A deep learning approach to predict crypto-currency price by evaluating sentiment and stock market correlations

by

Miftahul Zannat Maliha
22341041
Ananya Subhra Trisha
20241062
Abu Mauze Tamzid Khan
19301045
Prasoon Das
18101603
Shuhanur Rahman Shakil
18301243

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

Department of Computer Science and Engineering
BRAC University
January 2023

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at BRAC University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Miftahul Zannat Maliha.

Miftahul Zannat Maliha
22341041

Ananya Subhra Trisha

Ananya Subhra Trisha
20241062

Tamzid Khan

Abu Mauze Tamzid Khan
19301045

Prasoon Das

Prasoon Das
18101603

Shuhanur Rahman

Shuhanur Rahman Shakil
18301243

Approval

The thesis/project titled “A deep learning approach to predict crypto-currency price by evaluating sentiment and stock market correlations” submitted by-

1. Miftahul Zannat Maliha (22341041)
2. Ananya Subhra Trisha (20241062)
3. Abu Mauze Tamzid Khan (19301045)
4. Prasoon Das (18101603)
5. Shuhanur Rahman Shakil (18301243)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January 19, 2023.

Examining Committee:

Supervisor:
(Member)

Iqbal

Muhammad Iqbal Hossain, PhD
Associate Professor
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

RRahman

Rafeed Rahman
Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

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

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

For the technological shift, advancing epoch towards cryptocurrency intensified the impactful method. Metaverse can originate the base operation into a diversified level. The extension of digital marketing contributes to blockchain technology more.Our research demonstrates, attested cryptocurrency price evaluation associated with the stock and sentiment. In our research, we have implemented various techniques to predict cryptocurrency prices. Crypto like bitcoin, ethereum and litecoin are the primary focus in this paper. Our research observes the fluctuation into the cryptocurrency prices. In our research procedure, we used the LSTM-GRU hybrid, ARIMA for time series prediction. The research follows sentiment analysis from the twitter scrapped data. The research provides cogent insights of cryptocurrency price prediction fluidity with the stock price and the twitter sentiment on following cryptocurrencies. Additionally, the data merge with the LSTM time series model depicts the cryptocurrency stock market and shows us the relationship between stock price, twitter sentiment and cryptocurrency price pertinence.

Keywords: Crypto-currency; Machine Learning; Bit Coin; Sentiment Analysis; Prediction; Stock Market;

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our Supervisor and Co-Supervisor Muhammad Iqbal Hossain Sir and Rafeed Rahman Sir for their kind support and advice in our work. They helped us whenever we needed help.
And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$ARIMA$ Autoregressive Integrated Moving Average

BTC BitCoin

ETH Ethereum

$GRNN$ A general regression neural network

GRU Glavnoye Razvedyvatelnoye Upravlenie

$LSTM$ Long short-term memory

LTC LiteCoin

NNR Neuronal Nicotinic Receptor

RNN A recurrent neural network

$VADER$ Valence Aware Dictionary and Sentiment Reasoner

Chapter 1

Introduction

Cryptocurrency enhanced the probability over the market with growing solubility behind the technology. The market instability and fluctuations of this digital currency have drawn a lot of attention. While the safety and security of online transactions are obvious selling points, it is the investment potential that has sparked the most alarm among the general population. However, because the value of cryptocurrencies is not backed by anything, investments in them may be extremely volatile. The uncertainty produced by the volatility of cryptocurrency pricing affects both investors and consumers, who now want to use cryptocurrencies more as a currency than an investment. Given these circumstances, cryptocurrency price prediction has been obscure. Additionally, the main goal of this research was to establish the reality of a turbulent price change in the years 2020, 2021, and 2022.The entire economy agonized due to the pandemic and global events. In 2022, cryptocurrency and stocks went into turmoil with a drastic price shift. To conduct our research, we have used deep learning techniques. We provide an efficient technique for anticipating cryptocurrency price swings in this study. Given their unpredictability and intricacy, projecting the price of cryptocurrencies is always challenging. The findings in this article demonstrate that cryptocurrencies play a key role in investor portfolios since they offer a diversification alternative for investors, indicating that cryptocurrencies represent a new financial asset class [23].The investment opportunity in cryptocurrencies and the assortment of investments have influenced our research. For preceding our research, we used LSTM -GRU hybrid for price prediction. We have implemented ARIMA to see the time series prediction. The research reaches to the final outcome with time series LSTM. The research is insightful to convey the outcome into the cryptocurrency oriented market.

1.1 Motivation

Prime objective to address crypto-currencies price prediction with the stock and the sentiment. The year 2022 and 2021 went on with a rigorous shift in the crypto currency over the prices in the year 2022. The price disruption caused was throughout the market from the beginning of 2022. The relevant research we did to get the outcome of crypto-currency prices correlation with the twitter sentiment and the stock market.

1.2 Problem Statement

- We have executed deep learning algorithmic approach to detect the crypto-currency price.
- This research was primarily cryptocurrency data-driven. The bulk amount of crypto-currency historical data and predicting outcomes was tedious. To add up on, the shift and out come the prediction result are pompous.
- The prediction with stock sentiment and crypto-currency price got more accurate results with the data of 2022.
- We implemented LSTM-GRU hybrid, ARIMA for time series prediction Vader for predicting sentiment and emotions.
- Better accuracy and computational time decreases with the time series LSTM model.
- Initially we have used bitcoin to predict the price deterioration then we have seen the similar deficit over Litecoin and Ethereum.

1.3 Objective and Contribution

In this study, we evaluate and compare a range of cryptocurrency price prediction models, accounting for correlations with stock market shares and sentiment analysis on Twitter. Specifically, the following is a summary of the major contributions that the paper makes:

- We compared many models across three distinct crypto-currencies like Bitcoin, Ethereum, and Litecoin and identified the most accurate model capable of handling huge volumes of data.
- We establish a correlation between the price fluctuations of cryptocurrencies and the stock market. Our analysis reveals a correlation between crypto-currency price fluctuations and stock market values beginning in 2022.
- We collect the data for sentiment analysis using an API called Twint that has not before been utilized in research. This API can scrape more precise data for an endless amount of time. The information is collected daily and then combined for an entire year.
- We implement the LSTM time series model on a dataset comprising sentiment scores, cryptocurrency, and stock market prices. As the currency of crypto-currencies is influenced by the stock market and public perceptions, the accuracy of price predictions improves.

1.4 Thesis Structures

The majority of the research thesis has been structured in the manner given below, stepwise:

- Chapter 2 analyzes the current work on cryptocurrency price prediction through various deep learning models. Also, fetch the perception using a sentiment-based approach to detect the price of cryptocurrencies. Chapter 2 demonstrates the background of the research and the models used to predict cryptocurrency prices.
- Chapter 3 illustrates the proposed models for data collection and data pre-processing for the research work. Additionally, it shows the blueprint of the research work.
- Chapter 4 displays the evaluation and assessment of the following outcome.
- Chapter 5 presents the entire research in a concise manner with an insight into future work.

Chapter 2

Literature Review

The center of focus in this research is predicting cryptocurrency prices. Bitcoin became roughly twice as expensive in January 2022 as it was in January 2021. The price fluctuation of cryptocurrencies has been routinely followed since 2009. The fourth industrial revolution gave birth to augmented reality as well as virtual reality. Digital marketing is another key area to implement cryptocurrencies. Cryptocurrency security and surveillance rely on blockchain technology, making their long-term viability less certain. Even as the coronavirus outbreak slowed the economy and raised concerns about rising inflation on the US dollar, bitcoin's price began to accelerate its upward trend. Bitcoin's value had climbed by more than 300 percent from January 2019 to December 2020. The year finished with a price of around 29,374 dollars, the highest ever. The shift highlights the instability of cryptocurrencies. The previous year, 2022, saw a dramatic shift in several other cryptocurrencies as well. Profound transposition in crypto currencies like litecoin, ethereum, tether, and dogecoin is also noticeable. The movement in cryptocurrencies prices conveys the gist of the oscillation. In our research, we have distinguished the past year's prices of cryptocurrencies. The offshoot of the research found a distinctive variation over the cryptocurrency prices. Comparatively, we have observed that ethereum is 52.08 percent higher than it is now. Litecoin is 29.69 percent less than this year. Dogecoin perceived the outcome at 43.58 percent. The indistinguishable pattern is spotted over stable cryptocoins like tether, on account of the US dollar price shift. This research features cryptocurrency price prediction through sentiment and stock price analysis. In addition, the variance between cryptocurrency and stock market is in the regulation criteria. The cryptocurrency market is decentralized, whereas the stock market has its own regulatory body. Likewise, cryptocurrency investment can provide short-term profits due to the price oscillation. Cryptocurrency does not track market fundamentals analysis, cryptocurrency price is more transitional, the currency is not regulated and technology dependent. Conversely, industry experts and influencers anticipate cryptocurrency prices will stay extremely low until the end of 2022. Consequently, the tendency is projected to continue until the first few weeks of 2023. Every cryptocurrency specialist has a unique analysis or forecast of the Bitcoin price. Despite the recent volatility and price drop, several of them remain optimistic that Bitcoin will exceed 100,000 US dollars.

2.1 Related Work

Cryptocurrency price prediction and sentiment analysis had been prioritized on the basis of the growing digitalization. Machine learning technology has been used as a cryptocurrency price prediction model. ML is significantly used in terms of analysis and prediction accuracy; the model takes data as an input, trains the proper data set, and forecasts the outcomes. Techniques which are significantly used are Artificial Intelligence to speculate the stock market for decades. Support vector machines, random forests, and neural networks are the methods that are most frequently used. In the case of the NASDAQ composite index, recurrent neural networks were employed to anticipate market direction. Additionally, prosperous models handle stock market price forecasting as a classification problem rather than a regression problem, as one might think. The predictive model was proposed by incorporating the SVR and GARCH models[16]. Therefore, the SVR-GARCH model outperforms the GARCH, EGARCH, and GJR-GARCH models[16]. Significant progress is being made every day in determining which way the Standard and Poor's 500 stock index futures' prices will move [3][2][4].Back propagation algorithms, a multilayer feedforward network, have been shown in numerous studies for neural networks to be effective in overcoming stock market prediction difficulties[21][1].Additionally, how easily back propagation methods can outperform the best regression models for pricing projection[5]. A widely used crypto currency is Bitcoin, which influences investors. However, to predict Bitcoin prices, ARIMA and NN are widely used. Hence, for the training sample of data, it was discovered that the NNAR surpassed the ARIMA in terms of forecasting bitcoin prices [19]. In the test sample of the data, the reverse was true. Overall, the ARIMA price projection findings were shown to be more accurate than the NNAR [19]. Additionally, the application of LSTM networks as well as generalized regression neural networks (GRNN) for determining the value of cryptocurrencies has been researched. [20]. Meanwhile, researchers attempted to perform blockchain research using machine learning models where, machine learning models to identify user reviews and discovered that the amount of answers to user reviews and online comments had an impact on the number of transactions between users [2]. Financial sector forecasting is a well-known and thoroughly researched topic of future cash flows [26]. According to how signals suggest market changes, linear regressions impose strict assumptions and are unable to take into account a wide variety of characteristics. Contrarily, the use of machine learning technologies for stock market prediction is common and frequently does not impose such restrictions [13], [24].As a result, conclusions about traditional financial markets' low integration may not apply to the cryptocurrency market. The level of market efficiency in the bitcoin retail space is the subject of several studies covering a range of time horizons [8], [10], [18], [25]. Urquhart examines the time series of daily bitcoin prices. This research provided an efficient outlook, as the outcome of the research asserts the bitcoin market is not even inefficient. The cryptocurrency market, however, becomes more meticulous with time, according to a time division study. Since the beginning of the bitcoin market, its size and complexity have increased, and it would seem that market competency is rising as well.

The cryptocurrency market is frequently seen as unpredictable and subject to the emergence of price dynamics bubbles, particularly in the case of Bitcoin [14]. Sources of textual data Various social media platforms, including Twitter, Wikipedia, and

Reddit forums, disseminate cryptocurrency knowledge. Social media is quite useful for forecasting future events and developments [28]. Particular assistance is vital for knowledge diffusion in online communities because network effects are considerable [6]. After examining internet forums, several cryptocurrencies are exchanged online [7]. Tweets that enhance sentiment polarization have been proven to favorably affect Bitcoin prices. Reddit activity has been linked to the propagation of epidemic-like investing ideas, which has aided in the discovery of cryptocurrency price spikes [16]. Also, each platform is very specialized when it comes to making its own kind of text content. The interaction with their audience in a specific way, whether through simple comments on Twitter or rigorously written articles or online posts that can range in length from short replies to more tangled texts. Each venue has a distinct impact on investors and traders. The use of many data sources can lead to more accurate price forecasts, although it is unclear how each social media site affects the final price. Until date, there has been no mention of how different data sources' properties impact forecast performance for bitcoin prices[11]. Furthermore, the field of market prediction, and particularly cryptocurrencies, suffers from a scarcity of high-quality datasets[9]. A variety of machine learning techniques are used for sentiment analysis on social media. An example would be a twitter sentiment analysis[19] have been developed during the last decades to anticipate stock market price movement using social media. Furthermore, it was discovered that using Twitter sentiment analysis in conjunction with the LSTM model outperformed other machine learning models at predicting stock price. The Support Vector Machine is utilized in these models [12]. Moreover, it has been established that Google's page views of essential points might improve cryptocurrency price predicting ability[17], [22]. RNN-based valuation is shown to have an overall accuracy rate of 77.62 percent. The predicted percentage for positive and negative sentiment segmentation of tweets is 81.39 percent [15]. Above works manifest sentiment analysis and market both have equal consequences into the cryptocurrency price movement.

2.2 Background Studies

Numerous researches were conducted for the determination of cryptocurrency price. The sentiments on social media impact the price prediction. Various Machine learning based researches are seen. Cryptocurrency founder Satoshi Nakamoto introduced it as a decentralized currency. The method used to decentralize cryptocurrency is blockchain technology. Correspondingly, the decentralization policy cannot regulate the price of cryptocurrency. Cryptocurrency prices cannot be effortlessly monitored. This price fluctuation mostly affects investors worldwide. The indicators depending on cryptocurrency prices are many. Events based on sentiments have been impactful in analyzing cryptocurrency prices. Although, the index of the price fluctuations can be seen from the market stock assessment. Sentiment analysis has an effective implication for predicting prices in the cryptoworld. LSTM based sentiment analysis over a Chinese website Sina Weibo describes the results functionally. According to the research, the proposed method outperforms the most advanced auto regressive-based model in precision and recall by 18.5 percent and 15.4 percent, respectively [23]. Utilizing fractional integration techniques, investigate the stochastic features of six main cryptocurrencies and their bilateral links with six stock market indexes [27].

The models that were used to predict/forecast the price of cryptocurrencies are described below:

2.2.1 ARIMA

The data in this model demonstrate non-stationarity. To get rid of this non-stationarity, apply the first differentiation step once or more. If all of a random variable's statistical characteristics remain consistent across time, it is said to be stationary. A stationary series has no trend, variations that revolve around its mean constant amplitude, and regular wriggling. These latter circumstances indicate that its autocorrelations are stable throughout time. This form's random variable can be thought of as a mix of noise and signal. The ARIMA model can be thought of as a "filter" that seeks to isolate the signal from the background noise before extrapolating the signal into the future to produce projections. The model's equation is linear and of the regression type. The dependent variable delays or the forecast error lags make up the equation's predictors.

$$Predicted\ Value = \begin{cases} \text{constant} \\ \text{weighted average of a few recent values} \\ \text{weighted average of the most recent values of the errors} \end{cases} \tag{2.1}$$

An observation and a certain number of lag observations are used in autoregressive models as a dependent relationship. The acronym, which stands for "Integrated," also shows how the raw observations are differentiated. It makes the time series stable by deducting one observation from another observation from the preceding time step. The moving average, or MA of ARIMA, is a technique that makes use of the relationship between an observation and the residual errors of a moving average model applied to lag observations. Each element is specifically described in ARIMA as a parameter. For these aspects, three common notations, such as p, d, and q, were used. However, the parameters are replaced with integer values to quickly identify the particular ARIMA model.

2.2.2 GRU

GRUs resemble Long Short Term Memory (LSTM). Gates are used by GRU and LSTM to regulate the information flow. They are a more recent technology than LSTM. Because of this, they perform better than LSTM and have a simpler architecture. The gated recurrent unit (GRU)[4] offered a condensed variant of the LSTM memory cell that frequently achieves equal performance while being quicker to compute [4]. Even though it uses LSTM, it just has three gates and doesn't maintain an internal cell state. The Gated Recurrent Unit concealed state incorporates the data from an LSTM recurrent unit's internal cell state. To the following Gated Recurrent Unit, this aggregate data is forwarded.

Update Gate: The amount of earlier knowledge that must be transmitted into the future is specified by this gate. It functions like the Output Gate in an LSTM recurrent unit.

Reset Gate: This controls how much past knowledge should be forgotten. It is equivalent to the union of the Input Gate and the Forget Gate in an LSTM recurrent unit.

Current Memory Gate: It is typically ignored during a standard discussion of a gated recurrent unit network. Similar to how the Input Modulation Gate is a component of the Input Gate, it is incorporated into the Reset Gate and is used to make the input Zero-mean as well as add nonlinearity to it. To lessen the impact of prior knowledge on information being broadcast into the future, it is included in the Reset gate.

2.2.3 Vader

Specifically adapted to the sentiments conveyed in social media, VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that also performs well on texts from other domains. Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, developed by C.J. Hutto and Eric Gilbert, was published in the Proceedings of the Eighth International Conference on Weblogs and Social Media. (ICWSM-14).
The VADER model uses a combination of lexical features and rules-based heuristics to identify the sentiment of a given text. It is particularly effective at correctly identifying the sentiment of texts that are expressed in a very positive or very negative way, as well as texts that contain emotional or exaggerated language.
One of the key features of the VADER model is its ability to take into account the context in which words are used, and to weight the sentiment of individual words based on this context. This allows it to identify the overall sentiment of a text even if it contains words with conflicting sentiment.
VADER has been widely used in a variety of applications, including social media analysis, customer service, and market research. It is implemented in the Python Natural Language Toolkit (nltk) library, and is available for use in other programming languages as well.

Vader architecture: The lexicon-based sentiment analysis program VADER (Valence Aware Dictionary and sEntiment Reasoner) combines lexical features with rules-based heuristics to determine the sentiment of a given text. It is implemented in the Python Natural Language Toolkit (nltk) library, and is available for use in other programming languages as well.
The VADER model consists of three main components:

1. A list of terms and expressions that are connected to sentiments, either positive or negative. In addition to a list of words and phrases that are known to be connected with a text's tone, this dictionary also provides a list of words and phrases that can heighten a text's sentimental impact.

2. A set of rules-based heuristics that take into account things like punctuation, capitalization, and the presence of certain emoticons to identify the sentiment of a text.

3. A scoring algorithm that combines the sentiment scores from the dictionary and the heuristics to produce an overall sentiment score for the text. The scoring algorithm takes into account the context in which words are used and weights the sentiment of individual words based on this context.

The VADER model requires only the text you wish to analyze as input, and it will return a sentiment score for the text. The emotion score runs from -1 to 1, with a score of -1 denoting an extremely unfavorable opinion, a score of 0 a neutral opinion, and a score of 1 an exceptionally positive one.

2.2.4 LSTM

We are going to use LSTM model on our dataset to predict the price. LSTM stands for Long ShortTerm Memory. Recurrent neural networks are a subtype of this. An RNN's current step is fed data from the previous phase's output. It works with the issues of long-term RNN dependence, where the RNN can assume or predict data or words from the current data, but it's impossible for this RNN to predict words that are stored in the long-term memory. RNN's completion becomes less effective after increasing the gap size. This LSTM model can automatically save the data for a longer time frame. It is applied to classification, time-series data processing, and forecasting. Additionally, gates and cell states are utilized in this LSTM model. LSTM model can give a proper solution to the problems faced by the RNN where Long term dependency problems in RNNs and Vanishing Gradient Exploding Gradient are the two major ones. The cells are responsible for information retention, and the gates control memory manipulation. In LSTM, gates come in three different varieties.

Forget Gate: The LSTM architecture's initial component is illustrated by the forget gate (ft). The information from the current input (Xt) and the previous hidden state is communicated using the sigmoid activation function (ht). When the output value is close to 0, it suggests forgetting, whereas when it is close to 1, it suggests keeping.

$$f_t = \sigma(W_f.[h_t - 1, x_t] + b_f) \tag{2.2}$$

Input Gate: The input gate changes the cell state by including relevant information. The sigmoid function, which is utilized in a way similar to the forget gate, is first used to manage the information and filter the values that need to be remembered using the inputs h(t)-1 and x(t). Then, using the tanh function, a vector is generated with an output range of -1 to +1 that comprises all possible values for h(t)-1 and x(t). To create usable information, multiply the vector's values by the prescribed values in the last step.

$$i_t = \sigma(W_i.[h_t - 1, x_t] + b_i) \tag{2.3}$$

$$C_t = tanh(WC.[h_t - 1, x_t] + b_C) \tag{2.4}$$

Output Gate: The output gate chooses the value of the subsequent concealed state. This state stores information from previous inputs. The third sigmoid function is first supplied with the values of the previous and current hidden states. The tanh function then receives the new cell state that the cell state produced.

$$O_t = \sigma(W_o.[h_t - 1, x_t] + b_o) \tag{2.5}$$

$$h_t = o_t * tanh(C_t) \tag{2.6}$$

Here,

$$i_t = \text{input gate} \tag{2.7}$$
$$f_t = \text{forget gate} \tag{2.8}$$
$$o_t = \text{output gate} \tag{2.9}$$
$$\sigma = \text{sigmoid function} \tag{2.10}$$
$$w_x = \text{weight for the respective gate(x) neurons} \tag{2.11}$$
$$x_t = \text{input at current timestamp} \tag{2.12}$$
$$b_x = \text{biases for the respective gates(x)} \tag{2.13}$$
$$h_t - 1 = \text{output of previous LSTM block} \tag{2.14}$$
$$\tag{2.15}$$

Chapter 3

Proposed Model

3.1 Proposed System

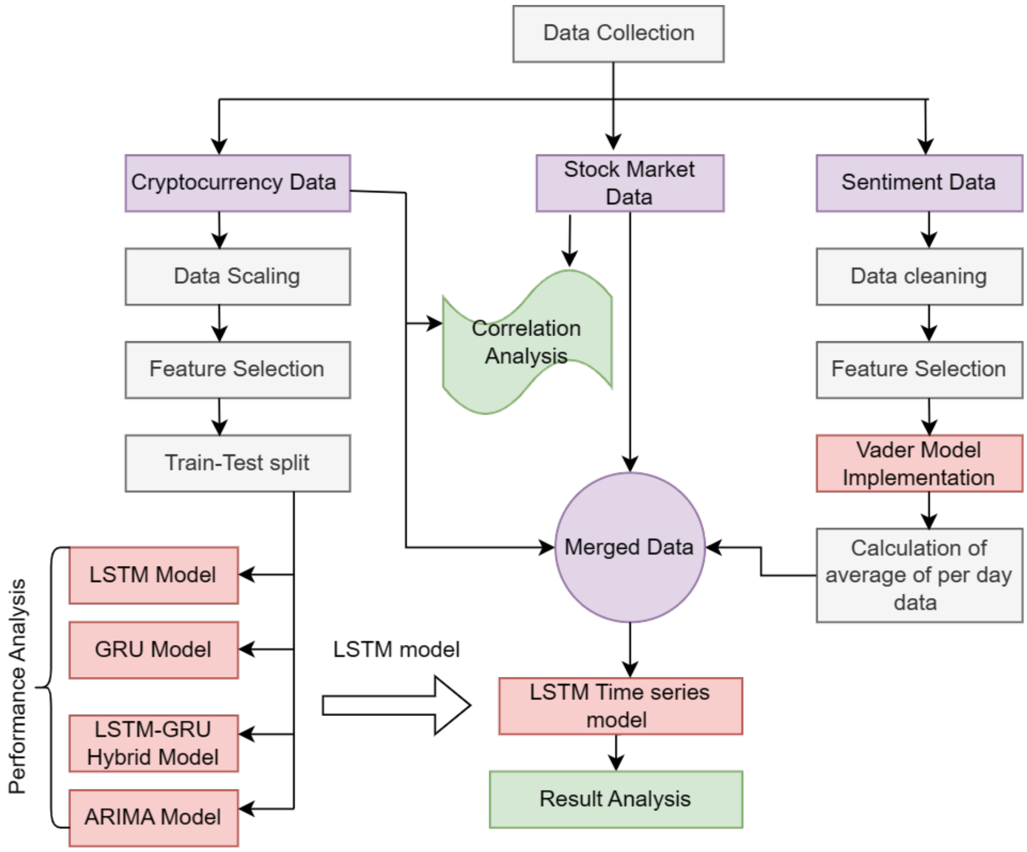

Figure 3.1: Top level view of the system

Figure 3.1 represents the workflow of the proposed process, from data collection to final result analysis. Following the collection of data from three different streams, including cryptocurrency, the stock market, and tweeter sentiment, the most accurate model to predict the price of cryptocurrency was initially developed. The dataset was processed using an LSTM model, a GRU model, an ARIMA model, and a hybrid LSTM-GRU model. The second finding was the possibility of a correlation between the price of cryptocurrencies and the price of the stock market. Thirdly, the sentiment analysis was carried out, and using the VADER model, a sentiment score was calculated. In the end, after combining the sentiment score with the price of the cryptocurrency and the price of the stock market, a combined dataset was created. This dataset was then fed into the LSTM model, which predicted the price of the cryptocurrency.

3.2 Data-set Description

This study uses the information collected from Yahoo Finace for cryptocurrency and stock market. It offers real-time information, like financial market analysis. It connects to an API which returns the historical data of BTC-USD, ETH-USD, and LTC-USED for Bitcoin, Ethereum and Letecoin respectively, in the format of a csv file. The stock market price of Nasdaq was collected in the same manner from Yahoo Finance. Numerous issues have arisen, such as the fact that the rate of cryptocurrency conversion fluctuates depending on the chosen currency and that data is available for different currencies at different times. The following feature is utilized to collect and filter the daily pricing data for every dataset in the cryptocurrency and stock markets:

(i) Open: The daily opening price

(ii) Close: The daily closing price

(iii) High: Each day's highest price

(iv) Low: Each day's lowest price

(v) Adj Close: Each day's Adjusted closing price

(vi) Volume: Total collection of each day

Twint and the Twitter API are used to gather tweets on bitcoin. Without using Twitter's API, users can scrape tweets from Twitter profiles using Twint, an advanced Twitter scraping program created in Python. From January 22 until November 22, the framework was used to collect 50 data every three hours. We scrape the tweeter data by searching with the keyword "bitcoin" to get all the tweets that were made using the hashtag. Similarly, for the Ethereum and Litecoin dataset, we search with "ethereum" and "litecoin" respectively. Only the English-language tweets were taken, as the language configuration was kept at "en".

3.3 Data Preprocessing

As there were not any null values in the datasets of Bitcoin, Litecoin, Ethereum, and Nasdaq Daily, they did not need any preprocessing. However, for the sentiment data, because the tweets from Twitter are scraped directly from Twitter, there is a requirement that the data be pre-processed before it can be used. The tweets that include hashtags, special symbols, emoticons, and a wide variety of other unwelcome content have been removed from the dataset provided by Twitter. In addition to this, it makes use of a regular expression in order to get rid of the undesirable characters. Since the data were collected on a monthly basis, it was subsequently compiled into a single data set. Finally, the unnecessary columns were dropped, and only the three columns "Date", "Time", and "Tweet" were kept. Tables 3.1, 3.2, and 3.3 show the cleaned data against each Date and Datetime for Bitcoin, Ethereum and Litecoin.

DateTime	Date	Tweet	CleanedTweets
2022-01-31 23:59:57	2022-01-31	#Bitcoin is one	Bitcoin is one of......
2022-01-31 23:59:52	2022-01-31	Lol... Oh, you ju.....	Lol Oh you just r ...
2022-01-31 23:59:46	2022-01-31	One of the larges.....	One of the larges.....

Table 3.1: Preprocessed tweeter data (BTC)

DateTime	Date	Tweet	CleanedTweet
2022-01-30 23:59:56	2022-01-30	#NFTGiveaway...	NFTGiveaway 1x...
2022-01-30 23:59:29	2022-01-30	Reply another #ET...	Reply another ETH...
2022-01-30 23:59:06	2022-01-30	Hot New Avail...	Hot New Available..

Table 3.2: Preprocessed tweeter data (ETH)

DateTime	Date	Tweet	CleanedTweet
2022-01-30 23:59:17	2022-01-30	Check out my $LTC...	Check out my LTC ...
2022-01-30 23:50:11	2022-01-30	Confidential tran...	Confidential tran...
2022-01-30 23:49:43	2022-01-30	Shout out to the ...	Shout out to the ...

Table 3.3: Preprocessed tweeter data (LTC)

After running the data through the Vader model, the polarity score "Positive", "Negative", "Neutral" and "Compound" is obtained for each DateTime. Later on, the columns are merged with the crypto price and bitcoin price in Tables 3.4, 3.5 and 3.6. Here, the "StockPrice" column is the price taken from stock data, and the "Price" column is the Adjusted Volume of cryptocurrency price.

DateTime	Negative	Neutral	Positive	Compound	Price	StockPrice
2022-01-03	0.0726	0.8002	0.1270	0.1079	46458.1171	15832.7998
2022-01-04	0.0467	0.8399	0.1133	0.1207	45897.5742	15622.7197
2022-01-05	0.0454	0.8392	0.1153	0.2100	43569.0039	15100.1699
2022-01-06	0.0707	0.8415	0.0875	0.0669	43160.9296	15080.8603
2022-01-07	0.0609	0.8430	0.0960	0.0921	41557.9023	14935.9003

Table 3.4: Merged data of crypto, stock and tweeter sentiment (BTC)

DateTime	Negative	Neutral	Positive	Compound	Price	StockPrice
2022-01-03	0.0205	0.9142	0.0651	0.1586	3761.3803	15832.7998
2022-01-04	0.0366	0.8910	0.0723	0.0874	3794.0566	15622.7197
2022-01-05	0.0437	0.8771	0.0790	0.0924	3550.3869	15100.1699
2022-01-06	0.0217	0.8859	0.0922	0.1960	3418.4082	15080.8603
2022-01-07	0.0305	0.8928	0.0766	0.1299	3193.2104	14935.9003

Table 3.5: Merged data of crypto, stock and tweeter sentiment (ETH)

DateTime	Negative	Neutral	Positive	Compound	Price	StockPrice
2022-03-02	0.0000	0.8345	0.1654	0.4869	110.3513	13752.0195
2022-03-30	0.0000	0.8684	0.1315	0.4929	131.1693	14442.2695
2022-04-29	0.0063	0.8746	0.1190	0.4186	100.3780	12334.6396
2022-05-02	0.0084	0.8774	0.1140	0.3826	100.6932	12536.0195
2022-05-03	0.0018	0.8739	0.1242	0.4339	99.3430	12563.7597

Table 3.6: Merged data of crypto, stock and tweeter sentiment (LTC)

The features are described below:

(i) Negative: Negative sentiment

(ii) Neutral: Neutral sentiment

(iii) Positive: Positive sentiment

(iv) Compound: Scaled sum of positive, negative, neutral sentiment

(v) Price: Closing Price of Cryptocurrency

(vi) StockPrice: Closing price of Stock market

3.4 Experimental Setup

3.4.1 Model for cryptocurrency data

The MiniMax Scaler was used to scale the data once it had been loaded into the dataframe. The goal of normalization is to scale down the values of numerical columns in a dataset without losing information or distorting range differences. MinMaxScaler divides by the range after dividing each value in a feature by its minimal value. The range is the variation between the highest and lowest initial values. The original distribution's shape is maintained by MinMaxScaler. Here, a scale from 0 to 1 was employed. The scaled data was then divided into a training set comprising 60 percent of it and a testing set comprising the remaining 40 percent. How well a model matches its training data is measured by the validation loss, whereas how well it performs on test data is measured by the training loss. Figures 3.2, 3.3 and 3.4 shows the training and validation loss for training by LSTM, GRU and LSTM-GRU hybrid models respectively.

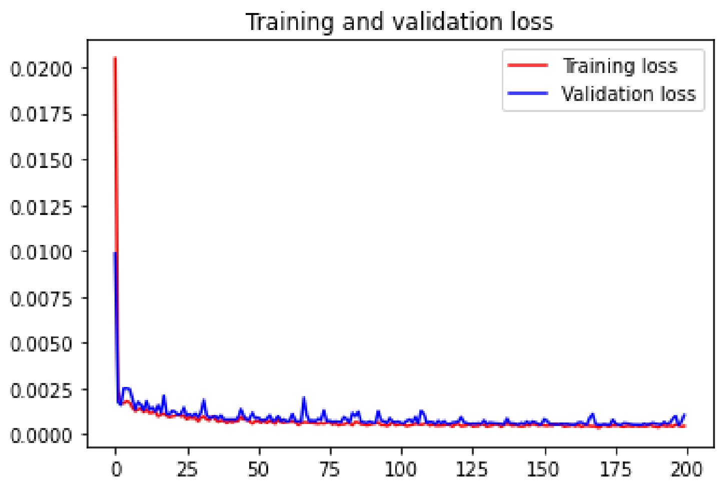

Figure 3.2: Loss of training and validation for LSTM

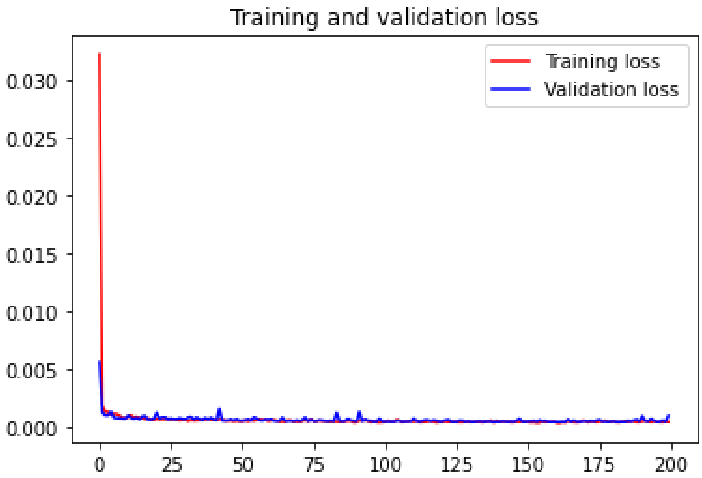

Figure 3.3: Loss of training and validation for GRU

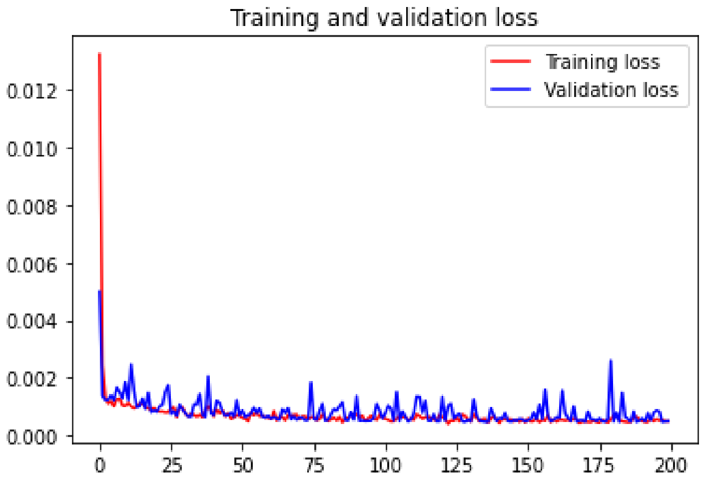

Figure 3.4: Loss of training and validation for LSTM-GRU

Firstly, after building the LSTM model, the training set is fitted by the model with an epoch of 200 and 100 neurons. Later, the data was fitted into the GRU model and the LSTM-GRU hybrid model with the same number of epochs and neurons. Lastly, the data is fitted into the ARIMA model. Since each element is specifically described in ARIMA as a parameter, three common notations, such as p, d, and q, were used for these features. The values specified in our proposed model were taken as (p, d, q) as (4, 1, 0). Finally all the predicted price through models and actual price of BTC along with stock market data of Nasdaq are plotted onto the same graph to find out the correlations.

3.4.2 Model for Sentiment Analysis

1. Vader: The Vader algorithm is initially used to find the polarity of positive, negative, and neutral attitudes. A three-dimensional vector is created as a consequence, representing the polarity of each sentiment type for a certain tweet. There are 0 to 1 values for each polarity. The polarity score of the input tweet is used to categorize the sentiment; the higher the polarity score, the more likely that sentiment is to be present. The VADER code snippet is as follows:

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyser = SentimentIntensityAnalyzer()
def senti_score_udf(sentence):
    snt = analyser.polarity_scores(sentence)
    return ([snt['neg'], snt['neu'], snt['pos'], snt['compound']])
func_udf2 = udf(senti_score_udf, ArrayType(FloatType()))
CleanDF = CleanDF.withColumn('p_neg',
    func_udf2(CleanDF['CleanedTweets'])[0])
CleanDF = CleanDF.withColumn('p_neu',
    func_udf2(CleanDF['CleanedTweets'])[1])
CleanDF = CleanDF.withColumn('p_pos',
    func_udf2(CleanDF['CleanedTweets'])[2])
CleanDF = CleanDF.withColumn('p_comp',
    func_udf2(CleanDF['CleanedTweets'])[3])
CleanDF = CleanDF.withColumn("DateTime",CleanDF['DateTime'])
CleanDF.show(3)
```

Given that the preprocessed data was a collection of 50 tweets every three hours, the average of each day was quantified, and data per day was derived. Later, the dates were arranged according to ascending date order. This data, along with the sentiment polarity, is merged with the crypto and stock data using the Python library PySpark to combine the data as shown in Figure 3.6. The data sent after plotting in the graph resembles figures 3.5, 3.6, and 3.7, which show a data analysis for the polarity scores and price fluctuation of each date index.

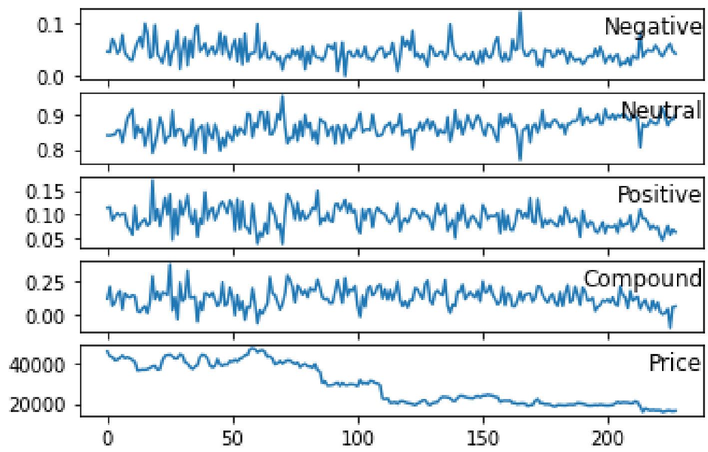

Figure 3.5: Data analysis of Sentiment score and crypto data of Bitcoin

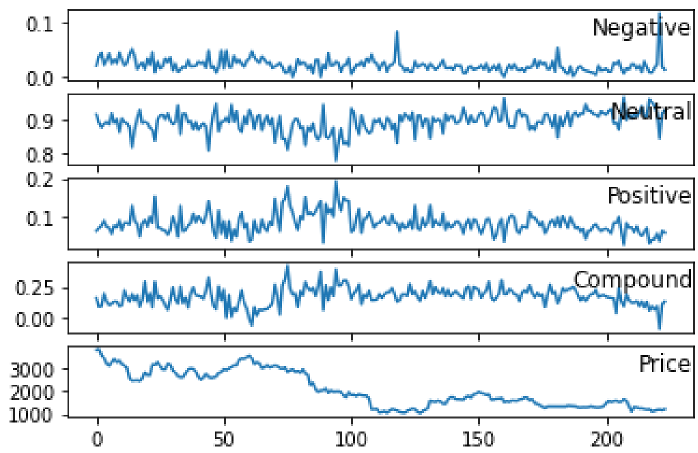

Figure 3.6: Data analysis of Sentiment score and crypto data of Ethereum

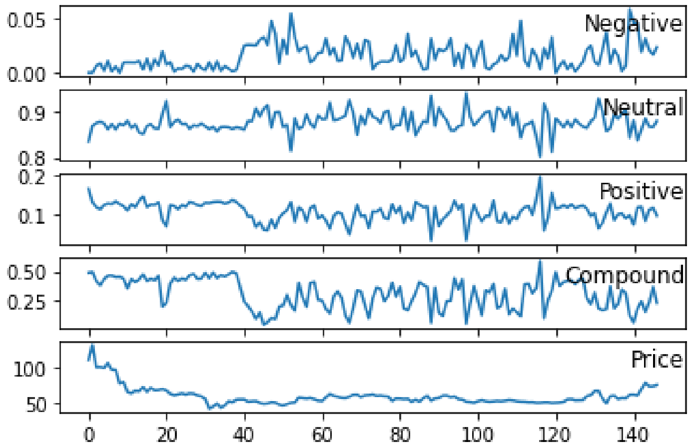

Figure 3.7: Data analysis of Sentiment score and crypto data of Litecoin

2. LSTM Time Series: The data is scaled using the MiniMaxScaler, after which it is divided into a train set, which contains 90 percent of the data, and a test set, which contains 10 percent of the data. The DateandTime is then set as the index. With an epoch value of 100, the train-related data is implemented into an LSTM time series model.

Chapter 4

Experimentation and Result Analysis

4.1 Comparing models for price prediction

This section discusses the performance evaluation of several implemented models. Basically, after fitting the data into the models, some statistical measures are taken to identify the better performance of the models. Two of them are R^2 score and the variance regression score. The percentage of a dependent variable's fluctuation that is explained by independent variables in a regression model is represented by the mathematical statistic R^2 (R2). Correlation measures the strength of the relationship between a dependent and independent variable, whereas R^2 measures the extent to which the variation of one variable explains the variance of the other. A model's inputs can account for almost half of the observed variation if its R2 is 0.50.So a better value or R2 score signifies better performance of the model. Formulae of R^2:

$$R^2 = 1 - \frac{UnexplainedVariation}{TotalVariation} \tag{4.1}$$

Table 4.1 explains the comparison among the R^2 score of three individual models on different data-sets of cryptocurrency that are mentioned below:

	LSTM	GRU	LSTM-GRU Hybrid
Bitcoin	0.9769	0.9785	0.9889
Ethereum	0.9844	0.9845	0.9444
Litecoin	0.98308	0.9810	0.9760

Table 4.1: R2 score comparison of different models on different data-sets

In the Arima model, the measure that is considered is the Mean Absolute Percentage Error(MAPE) score. For measuring forecasting accuracy, it is one of the most significant KPIs. MAPE is the sum of the individual absolute errors divided by the demand (each period separately). Since it determines the loss, a lower MAPE score signifies better performance by the model. In this study, the MAPE score for bitcoin was 1.12 percent, which means it has an accuracy of 98.88 percent. For Ethereum and Litecoin, the MAPE score was 2.97 percent and 3.27 percent, respectively. Hence, the accuracy was 97.03 and 96.73 percent. Even though both ARIMA and

LSTM are able to accurately predict the currency's price, LSTM requires more time to train its neural network model. However, after training, the LSTM could generate more accurate predictions and with greater efficiency. In general, LSTM prediction could be improved by utilizing less historical information. ARIMA is highly effective in making short-term predictions, but as the time period increases, the precision rate decreases.

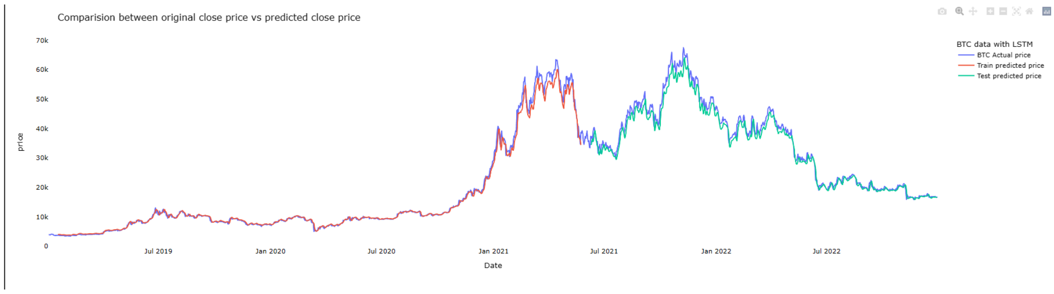

Figure 4.1: Original price vs. Predicted price for LSTM model on Bitcoin

Figure 4.1 displays a graphical representation of the predicted price of bitcoin from 2019 to 2022 in comparison to its original price. We may deduce from the graph that the model can accurately predict the price for both the training and testing sets of data. We can also observe that the price variation is greater in 2021, with prices rising in January and markedly falling in July. Even so, the fluctuation of the pricing model was able to accurately anticipate the price by more than 97 percent.

Figure 4.2 shows a graphic comparison of Ethereum's predicted price from 2019

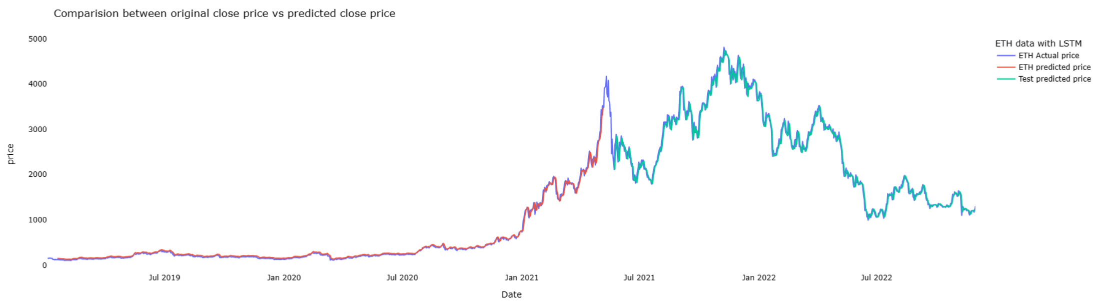

Figure 4.2: Original price vs. Predicted price for LSTM model on Ethereum

to 2022 to its original price. The graph indicates that the model can predict prices accurately for both the training and testing sets of data. We can also see that Ethereum's price variance in 2021 is somewhat less than that of Bitcoin because the graph shows fewer ups and downs. But as the graph shows, there are more peaks and valleys in the price of Ethereum in 2022. Even still, the pricing model's variation was able to predict the price with greater than 98 percent accuracy.

Figure 4.3 shows a graphic comparison between the expected price of Litecoin from 2019 to 2022 and its initial price. The graph shows that for both the training and testing sets of data, the model can accurately forecast prices. As can be seen, the price of Litecoin fluctuates less than the prices of Bitcoin and Ethereum. The start of 2021, however, seems to have seen a considerable surge in pricing. However, the pricing model's variation was able to predict prices to a greater extent than 98 percent.

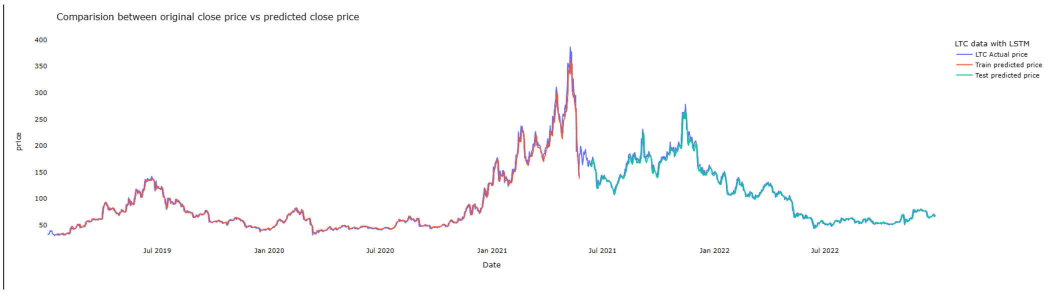

Figure 4.3: Original price vs. Predicted price for LSTM model on Litecoin

Figure 4.4 shows a graphic comparison of Bitcoin's predicted price from January

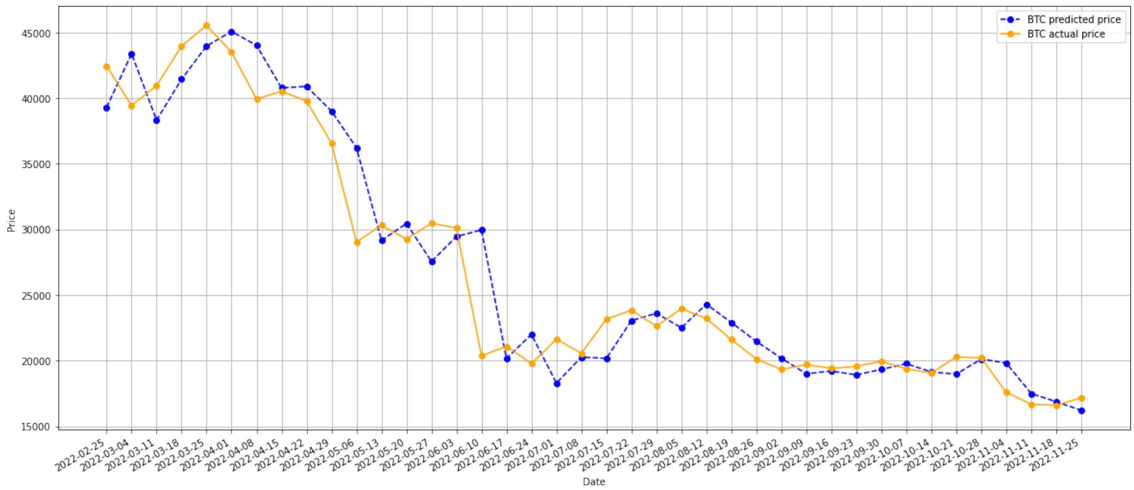

Figure 4.4: Original price vs. Predicted price for ARIMA model on Bitcoin

2022 to November 2022 in relation to its original price. Here, weekly Bitcoin data from 2021 to 2022 is used. Data from earlier years is disregarded because 2021 will see considerable price increases. The graph shows that the price decreased significantly between February 2022 and June 2022. However, the price fluctuation model can anticipate the price with an accuracy of more than 97 percent.

The predicted price of Ethereum from January 2022 to November 2022 with respect to its original price is represented graphically in Figure 4.5. The graph demonstrates that between February 2022 and June 2022, the price decreased considerably, just like it did for Bitcoin. But the price fluctuation model, which suggests that Ethereum has less price volatility than Bitcoin, can predict prices with an accuracy of more than 98 percent.

Figure 4.6 shows a graphic representation of the predicted price of Litecoin from January 2022 to November 2022, relative to its original price. The graph shows that, similar to Bitcoin and Ethereum, the price fell noticeably between February 2022 and June 2022. But starting in October 2022, there will be a substantial price increase. Over 97 percent of predictions made by the model are accurate.

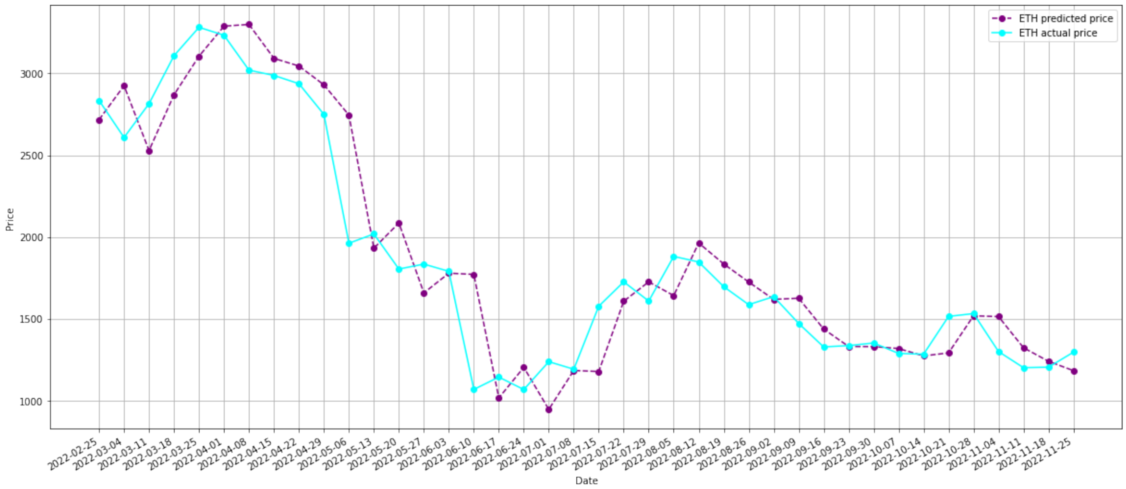

Figure 4.5: Original price vs. Predicted price for ARIMA model on Ethereum

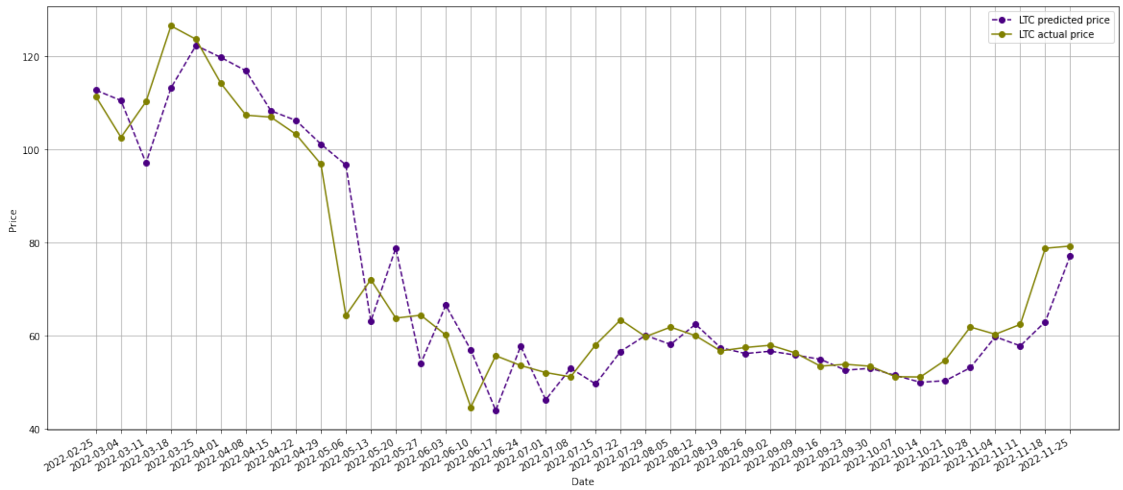

Figure 4.6: Original price vs. Predicted price for ARIMA model on Litecoin

4.2 Cryptocurrency and Stock market correlation

This section will analyze the relationship that has developed over time between the values of cryptocurrencies and stock market indices. Regardless of the fact that cryptocurrency data is significantly more volatile than stock market data. Because cryptocurrencies are decentralized, this indicates that the operations and transactions involving cryptocurrencies are not governed by any central bank. Because of this, the distances that separate the sites prior to January 22 are more unequal and asymmetric. After 2022, both lines are seen to fluctuate on certain days, resulting in a slight correlation between them.

From February 2022 through November 2022, there was a correlation between Bitcoin prices and NASDAQ prices, as seen in Figure 4.7. We can see from the graph that there is a major relationship between NASDAQ prices and Bitcoin pricing because the graph's peaks and troughs have similarities for both. For instance, the NASDAQ price and Bitcoin price both see huge decreases on June 10, 2022, while a rise in both prices is shown on March 25, 2022.

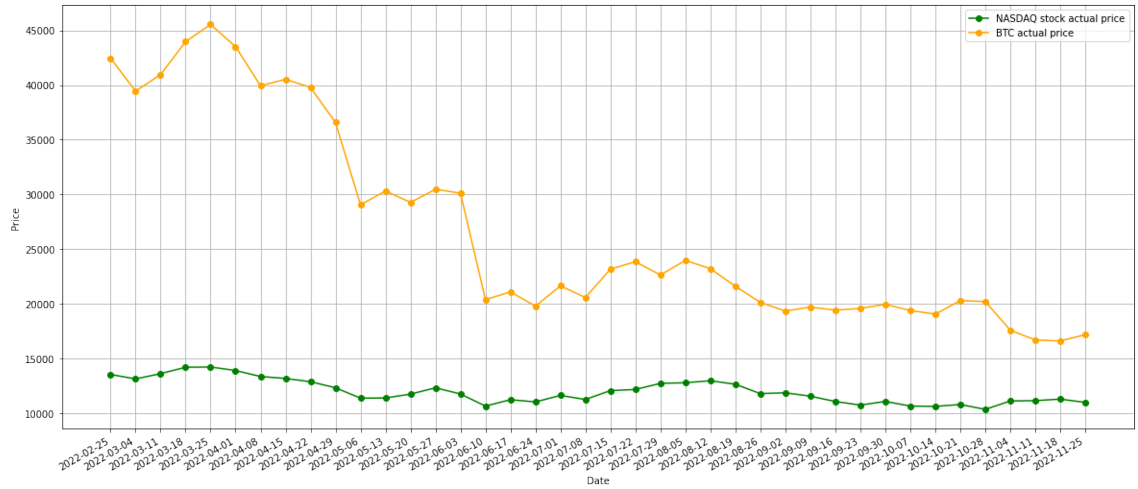

Figure 4.7: Comparison of Bitcoin and Nasdaq

As shown in Figure 4.8, there existed a correlation between Ethereum pricing and NASDAQ prices from February 2022 to November 2022. Because the graph's peaks and troughs are similar for both the NASDAQ and Ethereum values, we can observe from the graph that there is a significant relationship between them. Although there were numerous different peaks and troughs on the graph from June 10, 2022, to July 8, 2022, the correlation was slightly smaller than that of Bitcoin. Nevertheless, a significant relationship is also seen here.

From February 2022 to November 2022, there was a link between the price of

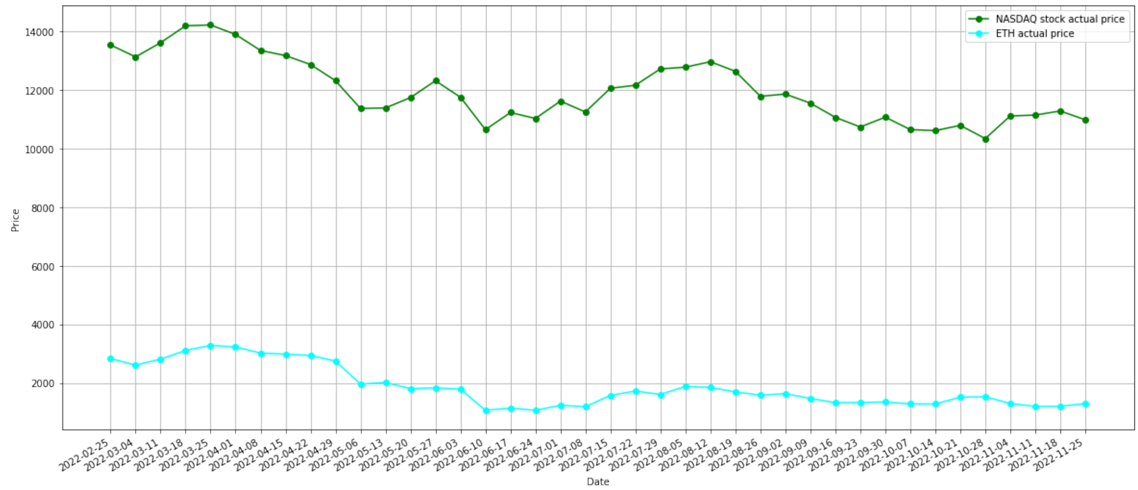

Figure 4.8: Comparison of Ethereum and Nasdaq

Litecoin and the NASDAQ prices, as depicted in Figure 4.9. Due to Litcoin's lower price and lower volume when compared to the NASDAQ, the relationship is not readily apparent. Even so, there are some connected peaks and troughs that show their connection.

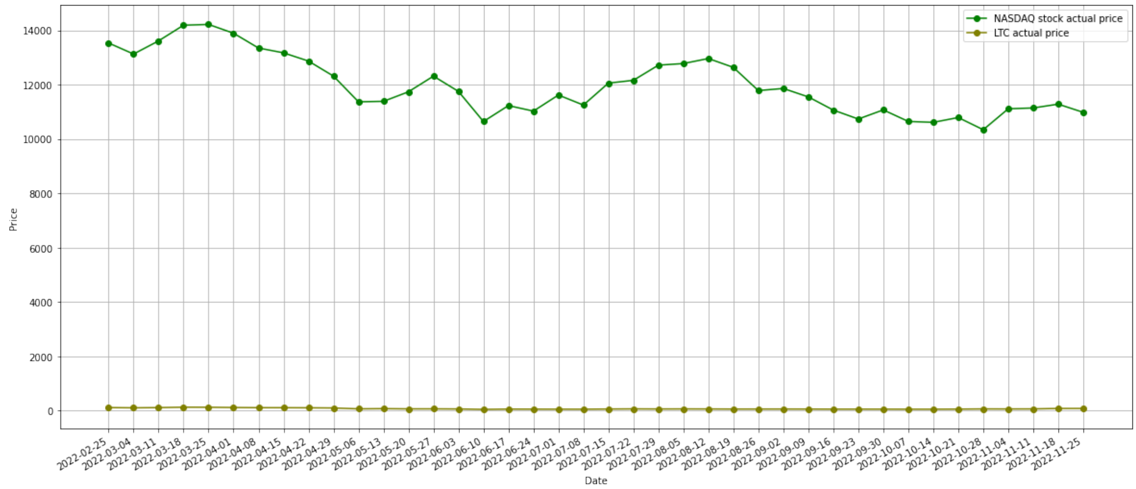

Figure 4.9: Comparison of Litecoin and Nasdaq

4.3 Price prediction using sentiment analysis

Following the sentiment analysis and the data's fitting into the LSTM time series model, the outcome was an R2-score of 99.51 percent, 98.09 percent, and 99.15 percent for Bitcoin, Ethereum, and Litecoin, respectively. This shows that it outperforms the LSTM model when used only with the various cryptocurrency datasets. Figures 4.10, 4.11, and 4.12 show a comparison of the actual price and the sentiment analysis-predicted price for Bitcoin, Ethereum, and Litecoin, respectively. Figure 4.10 shows that the highest peak, which happens on November 9, 2022, is nearly identical for both the predicted and actual price graphs, indicating a more accurate prediction. The identical situation will play out for Ethereum (Figure 4.11) on November 28, 2022, and for Litecoin (Figure 4.12) on November 18, 2022. Therefore, it's obvious that adding sentiment data to cryptocurrency price predictions makes them more accurate.

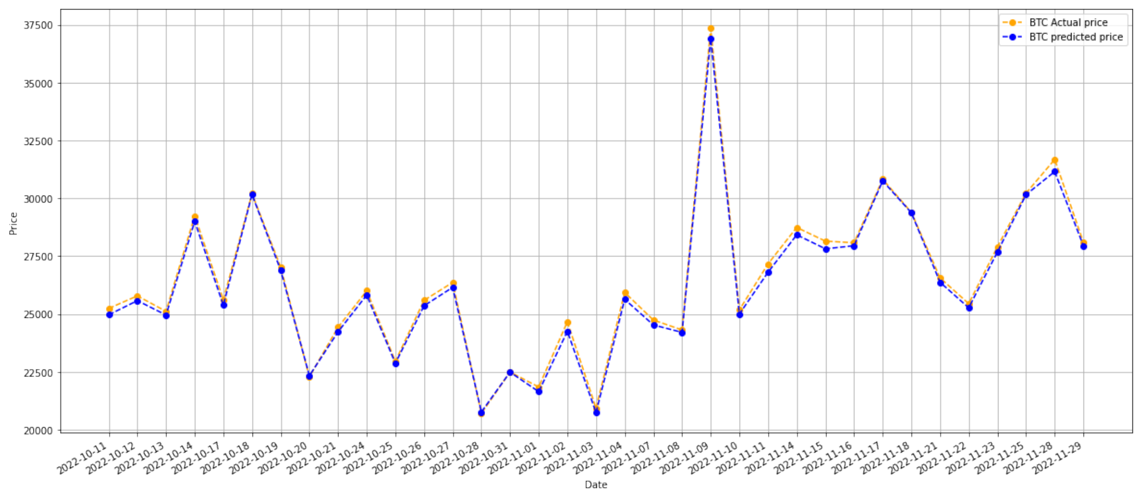

Figure 4.10: Actual and predicted price on Bitcoin through sentiment analysis

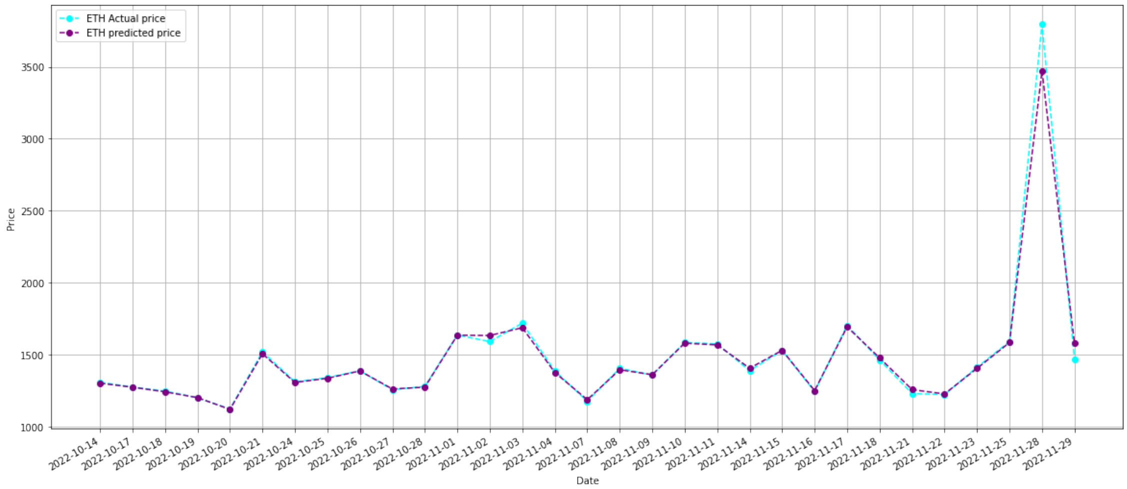

Figure 4.11: Actual and predicted price on Ethereum through sentiment analysis

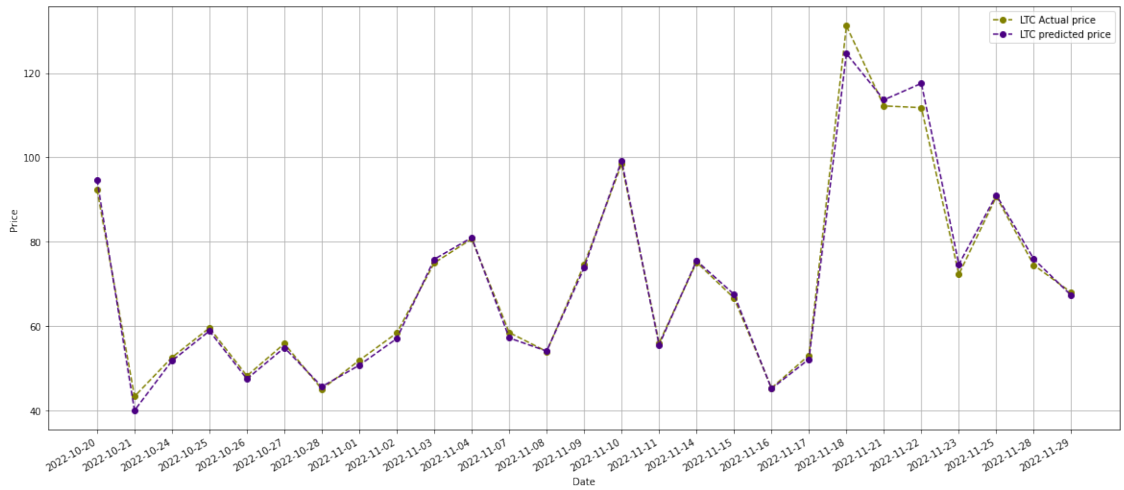

Figure 4.12: Actual and predicted price on Litecoin through sentiment analysis

The sentiment data's R2-score for Bitcoin, Ethereum, and Litecoin was 99.29, 98.28, and 99.75 percent when combined with stock and cryptocurrency price data. Combining these two sources of data can result in a more accurate prediction of the price since it has been seen that variations in the price of Bitcoin, Ethereum, and Litecoin are associated with both the stock market and the tweets of individuals starting from the year 2022. Figures 4.13, 4.14, and 4.15 compare the predicted price obtained from the LSTM time series model on the combined dataset with the corresponding prices for Bitcoin, Ethereum, and Litecoin. Figure 4.13 shows the peak point in the graph where the predicted and actual prices are the most close, as opposed to figure 4.10, which just used sentiment data to illustrate. This peak point occurs on November 9, 2022. For both Ethereum (Figure 4.14) and Litecoin, the fact is equally obvious (Figure 4.15). The study suggests that using sentiment data as well as stock data to predict cryptocurrency prices outperforms using sentiment data alone, which simply shows the relationship between stock and cryptocurrency prices.

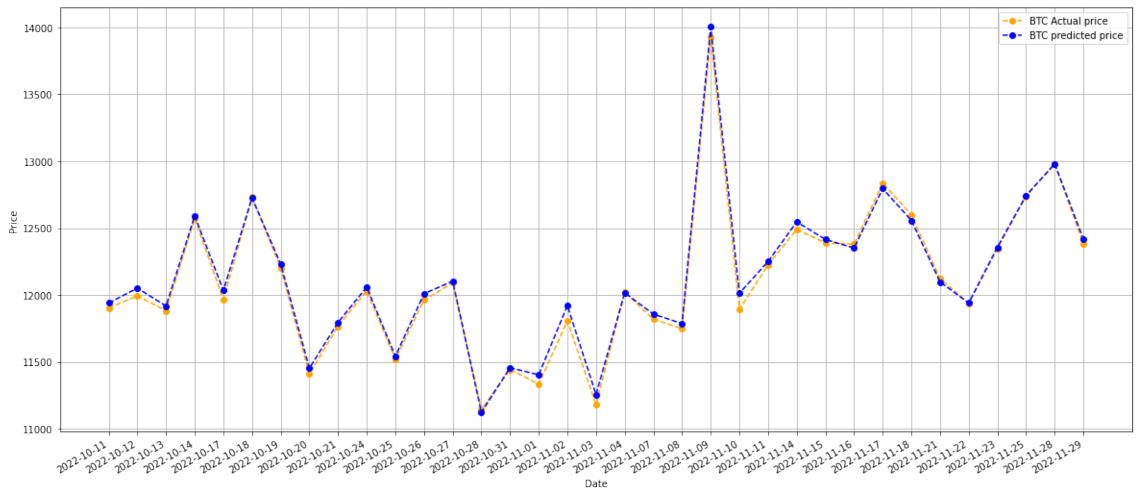

Figure 4.13: Actual and predicted price on Bitcoin through LSTM time series model

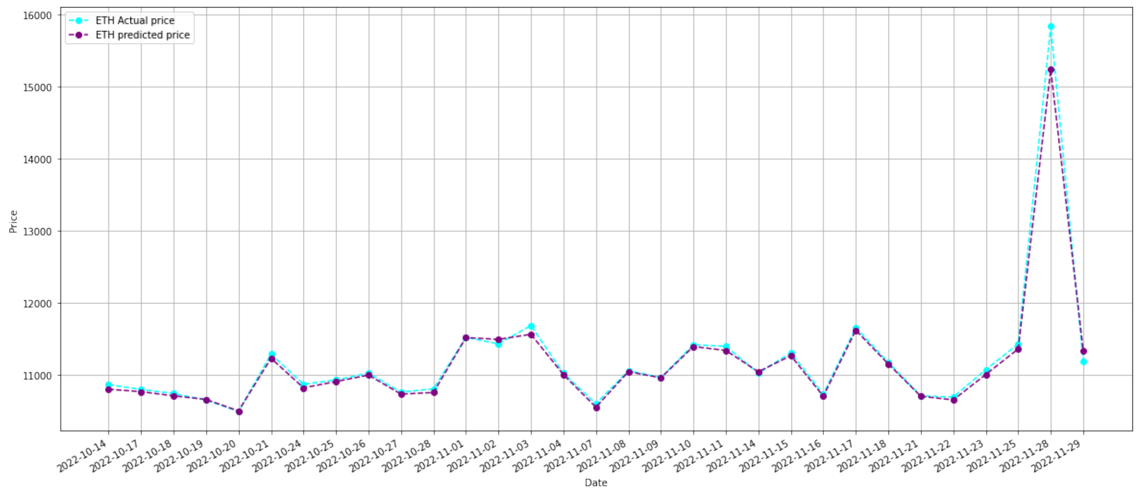

Figure 4.14: Actual and predicted price on Ethereum through LSTM time series-model

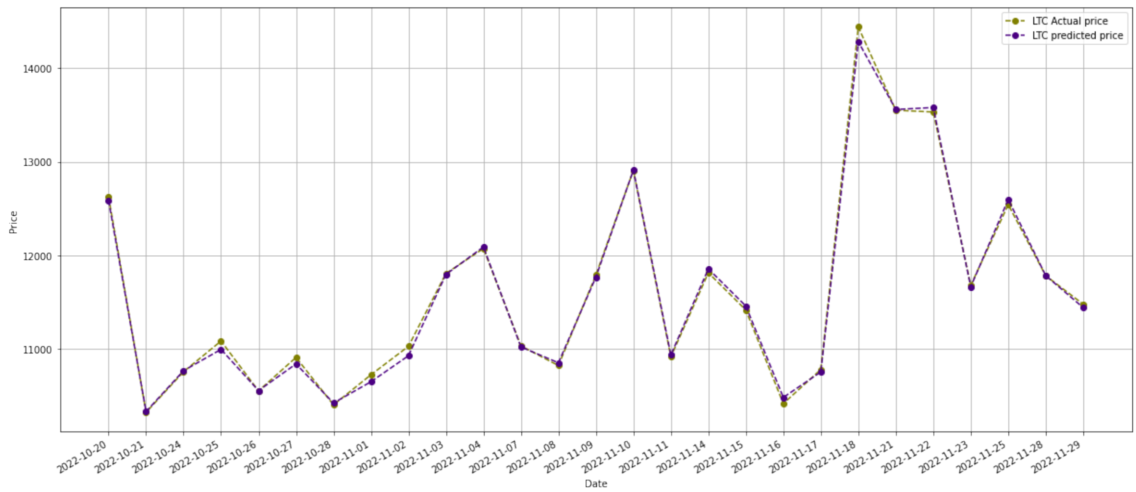

Figure 4.15: Actual and predicted price on Litecoin through LSTM time seriesmodel

Chapter 5

Conclusion and Future Work

Cryptocurrency price prediction has a great impact due to the technological advancement of blockchain. Cryptocurrencies have a great deal of value, such as faster and less expensive financial transactions and decentralized systems that do not crash at an individual point of failure. This technology can ensure safe transactions. Also, the reliability of the technology is making cryptocurrency an investment medium for the general people. Stock market has been an investment opportunity throughout the global market. The absolute goal is that the proposed model of this paper will evoke cryptocurrency price prediction more accurately. Usually, the price of cryptocurrency gets affected by social and psychological factors. That is why it has been difficult for the researchers to predict the price of cryptocurrency. The financial market is predicted and studied using a variety of time series models, including GARCH and ARIMA. In this study, we also predicted the price of cryptocurrency using the ARIMA model and compared the outcome to other methods.But in the case of the time series models, there are some weaknesses, such as the fact that ARIMA doesn't act better in the long run. Also, practicality is a major issue. For these reasons, with nonuniform data, the accuracy decreases. Some ML algorithms are also widely used to forecast the price of crypto. Random forest, KNN, SVM are some of the examples. But now-a-days, some DL algorithms have been so popular in many financial markets to predict the prices. GRU, LSTM, and the hybrid model of GRU and LSTM are examples of some of the algorithms. The LSTM and LSTM-GRU hybrid models have been compared in this paper. Accuracy is the primary factor. The LSTM model has been proposed for predicting after comparison. Additionally, we have focused on sentimental aspects like tweets. The proposed model has demonstrated the accuracy of the prediction results in the case of sentiment.The prediction was made more precise by including sentiment and stock market data. That suggests that, when compared to other indicators like sentiment and the stock market, cryptocurrencies are becoming more relevant and stable.
In this study, we used sentiment data from English-language tweets. We'll be attempting to scrape data from more social media sites in the future, like Facebook, LinkedIn, and Reddit, and use it as sentiment data. In addition, our proposed model demonstrates the prediction. We'll add forecasting functionality in the future.

Bibliography

[1] G. Tsibouris and M. Zeidenberg, "Back propagation as a test of the efficient markets hypothesis," in *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, vol. iv, 1992, 523–532 vol.4. DOI: 10.1109/HICSS.1992.183443.

[2] *Proceedings of International Conference on Neural Networks (ICNN'96), Washington, DC, USA, June 3-6, 1996*, IEEE, 1996. [Online]. Available: https://ieeexplore.ieee.org/servlet/opac?punumber=3927.

[3] K. KOHARA, T. ISHIKAWA, Y. FUKUHARA, and Y. NAKAMURA, "Stock price prediction using prior knowledge and neural networks," *Intelligent Systems in Accounting, Finance and Management*, vol. 6, no. 1, pp. 11–22, 1997. DOI: https://doi.org/10.1002/(SICI)1099-1174(199703)6:1<11::AID-ISAF115>3.0.CO;2-3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/.

[4] R. Tsaih, Y. Hsu, and C. C. Lai, "Forecasting sp 500 stock index futures with a hybrid ai system," *Decision Support Systems*, vol. 23, no. 2, pp. 161–174, 1998, ISSN: 0167-9236. DOI: https://doi.org/10.1016/S0167-9236(98)00028-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923698000281.

[5] M. Zekic-Susac, N. Sarlija, and M. Benšić, "Small business credit scoring: A comparison of logistic regression, neural network, and decision tree models," Jul. 2004, 265–270 Vol.1, ISBN: 953-96769-9-1. DOI: 10.1109/ITI.2004.241696.

[6] P. Panzarasa, T. Opsahl, and K. M. Carley, "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 911–932, 2009. DOI: https://doi.org/10.1002/asi.21015. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21015.

[7] R. Grinberg, "Bitcoin: An innovative alternative digital currency," *Hastings Science Technology Law Journal*, vol. 4, Dec. 2011.

[8] F. Glaser, K. Zimmermann, M. Haferkorn, and M. C. Weber, "Bitcoin - asset or currency? revealing users' hidden intentions," *PSN: Exchange Rates & Currency (International) (Topic)*, 2014.

[9] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653–7670, 2014, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.06.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414003455.

[10] A. H. Dyhrberg, "Bitcoin, gold and the dollar – a garch volatility analysis," *Finance Research Letters*, vol. 16, pp. 85–92, 2016, ISSN: 1544-6123. DOI: https://doi.org/10.1016/j.frl.2015.10.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1544612315001038.

[11] C. Lamon, E. Nielsen, and E. Redondo, "Cryptocurrency price prediction using news and social media sentiment," 2017.

[12] R. Pimprikar, S. Ramachandran, and K. T. Senthilkumar, "Use of machine learning algorithms and twitter sentiment analysis for stock market prediction," 2017.

[13] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018, ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2017.11.054. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221717310652.

[14] J. C. Gerlach, G. Demos, and D. Sornette, "Dissection of bitcoin's multiscale bubble history," *Econometric Modeling: Capital Markets - Asset Pricing eJournal*, 2018.

[15] D. R. Pant, P. Neupane, A. Poudel, A. Pokhrel, and B. Lama, "Recurrent neural network based bitcoin price prediction by twitter sentiment analysis," *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, pp. 128–132, 2018.

[16] R. C. Phillips and D. Gorse, "Mutual-excitation of cryptocurrency market returns and social media topics," in *Proceedings of the 4th International Conference on Frontiers of Educational Technologies*, ser. ICFET '18, Moscow, Russian Federation: Association for Computing Machinery, 2018, pp. 80–86, ISBN: 9781450364720. DOI: 10.1145/3233347.3233370. [Online]. Available: https://doi.org/10.1145/3233347.3233370.

[17] X. Sun, M. Liu, and Z. Sima, "A novel cryptocurrency price trend forecasting model based on lightgbm," *Finance Research Letters*, vol. 32, Dec. 2018. DOI: 10.1016/j.frl.2018.12.032.

[18] A. S. Hu, C. A. Parlour, and U. Rajan, "Cryptocurrencies: Stylized facts on a new investible instrument," *Financial Management*, vol. 48, no. 4, pp. 1049–1068, 2019. DOI: https://doi.org/10.1111/fima.12300. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/fima.12300. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/fima.12300.

[19] G. Swankie, "Examining the price dynamics of the cryptocurrency market and predicting bitcoin price through the application of statistical analysis and deep learning," Ph.D. dissertation, Aug. 2019. DOI: 10.13140/RG.2.2.13654.40003.

[20] ——, "Examining the price dynamics of the cryptocurrency market and predicting bitcoin price through the application of statistical analysis and deep learning," Ph.D. dissertation, Aug. 2019. DOI: 10.13140/RG.2.2.13654.40003.

[21] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, "Price movement prediction of cryptocurrencies using sentiment analysis and machine learning," *Entropy*, vol. 21, no. 6, 2019, ISSN: 1099-4300. DOI: 10.3390/e21060589. [Online]. Available: https://www.mdpi.com/1099-4300/21/6/589.

[22] E. Abakah, L. Gil-Alana, G. Madigu, and F. Romero-Rojo, "Volatility persistence in cryptocurrency markets under structural breaks," *International Review of Economics amp; Finance*, vol. 69, no. C, pp. 680–691, 2020. [Online]. Available: https://EconPapers.repec.org/RePEc:eee:reveco:v:69:y:2020:i:c:p:680-691.

[23] L. A. Gil-Alana, E. J. A. Abakah, and M. F. R. Rojo, "Cryptocurrencies and stock market indices. are they related?" *Research in International Business and Finance*, vol. 51, p. 101 063, 2020, ISSN: 0275-5319. DOI: https://doi.org/10.1016/j.ribaf.2019.101063. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0275531919303472.

[24] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020. [Online]. Available: https://EconPapers.repec.org/RePEc:oup:rfinst:v:33:y:2020:i:5:p:2223-2273..

[25] O. Holovatiuk, "Cryptocurrencies as an asset class in portfolio optimisation," *Central European Economic Journal*, vol. 7, pp. 33–55, Sep. 2020. DOI: 10.2478/ceej-2020-0004.

[26] P. Jaquart, D. Dann, and C. Martin, "Machine learning for bitcoin pricing - a structured literature review," in *Wirtschaftsinformatik*, 2020.

[27] X. Huang, W. Zhang, Y. Huang, X. Tang, M. Zhang, J. Surbiryala, V. Iosifidis, Z. Liu, and J. Zhang, "Lstm based sentiment analysis for cryptocurrency prediction," in *International Conference on Database Systems for Advanced Applications*, 2021.

[28] H. Schoen, D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor, *The power of prediction with social media*, Post-print.