Aldegunde, M. (2007). Peripheral serotonin dynamics in the rainbow trout (*Oncorhynchus mykiss*). *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology, 145*(2): 245 255

Cakir, Y., and Strauch, S. M. (2005). Tricaine (MS-222) is a safe anesthetic compound compared to benzocaine and pentobaritol to induce anesthesia in leopard frogs (*Rana pipiens*). *Pharmacological Reports, 57*: 467-474.

Cameron, A. A., Plenderleith, M. B. and Snow, P. J. (1990). Organization of the spinal cord in four species of elasmobranch fishes: cytoarchitecture and distribution of serotonin and selected neuropeptides. *The Journal of Comparative Neurology, 297*: 201-218

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. Neurocomputing, 192, 38-48.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 205-208). IEEE.

Lundmark, L. W., Oh, C., & Verhaal, J. C. (2017). A little Birdie told me: Social media, organizational legitimacy, and underpricing in initial public offerings. Information Systems Frontiers, 19(6), 1407-1422.

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1).

Roy, S., Nanjiba, S., & Chakrabarty, A. (2018, December). Bitcoin price forecasting using time series analysis. In 2018 21st International Conference of Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.

Huang, X., Zhang, W., Huang, Y., Tang, X., Zhang, M., Surbiryala, J., ... & Zhang, J. (2021). LSTM Based Sentiment Analysis for Cryptocurrency Prediction. arXiv preprint arXiv:2103.14804.

Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. Ethereum project yellow paper, 151(2014), 1-32.

Chen, R., & Lazer, M. (2013). Sentiment analysis of twitter feeds for the prediction of stock market movement. stanford edu Retrieved January, 25, 2013.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.

Saad, M., Choi, J., Nyang, D., Kim, J., & Mohaisen, A. (2019). Toward characterizing blockchain-based cryptocurrencies for highly accurate predictions. IEEE Systems Journal, 14(1), 321-332.

Pimprikar, R., Ramachadran, S., & Senthilkumar, K. (2017). Use of machine learning algorithms and twitter sentiment analysis for stock market prediction. Int J Pure Appl Math, 115(6), 521-526.

Stenqvist, E., & Lönnö, J. (2017). Predicting Bitcoin price fluctuation with Twitter sentiment analysis.

Ethereum Foundation. 2017. What is Ether. Retrieved May 21, 2021 from https://www.ethereum.org/ether

Buterin, V. A next generation smart contract & decentralized application platform (2013) whitepaper. Ethereum Foundation.

Buterin, V. (2014). A next-generation smart contract and decentralized application platform. white paper, 3(37).

Vuleta, B. (2021, Jan 28). How Much Data is Created Every Day? Retrieved from https://seedscientific.com/how-much-data-is-created-every-day/

Algorithmia.(2016, Aug 11). What is natural language processing? Introduction to NLP. Retrieved from https://algorithmia.com/blog/introduction-natural-language-processing-nlp

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3), 1.

Franco, P. (2014). The Wiley Finance Series: Understanding Bitcoin: Cryptography, Engineering and Economics (1). Somerset: Wiley.

Yamak, P. T., Yujian, L., & Gadosey, P. K. (2019, December). A comparison between arima, lstm, and gru for time series forecasting. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (pp. 49-55).

Judmayer, A., Stifter, N., Krombholz, K., & Weippl, E. (2017). Blocks and chains: introduction to bitcoin, cryptocurrencies, and their consensus mechanisms. Synthesis Lectures on Information Security, Privacy, & Trust, 9(1), 1-123.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Dai, W. (1998). B-money proposal. White Paper.

Back, A. (1997). Hashcash. 1997. URL: http://www. cypherspace. org/hashcash.

Finney, H. (2004). RPOW—Reusable Proofs of Work. 2004. URL: https://nakamotoinstitute. org/finney/rpow/index. html.

Squark (n.d). Glossary of AI Terminology. Retrieved May 21, 2021 from: https://squarkai.com/glossary/#.XWVCLegzZPY.

APPENDIX

<Appendix A>

<u>“TwitterExtractionData-Tweepy.py”</u>

```
“#import library
import tweepy
import pandas as pd
import csv
from tweepy.cursor import Cursor
#twitter credential
#CAN USE FOR 1 MONTH AS BASIC PLAN FOR TWITTER DEV
consumerKey = "mdmzMBrRM2nNnTdEoN7jZuxmY"
consumerSecret = "3D6YYKcs6kheu1VjqZzkHOu9Fu5RBXKYOG947t0PGykr39btce"
accessToken = "525051045-gOq1C7h2YUbz1gFhRuehci0atHZUPwFC5nwlv0VP"
accessTokenSecret = "eXvSLyGa9OrzS6R1EiZ86nxvHGpP0RsxelqyVWkxZqmYV"
#cre_details = pd.read_csv('Login.csv')
#get Twitter API cendential
#consumerKey = cre_details["key"][0]
#consumerSecret = cre_details["key"][1]
```

```
#accessToken = cre_details["key"][2]

#accessTokenSecret = cre_details["key"][3]

#aunthentication part

authenticate = tweepy.OAuthHandler(consumerKey, consumerSecret)

authenticate.set_access_token(accessToken, accessTokenSecret)

api = tweepy.API(authenticate, wait_on_rate_limit=True)

#collect 2000 data about Bitcoin and Ethereum

first_searchtag = 'Bitcoin'

second_searchtag = 'Ethereum'

tweets_bitcoin = tweepy.Cursor(api.search, q=first_searchtag, lang='en', since= '2020-12-01', tweet_mode= 'extended').items(2000)

tweets_ethereum = tweepy.Cursor(api.search, q=second_searchtag, lang='en', since= '2020-12-01', tweet_mode= 'extended').items(2000)

for status in tweets_bitcoin:

    print(status.text)”
```

<Appendix B>

“Sentiment Analysis with Tweet API and TextBlob.py”

“# This program to analyse the sentiment about crypto(Bitcoin and Ethereum) using twitter extracted data

```
#!pip install TextBlob

#Import libraries

import pandas as pd

import re

import matplotlib.pyplot as plot

from textblob import TextBlob

from datetime import datetime

plot.style.use('fivethirtyeight')

#please ensure it run the correct file

df = pd.read_csv('Tweets - BTC.csv')

#df = pd.read_csv('Tweets - ETH.csv')

date_sr = pd.to_datetime(df['date'].values)

change_format = date_sr.strftime('%d-%m-%Y')

df = df.set_index(pd.to_datetime(change_format))
```

```
#create function to clean the content
def cleanTweetContent(text):
    text = re.sub('#bitcoin', 'bitcoin', text)
    text = re.sub('#Bitcoin', 'Bitcoin', text)
    text = re.sub('#ethereum', 'ethereum', text)
    text = re.sub('#Ethereum', 'Ethereum', text)
    text = re.sub('#[A-Za-z0-9]+', '', text)
    text = re.sub('@[A-Za-z0-9]+', '', text)
    text = re.sub('\\n', '', text)
    text = re.sub('https?:\/\/\S+', '', text)
    return text
#create clean_content column
df['clean_content'] = df['content'].apply(cleanTweetContent)
#get subjectivity
def getSubjectivity(text):
    return TextBlob(text).sentiment.subjectivity
#get polarity
def getPolarity(text):
```

```
    return TextBlob(text).sentiment.polarity

#create subjectivity column

df['Subjectivity'] = df['clean_content'].apply(getSubjectivity)

#create Polarity column

df['Polarity'] = df['clean_content'].apply(getPolarity)

#get negative, neutral and positive sentiments

def getSentiment(score):

    if score < 0:

        return 'Negative'

    elif score == 0:

        return 'Neutral'

    else:

        return 'Positive'

#create Sentiment column

df['Sentiment'] = df['Polarity'].apply(getSentiment)

#scatter plot ot show the subjectivity and polarity of the twitter content

plot.figure(figsize=(10,7))

for i in range(0, df.shape[0]):
```

```
    plot.scatter(df['Polarity'][i],df['Subjectivity'][i],color= 'Blue')
plot.title('Sentiment Analysis Scatter Plot')
plot.xlabel('Polarity')
plot.ylabel('Subjectivity (objective -> subjective)')
plot.show()
#plot bar graph for sentiment count for Bitcoin
plot.title('Sentiment Analysis')
df['Sentiment'].value_counts().plot(kind='bar')
plot.xlabel('Sentiments')
plot.ylabel('Counts')
plot.show()
df['Sentiment'].value_counts()
#plot the sum of polarity for each date for Bitcoin
plot.figure(figsize=(30,8))
plot.title('Sentiment Analysis vs Polarity')
polarityB = df.groupby(df.index).sum()['Polarity']
plot.plot(polarityB.index,polarityB)
#count polarity based on date
```

```
polarity_countB = df.groupby(df.index).count()['Polarity']

#get average sentiment based on date

avg_polarityB = polarityB / polarity_countB

df_cleanB=avg_polarityB

#please ensure to save the correct csv file

#bitcoin = PolarityAvgTweet-BTC.csv

#ethereum = PolarityAvgTweet-ETH.csv

df_cleanB.to_csv('PolarityAvgTweet-BTC.csv')

#df_cleanB.to_csv('PolarityAvgTweet-ETH.csv')

#end”
```

<Appendix C>

“YahooExtraction-yfinance.py”

```
“#import all library

import yfinance as yf

from datetime import datetime

import os

#extract the bitcoin and ethereum data

#1) the data will be based on the twitter data date From Feb 2021 - Apr 2021

#2) the data with 5 years details from January 2015 - December 2020

#1) the data will be based on the twitter data date From Feb 2021 - Apr 2021

#a) Bitcoin USD yahoo finance data

#set start date and end date

sd = datetime(2021, 2, 2)

ed = datetime(2021, 5, 1)



btc = yf.Ticker("BTC-USD")

btc_stockT = btc.history(

    start=sd,
```

```
    end=ed
).reset_index()
#b) Ethereum USD yahoo finance data
#set start date and end date
sd = datetime(2021, 2, 2)
ed = datetime(2021, 5, 1)
eth = yf.Ticker("ETH-USD")
eth_stockT = eth.history(
    start=sd,
    end=ed
).reset_index()
#2) the data with 5 years details from January 2015 - December 2020
#a) Bitcoin USD yahoo finance data
#set start date and end date
sd = datetime(2015, 1, 2)
ed = datetime(2021, 1, 1)
btc = yf.Ticker("BTC-USD")
btc_stockF = btc.history(
```

```
    start=sd,
    end=ed
).reset_index()
#b) Ethereum USD yahoo finance data
#set start date and end date
sd = datetime(2015, 1, 2)
ed = datetime(2021, 1, 1)
eth = yf.Ticker("ETH-USD")
eth_stockF = eth.history(
    start=sd,
    end=ed
).reset_index()
#save bitcoin and ethereum price data to csv
df_cleanBT=btc_stockT
df_cleanET=eth_stockT
df_cleanBF=btc_stockF
df_cleanEF=eth_stockF
os.chdir('C:\\Users\\User\\WQD7002\\data')
```

df_cleanBT.to_csv('BTC2M.csv', index=False)

df_cleanET.to_csv('ETH2M.csv', index=False)

df_cleanBF.to_csv('BTC5Y.csv', index=False)

df_cleanEF.to_csv('ETH5Y.csv', index=False)”

<Appendix D>

<u>“LSTM_Auto-Arima Prediction.py”</u>

```
“# import libraries to ignore warnings

import warnings

warnings.simplefilter(action='ignore', category=FutureWarning)

warnings.filterwarnings('ignore')

# import general libraries

import numpy as np

import pandas as pd

import math

import os

# import libaries for visualisation and define setting

import matplotlib.pyplot as plt

from matplotlib import gridspec

from pandas.plotting import lag_plot

plt.style.use("seaborn-white")

plt.rcParams["figure.figsize"]=20,10

plt.rcParams.update({'font.size': 14})
```

```
# import libraries for time series analysis

from datetime import datetime, date, timedelta

from sorted_months_weekdays import *

from sort_dataframeby_monthorweek import *

from pandas.tseries.holiday import USFederalHolidayCalendar

from pandas.tseries.offsets import CustomBusinessDay

from pandas.tseries.offsets import *

from statsmodels.tsa.stattools import adfuller

from statsmodels.tsa.seasonal import seasonal_decompose

# import libraries for modeling

from pmdarima import auto_arima

from sklearn.preprocessing import MinMaxScaler

from sklearn.metrics import r2_score

from keras.models import Sequential

from keras.layers import Dense, LSTM

# define function to read files:

#do note please change the file name at the 2 place: 1- '.csv' 2- ".csv"

#for 2 month data please use 2M.csv
```

```
#for 5 years data please use 5Y.csv
os.chdir('C:\\Users\\User\\WQD7002\\data')
def read_file(abbr):
    # read multiple .csv files in the directory
    file_list = []
    for file in os.listdir():
        if file.endswith('5Y.csv'):
            df = pd.read_csv(file,)
            df['Symbol'] = file.replace("5Y.csv", "")
            file_list.append(df)
    data = pd.concat(file_list, axis=0, ignore_index=True)
    if abbr == " ":
        data = data
    else:
        data = data[data['Symbol'].isin(abbr)]
    # rename columns
    data.columns = ['Date', 'Open', 'High', 'Low', 'Close', 'Volume', 'Dividends', 'Stock Splits', 'Symbol']
```

```
    return(data)

# read data

crypto_data = read_file(['BTC', 'ETH'])

# set date as index

crypto_data['Date'] = pd.to_datetime(crypto_data['Date'], format='%Y-%m-%d')

crypto_data.set_index('Date', inplace=True)

# subset Close column from crypto_data:

df_close = crypto_data[['Close', 'Symbol']]

# to check: display a random sampling of 5 rows

df_close.sample(5, random_state=0)

# reshape the data frame for visualisation:

df_close_viz = df_close.pivot(columns='Symbol')

df_close_viz.columns = df_close_viz.columns.droplevel(0)

df_close_viz.plot( grid=True, linewidth=2, alpha=0.8)

plt.title('Closing Price of BTC,ETH in USD ($)')

#so from this graph, we can see that the difference of close price between BTC-USD
and ETH-USD
```

```
### Missing Data Analysis

# is there any missing datapoints in the data?

df_close.isna().sum()

# define function to identify the missing trading day(s) for each company:

def missing_dates(data):

    grouped = data.groupby('Symbol')

    output = []

    for key, value in grouped:

        data_dates = value.index.sort_values()

        trading_days = len(data_dates)

        start_date = data_dates[0]

        end_date = data_dates[-1]

        business_dates = pd.date_range(start_date, end_date, freq=BDay())

        total_bdays = len(business_dates)

        dates_missing = pd.to_datetime([item for item in business_dates if item not in data_dates], format='%Y-%m-%d')

        missing_days = len(dates_missing)

        missing_perc = round(missing_days/total_bdays*100.0, 2)
```

```
        data_dict = {'Symbol':key,'Start_date':start_date, 'End_date':end_date, 'Business_days':total_bdays,
                     'Trading_days':trading_days, 'Missing_days':missing_days, 'Missing_percentage':missing_perc,
                     'Missing_dates': dates_missing}
        output.append(data_dict)
        df = pd.DataFrame(output)
        df.sort_values('Start_date').reset_index(drop=True)
    return df
missing_dates(df_close)
# subset data
#Please ensure the correct symbol will be use when
#need to run bitcoin or ethereum
#bitcoin=BTC
#ethereum=ETH
crypto_data = crypto_data[crypto_data['Symbol']=='BTC']
#crypto_data = crypto_data[crypto_data['Symbol']=='ETH']
```

```
data = crypto_data.filter(['Close'])

### Modeling Method 1: Auto-ARIMA

# Get the number of rows to train the model on

training_data_len = math.ceil(len(data) * .8)

# subset training data

train_data = data[0:training_data_len]

# subset testing data

test_data = data[training_data_len:]

# convert each dataset to a numpy array

train, test = train_data.values, test_data.values

# initialize Auto-ARIMA model:

model = auto_arima(train, start_p=1, start_q=1, max_p=3, max_q=3,

            m=12, start_P=0, seasonal=False, d=1, D=1, trace=True,

            error_action='ignore', suppress_warnings=True)

# fit model to the training series

model.fit(train)

# make predictions based on the fitted model

predictions = model.predict(n_periods = test.shape[0])
```

```
# define functions to output metrics for model evaluation:
def root_mean_squared_error(actual, pred):
    rmse = np.sqrt(np.mean(pred-actual)**2)
    return rmse
def mean_absolute_percentage_error(actual, pred):
    actual, pred = np.array(actual), np.array(pred)
    mape = np.mean(np.abs((actual - pred) / actual)) * 100
    return mape
# output model metrics for model evaluation
arima_rmse = root_mean_squared_error(test, predictions)
arima_mape = mean_absolute_percentage_error(test, predictions)
output=[]
output.append({'Model': 'Auto-Arima', 'RMSE': arima_rmse, 'MAPE': arima_mape, 'Predictions': predictions})
arima_metrics = pd.DataFrame(output)
arima_metrics
pred_data = pd.DataFrame(predictions, index=test_data.index, columns=['Prediction'])
```

```
plt.plot(train_data)

plt.plot(test_data)

plt.plot(pred_data)

plt.title('Auto-ARIMA Prediction')

plt.xlabel('Date')

plt.ylabel('Adjusted Closing Price in USD ($)')

plt.legend(['Train', 'Valid', 'Predictions'], loc='lower right')

### Modeling Method 2: Long short-term memory (LSTM)

# convert the dataframe to a numpy array

dataset = data.values

# Get the number of rows to train the model on

training_data_len = math.ceil(len(dataset) * .8)

# scale the data

scaler = MinMaxScaler(feature_range=(0,1))

scaled_data = scaler.fit_transform(dataset)

# create training dataset

# scale the train data
```

```
train_data = scaled_data[0:training_data_len, :]

# append the past-60-day values to X_train dataset and every 61st-day value to y_train

X_train = []

y_train = []

for i in range(60, len(train_data)):

    X_train.append(train_data[i-60:i, 0])

    y_train.append(train_data[i,0])

# convert X_train and y_train to numpy arrays

X_train, y_train = np.array(X_train), np.array(y_train)

# reshape X_train dataset

X_train = np.reshape(X_train, (X_train.shape[0], X_train.shape[1], 1))

# create testing dataset

# scale the test data and create X_test and y_test dataset

test_data = scaled_data[training_data_len-60:,:]

# Create the dataset X_test and y_test

X_test = []

y_test = dataset[training_data_len:,:]

for i in range(60, len(test_data)):
```

```
    X_test.append(test_data[i-60:i, 0])

# convert X_test and y_test to numpy array

X_test = np.array(X_test)

# reshape X_test dataset

X_test = np.reshape(X_test, (X_test.shape[0], X_test.shape[1], 1))

# Build the LSTM model

model = Sequential()

model.add(LSTM(50, return_sequences=True, input_shape=(X_train.shape[1], 1)))

model.add(LSTM(50, return_sequences=False))

model.add(Dense(25))

model.add(Dense(1))

# Compile the model

model.compile(optimizer='adam', loss='mean_squared_error')

# Train the model

history = model.fit(X_train, y_train, batch_size=1, epochs=4)

# Get the predicted price value from the model

predictions = model.predict(X_test)

# Transform the scaled predictions to the actual value
```

```
predictions = scaler.inverse_transform(predictions)

# output model metrics for model evaluation

lstm_rmse = root_mean_squared_error(y_test, predictions)

lstm_mape = mean_absolute_percentage_error(y_test, predictions)

output=[]

output.append({'Model': 'LSTM', 'RMSE': lstm_rmse, 'MAPE': lstm_mape, 'Predictions': predictions})

lstm_metrics = pd.DataFrame(output)

lstm_metrics

# Plot the data

train = data[:training_data_len]

valid = data[training_data_len:]

valid['Predictions'] = predictions

# Visualise the data

plt.figure(figsize=(16,8))

plt.title('LSTM Prediction')

plt.xlabel('Date')

plt.ylabel('Closing Price in USD ($)')
```

```
plt.plot(train['Close'])

plt.plot(valid[['Close', 'Predictions']])

plt.legend(['Train', 'Valid', 'Predictions'], loc='lower right')

plt.show()

## RMSE And MAPE Result <a name="analysis"></a>

# dataframe for model metrics

frames = [arima_metrics, lstm_metrics]

compare_metrics = pd.concat(frames).reset_index(drop=True)

compare_metrics

#Final Prediction

#Predict the future data based on the twitter data polarity

#data 2021-02-01 to 2021-04-30

#read the merge file for the prediction part

data = pd.read_csv('BTC2M.csv')

#data = pd.read_csv('ETH2M.csv')

# set date as index

data['Date'] = pd.to_datetime(data['Date'], format='%Y-%m-%d')

data.set_index('Date', inplace=True)
```

```
data = data.filter(['Close'])

# Predict the closing price for 2021-04-20

crypto_subset = data.loc['2021-02-18':'2021-04-19']

# Get the last 60-day closing price values and convert the dataframe to an array

last_60_days = crypto_subset[-60:].values

# Scale the data to be values between 0 and 1

last_60_days_scaled = scaler.transform(last_60_days)

# Create a list for the the last 60-day

X_test = []

X_test.append(last_60_days_scaled)

X_test = np.array(X_test)

X_test = np.reshape(X_test, (X_test.shape[0], X_test.shape[1], 1))

# Fit in the model to get the predicted scaled price

pred_price = model.predict(X_test)

# Unscale the data

pred_price = scaler.inverse_transform(pred_price)

print('2021-04-20')

print(pred_price)
```

```
# actual stock price for 2021-04-20

data.loc['2021-04-20':'2021-04-24']
```

<Appendix E>

“Mergerd_data.csv

Date	Open	High	Low	Close	Volume	Polarity
1/2/2021	1314.855	1373.846	1274.358	1369.041	29210670920	0
2/2/2021	1369.505	1542.991	1362.771	1515.194	45437142801	0
3/2/2021	1514.77	1660.91	1510.01	1660.91	41874566399	0
4/2/2021	1661.17	1689.187	1561.854	1594.763	44396871836	0
5/2/2021	1594.793	1756.511	1594.793	1718.651	40108628454	0.059511
6/2/2021	1717.797	1738.314	1649.069	1677.847	39873420648	0.056411
7/2/2021	1677.606	1690.037	1501.75	1614.228	39889440151	0.047816
8/2/2021	1613.642	1770.591	1571.58	1746.617	48012285956	0.0865
9/2/2021	1746.926	1815.964	1711.621	1768.035	44180727529	0.071601
10/2/2021	1768.04	1826.697	1686.542	1744.243	41916084617	0.070219
11/2/2021	1743.714	1806.539	1708.679	1783.798	36021495262	0
12/2/2021	1783.489	1861.357	1744.169	1843.533	37905036865	0
13/2/2021	1843.987	1871.604	1770.612	1814.11	35359490535	0.065278
14/2/2021	1814.372	1848.154	1789.914	1805.084	31439114900	0.08913
15/2/2021	1804.677	1833.831	1683.907	1779.791	38955610883	0.07802
16/2/2021	1778.946	1824.519	1729.642	1781.068	34269369268	0
17/2/2021	1781.35	1853.668	1736.706	1848.458	35955412703	0
18/2/2021	1848.206	1949.903	1848.206	1937.449	28255902969	0.076374
19/2/2021	1938.86	1969.547	1896.684	1960.165	26268814253	0.069375
20/2/2021	1959.903	2036.286	1830.531	1919.534	34696091102	0
21/2/2021	1918.673	1974.26	1890.368	1935.601	23626547717	0
22/2/2021	1935.558	1936.454	1580.627	1781.993	42409646036	0.06271
23/2/2021	1781.409	1781.409	1378.841	1570.204	52029864713	0
24/2/2021	1571.476	1710.984	1511.019	1626.576	31329000537	0
25/2/2021	1625.394	1670.224	1465.059	1475.704	24481681873	0
26/2/2021	1478.653	1559.029	1407.979	1446.034	31435997881	0
27/2/2021	1446.929	1524.932	1433.787	1459.973	20742103233	0
28/2/2021	1459.86	1468.391	1300.472	1416.049	27637026080	0.060825
1/3/2021	1417.151	1567.695	1416.416	1564.708	24032838645	0
2/3/2021	1564.063	1597.61	1461.325	1492.609	22523669722	0
3/3/2021	1491.451	1650.361	1481.906	1575.853	22674780680	0
4/3/2021	1574.624	1622.954	1511.103	1541.914	22906118718	0
5/3/2021	1541.542	1547.878	1450.891	1533.275	21067146937	0
6/3/2021	1532.373	1669.107	1519.141	1654.742	22746262366	0
7/3/2021	1655.392	1730.924	1636.564	1723.154	23809935410	0
8/3/2021	1724.229	1835.192	1670.942	1834.728	27630991158	0
9/3/2021	1835.148	1868.049	1804.266	1868.049	23461244507	0
10/3/2021	1868.489	1873.803	1766.49	1799.166	25154173185	0
11/3/2021	1798.034	1843.819	1734.617	1826.195	24013132909	0.093476
12/3/2021	1826.547	1839.497	1728.981	1772.102	22435821312	0.052921
13/3/2021	1772.166	1937.646	1733.64	1924.685	25014689475	0
14/3/2021	1923.864	1930.78	1845.12	1854.564	19344589211	0

15/3/2021	1854.087	1889.197	1749.606	1791.702	26244738810	0
16/3/2021	1792.414	1817.06	1720.053	1806.972	23828509590	0
17/3/2021	1807.056	1839.819	1749.18	1823.449	24512917348	0
18/3/2021	1823.158	1848.646	1705.716	1782.855	23263845504	0
19/3/2021	1782.569	1841.196	1746.473	1817.624	21249297710	0
20/3/2021	1817.523	1874.709	1811.729	1812.635	22677674970	0
21/3/2021	1812.607	1823.353	1764.139	1788.217	22977404620	0
22/3/2021	1788.362	1811.968	1674.3	1691.334	23599296129	0
23/3/2021	1690.872	1725.109	1662.54	1678.65	21998237965	0
24/3/2021	1678.003	1740.428	1570.788	1593.413	31228051473	0
25/3/2021	1593.123	1625.911	1560.37	1595.359	29650328701	0
26/3/2021	1595.21	1702.923	1594.737	1702.842	22548516548	0
27/3/2021	1703.036	1732.824	1674.319	1716.495	18102277710	0
28/3/2021	1716.406	1728.584	1672.66	1691.356	16599472938	0
29/3/2021	1691.263	1837.188	1683.717	1819.685	22796570548	0
30/3/2021	1819.466	1860.975	1793.922	1846.034	22512781703	0
31/3/2021	1846.098	1947.838	1793.002	1918.362	30226902621	0
1/4/2021	1919.157	1989.055	1912.178	1977.277	30914259795	0
2/4/2021	1976.933	2152.452	1960.679	2143.226	34862511022	0
3/4/2021	2142.896	2144.962	2028.422	2028.422	32011518871	0
4/4/2021	2027.671	2110.354	2007.112	2093.123	26006501902	0
5/4/2021	2093.261	2140.985	2032.388	2107.887	28889391170	0.053347
6/4/2021	2109.493	2151.223	2057.609	2118.379	29222865881	0.06326
7/4/2021	2117.729	2133.188	1945.442	1971.077	36116271935	0.047458
8/4/2021	1969.133	2091.516	1959.079	2088.574	25312956529	0.048841
9/4/2021	2088.772	2102.874	2055.163	2072.109	19812472092	0.043097
10/4/2021	2071.112	2196.996	2062.788	2135.942	24986243611	0.062834
11/4/2021	2136.157	2165.191	2119.866	2157.657	19692836132	0.101448
12/4/2021	2157.362	2199.719	2110.369	2139.353	21727936609	0.053171
13/4/2021	2139.364	2318.423	2138.56	2299.188	29456642939	0
14/4/2021	2299.348	2449.688	2284.564	2435.105	35592822986	0
15/4/2021	2436.035	2544.267	2409.924	2519.116	32325606817	0
16/4/2021	2516.602	2547.556	2318.675	2431.947	36196928256	0
17/4/2021	2429.981	2497.385	2333.683	2344.895	32349808978	0.146055
18/4/2021	2346.452	2365.46	2011.767	2237.137	50696368718	0.098617
19/4/2021	2238.033	2276.777	2086.689	2166.189	34060654971	0.104793
20/4/2021	2161.939	2345.835	2060.144	2330.211	39433483315	0.127021
21/4/2021	2331.16	2467.201	2238.367	2364.752	38899067643	0.116279
22/4/2021	2357.871	2641.095	2315.96	2403.535	53575904724	0.125713
23/4/2021	2401.256	2439.537	2117.04	2363.586	55413933925	0.111589
24/4/2021	2367.199	2367.741	2163.693	2211.626	31854226936	0.112625
25/4/2021	2214.414	2354.087	2172.515	2316.06	31814355546	0
26/4/2021	2319.478	2536.337	2308.315	2534.482	35208325408	0
27/4/2021	2534.031	2676.393	2485.375	2662.865	32275969215	0
28/4/2021	2664.686	2757.477	2564.082	2746.38	34269031076	0
29/4/2021	2748.65	2797.972	2672.107	2756.877	32578127990	0
30/4/2021	2757.734	2796.055	2728.17	2773.207	29777179889	0