



Arab American University
Faculty of Graduate Studies

**Multi-Channel Classifier for Analyzing External Influences and
Factors on Cryptocurrency Price Using Machine Learning
Techniques**

By

Mohammad Joudeh Atallah Hamayel

Supervisor

Dr. Amani Yousef Owda

Co- Supervisor

Dr. Majdi Owda

**This thesis was submitted in partial fulfillment of the requirements
for the Master`s degree in**

Data Science and Business Analytics

01 /2023

© Arab American University –2023. All rights reserved.

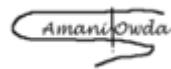
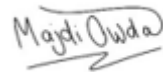


Thesis Approval

Multi-Channel Classifier for Analyzing External Influences and Factors on Cryptocurrency Price Using Machine Learning Techniques

By

Mohammad Joudeh Atallah Hamayel

This thesis was defended successfully on 21/1/2023 and approved by:

Committee members	Signature
1. Dr. Amani Owda: Supervisor	
2. Dr. Majdi Owda: Co- Supervisor	
3. Prof. Mohammad Awad: Internal Examiner	
4. Dr. Mohammad Jubran: External Examiner	

Declaration

I certify that this thesis titled “Multi-Channel Classifier for Analyzing External Influences and Factors on Cryptocurrency Price Using Machine Learning Techniques” is my work, has been completely done by myself, and does not contain work from other related works or research done in the same domain, and only submitted especially to complete a master’s degree in data science and business analytics.

The Name of The Student: Mohammad Joudeh Atallah Hamayel

ID: 202012260

Signature:

A handwritten signature in blue ink, appearing to be 'Mohammad Joudeh Atallah Hamayel', written in a cursive style.

Date: 07/10/2024

Dedication

I want to express my gratitude to my kids Ahmad, Hadi, and Dana for their support and for standing beside me, I present this work to my father, my mother, my brother's soul, The Arab American University for this opportunity, and to those who believe that there is nothing impossible so you have to work hard.

Acknowledgment

I would to exploit this small space to present big words of gratitude, thanks, and, appreciation to my supervisors Dr. Amani Owda and co-supervisor Dr. Majdi Owda for their support, and instructions that help and guide me to achieve this work and for putting me the correct direction, also share with me from their experience with pleasure.

Abstract

Trading in cryptocurrency markets is one of the most resources that the investor exploits to get a lot of profits. There are multiple techniques used such as speculation by exploiting the fluctuation in a cryptocurrency price to reach their goal. This investment is very risky so investors (players) are very careful when deciding to buy or sell in cryptocurrency markets. These types of markets attract attention from researchers, investors, and humans to study or invest or to make challenges, so they study the cryptocurrency markets to understand their behavior and nature.

This thesis aims to investigate the internal and external effects on cryptocurrency markets to predict the price. In general, the factors and influences are social media (Twitter), news articles, cost of production, demand, supply, and gold price. All of those are used to create a comprehensive, robust, and reliable environment and then extract the best feature from each influence (usually called a channel) and the best feature usually called the improved chromosome that forms the model.

To achieve this goal, this thesis targeted four types of cryptocurrency that are Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), and Dogecoin (DOGE). These cryptocurrencies are considered the most popular cryptocurrencies, and all of these have experimented with using the same channels and conditions. In addition to Twitter and news articles channels; for those I used two types of sentiment analysis namely; the VADER sentiment analysis and the Harvard IV-4. The outcomes indicate that the VADER lexicon and rule-based sentiment analysis tool are achieving a better result than Harvard IV-4 for the Twitter channel and the opposite for the news

articles channel. Therefore, the outcomes using VADER sentiment analysis represented the Twitter channel, and the outcome from Harvard IV-4 represented the news articles channels.

The data collection and preparation process was the most difficult and complex stage in this thesis. For example, there are more than 220 Million records collected from Twitter to represent the Twitter channel, on the other hand, interacting with this volume of data represents a big challenge. Data preprocessing for text was processed using multiple stages of processing such as removing duplication, null values processing, removing stop words, removing punctuations, and data completion to achieve this long and complex process. As a result, I have proposed a process called Divide, Clean, and Combine method (DCC).

Moreover, interacting with non-English Tweets shaped is another challenge from a sequence of challenges. To interact with non-English Tweets, I proposed a translation process that is complex and takes a long time of preprocessing. Although the translation process may be affected negatively in some scenarios; however, the results indicate that the model created by the Twitter channel with English and non-English (translated tweets) presents better accuracy than that the model depends on English tweets only.

The proposed technique that the thesis depends on is a data fusion technique. The technique aims to collect the data from multiple data sources and choose the represented feature from each channel to be fusion as a single dataset that is used to train and test the model by Long Short-Term Memory (LSTM). The simulation results show that the fusion techniques enhance the result rather than using a single channel. The evaluation was conducted by using mean absolute squared error

(MASE). In general, for all cryptocurrencies targeted in this thesis, the result was satisfactory for each targeted cryptocurrency, and the final result was 5.62%, 3.48%, and 2.88% for BTC, ETH, and LTC respectively, and for DOGE there were no effects for data fusion techniques in the result.

In future work, I will implement more algorithms and do a comparison between the targeted algorithms and try to get more channels, study the effects of these channels in the proposed model in this thesis model, and train the model on newly collected data.

Table of Contents

Contents	Page
Thesis Approval	i
Declaration	ii
Dedication	iii
Acknowledgment	iv
Abstract	v
List of Tables	xiii
List of Figures	xiv
List of Abbreviations	xxiii
Chapter One: Introduction	
1.1 Cryptocurrency introduction	1
1.2 Research objectives	3
1.3 Research limitation	4
1.4 Contribution	5
1.5 Thesis structure	6
Chapter Two: Background and literature review	
2.1 Literature review	7
2.2 Background	7
2.3 Cryptocurrency market	7
2.4 Cryptocurrency	9
2.4.1 Bitcoin	9

2.4.2 Ethereum	9
2.4.3 Litecoin	10
2.5 Blockchain (BC)	10
2.6 Cryptocurrency price prediction	14
2.6.1 Cryptocurrency price prediction through twitter channel	15
2.6.2 Cryptocurrency price prediction through news agency and financial forum channel	21
2.6.3 Cryptocurrency price prediction using google trends	22
2.6.4 Cryptocurrency price prediction using price historical data	24
2.6.5 Cryptocurrency price prediction and machine learning and statistical methods	27
2.6.6 Cryptocurrency price prediction using block-chain information	29
2.7 Data Fusion	30
2.8 Conclusion	31
Chapter Three: the data and data collection methods	
3.1 Introduction	33
3.2 Scrapping Tools	33
3.3 Data collection from twitter	35
3.4 Data collection from news agencies	44
3.5 Data collection for prices and trading information	45
3.6 Conclusion	46

Chapter Four: Exploratory and explanatory data analysis	
4.1 Introduction	47
4.2 Cryptocurrency financial information	47
4.2.1 Closing price	47
4.2.2 Open, high, and low prices	51
4.2.3 Volume	53
4.3 Blockchain and market information	56
4.3.1 Average hash rate	56
4.3.2 Average block size	61
4.3.3 Market capitalization	64
4.3.4 Mining difficulty	67
4.3.5 Transaction fees	69
4.4 Google trend	71
4.5 News agencies	73
4.6 Social media (Twitter)	78
4.7 Conclusion	80
Chapter Five: The proposed methodology and models	
5.1 Introduction	81
5.2 Feature selection	82
5.3 Long Short-Term Memory (LSTM) Algorithm	82
5.4 Data preprocessing and completion	84
5.4.1 Data Cleaning	85

5.4.2 Remove punctuation	85
5.4.3 Remove stop words	85
5.5 Price prediction based on English twitter data	85
5.5.1 Sentiment analysis based on VADER lexicon sentiment analysis for text analysis.	86
5.5.2 Sentiment analysis based on HARVARD IV-4 dictionary for text analysis	95
5.6 Price prediction based on news agency data	100
5.6.1 Predicting cryptocurrency using news channel based on VADER sentiment for text analysis	100
5.6.2 Predicting cryptocurrency using news channel based on Harvard IV-4 dictionary for text analysis	103
5.7 Price prediction based on google trends	105
5.8 Price prediction based on data fusion	109
5.9 Conclusion	114
Chapter Six: Results and evaluation	
6.1 Introduction	115
6.2 Cryptocurrency price prediction using social media (Twitter)	116
6.2.1 Prediction depends on English tweets	116
6.2.2 Prediction depends on non-English tweets	123
6.3 Cryptocurrency price prediction using news agencies	124
6.4 Cryptocurrency price prediction using google trend	130

6.5 Cryptocurrency price prediction using data fusion techniques	134
6.6 Conclusion	137
Chapter Seven: Conclusion and future work	
7.1 Conclusion.	138
7.2 Future work	139
References	140
المخلص	156

List of Tables

Number	Table	Page
3.1	The search keys are used in the Twitter scrapping process	36
3.2	Number of scrapped tweets vs clean tweets for Ethereum	37
3.3	Number of scrapped tweets vs clean tweets for Bitcoin	37
3.4	Number of scrapped tweets vs clean tweets for Litecoin	38
3.5	Number of scrapped tweets vs clean tweets for DOGE	38
3.6	Tweets distribution per language for the top 10 languages	39
3.7	Number of scrapped news	46
6.1	Predictions of cryptocurrency prices by Tweets channel	117
6.2	Predictions of cryptocurrency prices by news articles channel	125
6.3	Predictions of cryptocurrency prices by using Google trend	131
6.4	Predictions of cryptocurrency prices using data fusion techniques	135

List of Figures

Number	Figure	Page
1.1	A Block diagram shows the thesis structure	5
2.1	Cryptocurrency workflow through blockchain	12
2.2	Blockchain environment requirements	13
3.1	Attribute available through snsrape	35
3.2	Data collection method	36
3.3	Scrapped Tweets distribution per cryptocurrency	39
3.4	BTC, ETH, LTC, and DOGE Tweets Distribution per Language	42
3.5	BTC screenshot for undefined language	43
3.6	ETH screenshot for undefined language	43
3.7	LTC screenshot for undefined language	43
3.8	DOGE screenshot for undefined language	43
3.9	The scrapped news distribution per Cryptocurrency	45
4.1	Closing Price for BTC, ETH, LTC, and DOGE	48
4.2	Correlation matrix for the BTC, ETH, LTC, and DOGE	50
4.3	Boxplot for closing Price for BTC, ETH, LTC, and DOGE	51
4.4	Open, high, and low prices for BTC	52
4.5	Open, high, and low prices for ETH	52
4.6	Open, high, and low prices for LTC	53
4.7	Open, high, and low prices for DOGE	53

4.8	BTC Trading Volume	55
4.9	ETH Trading Volume	55
4.10	LTC Trading Volume	56
4.11	DOGE Trading Volume	56
4.12	BTC average hash rate	57
4.13	Correlation matrix between BTC closing price and average hash rate	58
4.14	ETH average hash rate	58
4.15	Correlation matrix between ETH closing price and average hash rate	59
4.16	LTC average hash rate	59
4.17	Correlation matrix between LTC closing price and average hash rate	60
4.18	LTC average hash rate	60
4.19	Correlation matrix between Doge closing price and average hash rate	61
4.20	BTC, ETH, LTC, and DOGE average BLOCK size	62
4.21-A	The correlation between trading volume and block size for BTC average block size	63
4.21-B	The correlation between trading volume and block size for ETH average block size	63
4.21-C	The correlation between trading volume and block size for LTC average block size	64

4.21-D	The correlation between trading volume and block size for DOGE average block size	64
4.22	BTC, ETH, LTC, and DOGE market capitalization	65
4.23-A	The correlation between trading volume, market capitalization, and the trading volume for BTC	66
4.23-B	The correlation between trading volume, market capitalization, and the trading volume for ETH	66
4.23-C	The correlation between trading volume, market capitalization, and the trading volume for LTC	67
4.23-D	The correlation between trading volume, market capitalization, and the trading volume for DOGE	67
4.24-A	BTC average mining difficulty	68
4.24-B	ETH average mining difficulty	69
4.24-C	LTC average mining difficulty	69
4.24-D	DOGE average mining difficulty	70
4.25	BTC, ETH, LTC, and DOGE average transaction fees	71
4.26	BTC, ETH, LTC, and DOGE correlation matrix for transaction fees, closing price, and transaction volume	72
4.27	BTC trend distribution over the world and top 10 countries	73
4.28	ETH trend distribution over the world and top 10 countries	73
4.29	LTC trend distribution over the world and top 10 countries	73
4.30	DOGE trend distribution over the world and top 10 countries	74
4.31	News Volume for BTC	74

4.32	The top 4 BTC news sources	75
4.33	News Volume for ETH	76
4.34	Top 4. Top 10 ETH news sources	76
4.35	News Volume for LTC	77
4.36	Top 4. LTC news sources	77
4.37	News Volume for DOGE	78
4.38	Top 4 Cryptocurrency news sources	78
4.39	BTC daily tweets	79
4.40	ETH daily tweets	79
4.41	LTC daily tweets	80
4.42	DOGE daily tweets	80
5.1	Proposed system to predict cryptocurrency prices	82
5.2	LSTM working mechanism architecture	84
5.3	The systematic process of tweets sentiment analysis	87
5.4	VADER sentiment analysis measurement concept	88
5.5	Positive and negative tweets distribution for BTC	89
5.6	Positive and negative tweets distribution for ETH	89
5.7	Positive and negative tweets distribution for LTC	90
5.8	Positive and negative tweets distribution for DOGE	90
5.9-A	BTC VADER Sentiment analysis output parameters	91
5.9-B	BTC VADER Sentiment analysis output parameters correlation matrix	92
5.10-A	ETH VADER Sentiment analysis output parameters	92

5.10-B	ETH VADER Sentiment analysis output parameters correlation matrix	93
5.11-A	LTC VADER Sentiment analysis output parameters	93
5.11-B	LTC VADER Sentiment analysis output parameters correlation matrix	94
5.12-A	DOGE VADER Sentiment analysis output parameters	95
5.12-B	DOGE VADER Sentiment analysis output parameters correlation matrix	95
5.13-A	BTC Harvard IV-4 Sentiment analysis output parameters	97
5.13-B	LTC Harvard IV-4 Sentiment analysis output parameters correlation matrix	97
5.14-A	ETH Harvard IV-4 Sentiment analysis output parameters	98
5.14-B	ETH Harvard IV-4 Sentiment analysis output parameters correlation matrix	98
5.15-A	LTC Harvard IV-4 Sentiment analysis output parameters	99
5.15-B	LTC Harvard IV-4 Sentiment analysis output parameters correlation matrix	99
5.16-A	DOGE Harvard IV-4 Sentiment analysis output parameters	100
5.16-B	DOGE Harvard IV-4 Sentiment analysis output parameters correlation matrix	100
5.17	The proposed model is to predict cryptocurrency prices using a news channel	101

5.18	BTC VADER Sentiment analysis output parameters correlation matrix for the news channel	102
5.19	ETH VADER Sentiment analysis output parameters correlation matrix for the news channel	102
5.20	LTC VADER Sentiment analysis output parameters correlation matrix for the news channel	103
5.21	DOGE VADER Sentiment analysis output parameters correlation matrix for the news channel	103
5.22	BTC Harvard IV-4 Sentiment analysis output parameters correlation matrix for the news channel	104
5.23	ETH Harvard IV-4 Sentiment analysis output parameters correlation matrix for the news channel	105
5.24	LTC Harvard IV-4 Sentiment analysis output parameters correlation matrix for the news channel	105
5.25	DOGE Harvard IV-4 Sentiment analysis output parameters correlation matrix for the news channel	106
5.26	Cryptocurrency price prediction using google trend	107
5.27-A	BTC Google trend correlation matrix with BTC price	107
5.27-B	BTC Google trend vs Price time series	108
5.28-A	ETH Google trend correlation matrix with ETH price	108
5.28-B	ETH Google trend vs Price time series	108
5.29-A	LTC Google trend correlation matrix with LTC price	109
5.29-B	LTC Google trend vs Price time series	109

5.30-A	DOGE Google trend correlation matrix with DOGE price	109
5.30-B	DOGE Google trend vs Price time series	110
5.31	Data fusion model to predict cryptocurrency price prediction	111
5.32	Data fusion correlation matrix for BTC	112
5.33	Data fusion correlation matrix for ETH	113
5.34	Data fusion correlation matrix for LTC	114
5.35	Data fusion correlation matrix for DOGE	115
6.1	The Result Extraction criteria	116
6.2	BTC price prediction using Twitter and VADER Sentiment analysis	118
6.3	BTC price prediction using Twitter and HARVARD IV-4 Sentiment analysis	119
6.4	ETH price prediction using Twitter and VADER Sentiment analysis	120
6.5	ETH price prediction using Twitter and Harvard Sentiment analysis	120
6.6	LTC price prediction using Twitter and VADER Sentiment analysis	121
6.7	LTC price prediction using Twitter and Harvard IV-4 Sentiment analysis	122
6.8	DOGE price prediction using Twitter and VADER Sentiment analysis	122

6.9	DOGE price prediction using Twitter and Harvard IV-4 Sentiment analysis	123
6.10	LTC price prediction using depends on English and Non-English Tweets	124
6.11	BTC price prediction using News channels VADER sentiment analysis	126
6.12	ETH price prediction using News channel - VADER sentiment analysis	126
6.13	LTC price prediction using News channel – VADER sentiment analysis	127
6.14	DOGE price prediction using News channel – VADER sentiment analysis	127
6.15	BTC price prediction using News channel – Harvard IV-4 sentiment analysis	128
6.16	ETH price prediction using News channel – Harvard IV-4 sentiment analysis	129
6.17	LTC price prediction using News channel – Harvard IV-4 sentiment analysis	129
6.18	DOGE price prediction using News channel – Harvard IV-4 sentiment analysis	130
6.19	BTC price prediction using Google trend	132
6.20	ETH price prediction using Google trend	132
6.21	LTC price prediction using Google trend	133

6.22	DOGE price prediction using Google trend	134
6.23	Comparison between actual and predicted values for BTC prediction using data fusion techniques	136
6.24	Comparison between actual and predicted values for LTC prediction using data fusion techniques	137
6.25	Comparison between actual and predicted values for ETH prediction using data fusion techniques	137

List of Abbreviations

ADA	Cardano.
ANN	Artificial Neural Networks.
ARMA	Autoregressive Moving average.
BCH	Bitcoin Cash.
BNN	Bayesian Neural Network.
BNN	Binarized Neural Network
BTC	Bitcoin.
CNN	Convolutional Neural Networks.
DCC	Divide, Clean, and, Combine method
DF	Data Fusion
DOGE	Dogecoin.
EMH	Efficient Market Hypothesis.
ETH	Ethereum.
GARCH	Generalized Autoregressive Conditional Heteroscedasticity.
KNN	K-Nearest Neighbors.
LR	Logistic Regression.
LSTM	Long-short Term Short Memory.
LSTM	Long Short-Term Memory.
LTC	Litecoin.
MAPE	Mean Absolute Percentage Error.
MLP	Multilayer Perceptron

NB	Naive Bayes.
NN	Neural Network.
RF	Random Forest.
RNN	Recurrent neural networks.
RRN	Recurrent Neural Network.
SC	Supply Chain.
SGBM	Stochastic Gradient Boosting Machine.
SVM	Support Vector Machine.
SVR	Support Vector Regression.
VADER	Valence Aware Dictionary and Sentiment Reason.
VAR	Vector Autoregression.
XLM	Stellar.
XMR	Monero.
XRP	Ripple.

Chapter One

Introduction

1.1 Cryptocurrency introduction

The history of financial investing began a decade with traditional stock and bond transactions in traditional financial markets, then the addition of financial derivatives, such as securities futures, that are primary components in the financial systems of developed countries [1]. With the revolution in technology, Virtual Currencies (VC) have been created, often known as cryptocurrencies, which are digital assets that are held electronically in digital wallets and traded over the internet via a peer-to-peer system [2]. VC has various features; volatility is the largest issue for cryptocurrency market players, thus it's important to idealize the technique that may be used to characterize cryptocurrency price fluctuations [3]. Cryptocurrency markets are complex and speculative systems. When investors decided to interact using strategies that generate some biasing for endogenous instabilities [4], it considers a type of virtual money that does not have a physical form, as well as a secure transaction currency that cannot be observed. Bitcoin, Ethereum, Litecoin, and Monero are all examples of this digital currency.

Digital Currencies were the first pure digital assets to be incorporated by asset managers. As a result, “summarizing recent research articles and findings on cryptocurrency trading, including trading platforms, trading signals, trading technique research, and risk management, is crucial” [5]. The enormous price fluctuations and intricacy of cryptocurrency markets have prompted a plethora of studies exploring cryptocurrency co-movements in both times and frequency [6].

Moreover, cryptocurrencies have attracted the interest of governments, financial institutions, commercial organizations, media, and academics from a variety of disciplines. Nonetheless, the majority of cryptocurrency research has concentrated primarily on Bitcoin's volatility and returns, neglecting the broader cryptocurrency market's possibilities [7]. Bitcoin and other cryptocurrencies have solidified their position in the financial system. More than 35 million BTC wallets exist, indicating that bitcoin is becoming a widely recognized means of e-payment [8].

Cryptocurrency also uses a ledger to keep digital transaction records, and it is open to the public. Transactions on the public ledger are now safe, transparent, efficient, and everlasting. Users don't need to trust a bank to store bitcoin, and they also don't need to trust the person performing the business to pay the user with the assistance of this ledger. Instead, users may watch money being given, received, confirmed, and recorded by thousands of individuals via them. Since its launch, which coincided with the global financial crisis of 2008 and the resulting lack of confidence in the financial system, bitcoin has occupied a significant position in the global financial landscape and gained the interest of regulators, governmental organizations, institutional and individual investors, academics, and the general public [9]. There is no need for a middleman in this system, and transactions are typically very cheap and quick.

Cryptocurrencies have grabbed the curiosity of many academics and policymakers since the launch of the first digital currency. It mainly depends on cryptography technology; their production and circulation are decentralized. Unlike fiat currencies, digital currencies do not need any intermediation by banks also not controlled or monitored by central banks [10]. Investing in

cryptocurrencies is often based on what is possible in the future rather than what is currently possible. Many scholars who are trying to figure out what cryptocurrencies are, or at least why they are so popular, have discovered that they have many of the features of speculative investment instruments [11].

1.2 Research objectives

In this research, one of the main objectives is to predict the cryptocurrency prices depending on the study the of cryptocurrency environment and the internal and external factors and influences that affect the cryptocurrency prices to make price predictions using machine learning algorithms called Long Short-Term Memory (LSTM) which is one of the most popular algorithms that interact with multivariate time series. The external factors that this work depends on are our social media (Twitter), news articles, demand and supply, and the cost of production and gold prices. The framework of this thesis is summarized as follows:

- Data collection process there are multiple data sources that I targeted to build the proposed model that depends on multiple channels; to achieve this stage I used multiple methods like web scrapping, historical data downloading like CSV files, and using python libraries such as beautifulsoup to collect the data from HTML pages.
- Interacting with big data is one of the challenges as the data file contains more than 220 M records using scrapping methods; to deal with such data I proposed multiple methods like dividing, processing, analyzing, and combining files, that I referred to it as Divide, Clean and Combines method (DCC).

- Data preprocessing stage; in this stage, I cleaned the data, and I used multiple techniques like removing punctuations and stop words and also interacting with the non-informative and noisy data especially the scrapped data from tweets.
- Data visualization using python libraries to get a deep and clear insight into the data shape and distribution and the features correlated with each other.
- Interacting with non-English tweets, because there are no sentiment analysis libraries that interact with non-English tweets, we proposed a special model that depends on the tweet's translation process that translates the text to English text using a python library called textblob, this process is also a complex and takes a long time for execution but it is very important to focus such this tweet.
- Training LSTM machine learning models to predict cryptocurrency prices depending on collected and combined features.
- Comparing the results while using multiple techniques such as VADER sentiment analysis, and Harvard IV-4 to analyze the tweets and news text and comparing the best models to extract the best features.
- Using data fusion techniques to study the cryptocurrency environment and internal and external factors and influences by collecting the data from multiple data sources and fusion them together to predict the cryptocurrency prices after selecting the best features that we called improved chromosome traits for the model.

1.3 Research limitation

The limitation of the study revolved around the limitation of data collection and the time-consuming for this process, another limitation is that related to the data preprocessing and the data quality.

1.4 Contribution

To the best of the author's knowledge; there are no studies in the open literature that used data fusion techniques to predict cryptocurrency prices. This study aims to bridge this gap by studying the cryptocurrency environment in general and internal and external factors and investigates their influences on cryptocurrency price fluctuation using big historical data. Then used the fused data feeds into the machine learning model, and also creates creative techniques in the data processing stage, which include deleting duplication, removing punctuation, and removing stop words. Further, studying every single factor separately we called a channel in our work to find the dominant factors and extract the best features from these channels that we also called enhanced chromosome traits that formulate the model. Finally, the output features from social media and news articles have been analyzed using two different techniques called VADER lexicon sentiment analysis and Harvard IV-4 dictionary, all models were implemented using Long Short-Term Memory (LSTM).

1.5 Thesis structure

The rest of this work is coordinated as illustrated in Figure 1.1.

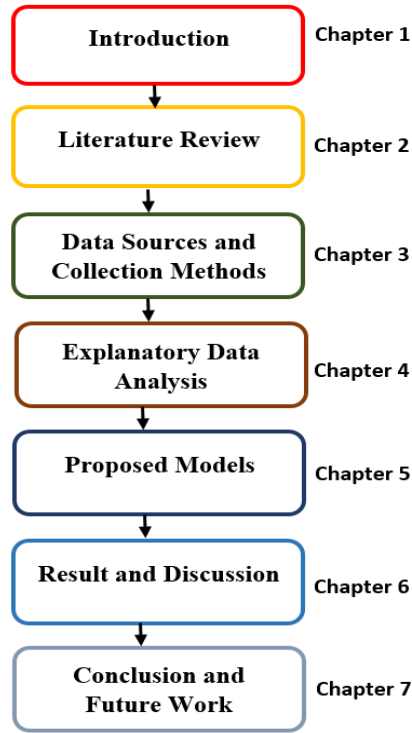


Figure 1.1. A Block diagram shows the thesis structure.

Chapter 2 literature review of the related work in machine learning and statistical analysis for predicting cryptocurrency prices and some other techniques that are used in the domain.

Chapter 3 clarifies the data sources and collection methods.

Chapter 4 presents the exploratory data analysis made and graphically shows the dataset and the element correlation to gain deeper insights from the dataset.

Chapter 5 presents the proposed method for building the prediction model.

Chapter 6 presents and discusses the results for single and fusion models using RMSE and MAPE evaluation measurements.

Chapter 7 represents the conclusion and brief outcomes representation.

Chapter Two

Background and Literature Review

2.1 Literature review

This chapter presents a summary analysis of the related publications of my thesis topics about cryptocurrency price prediction using machine learning algorithms, the factors that affected price fluctuation, the methods and channels that are used in the cryptocurrency price prediction domain, also the cryptocurrency markets components, and the technology that depends on.

2.2 Background

The appearance of blockchain technology in recent years, and the subsequent emergence of digital currencies and blockchain technology take the audience's attention to investment or research proposals, and this pushes them to study the environment and behavior of cryptocurrency markets. Researchers offer multiple methodologies and techniques to achieve their goals. Because of tremendous growth in the use of cryptocurrencies to get profits. However, owing to the market's unpredictable nature and significant price volatility some opinions sub-serve that cryptocurrencies are not viewed as an investment prospect because of the high risk in the investment process.

2.3 Cryptocurrency market

Nowadays financial markets are formed and recognized the characterized by a rapid information flow, a large number of participants with various investment horizons, and numerous feedback systems, all of which combine to produce complex phenomena such as speculative bubbles and collapses [12]. The understanding perspective of financial exposures and the broader dynamics of

financial system risk requires an understanding of inter-dependencies among financial variables that have ramifications for a variety of risk management, asset allocation, and regulation formulation challenges [13]. This is especially true in the cryptocurrency market, which has arisen as a new and exciting forum for both institutions and individuals from all over the world.

In today's market, around 1500 cryptocurrencies are being traded [14]. It is possible to create crypto money that may be used for online transactions. Even though it has no physical form, this money works similarly to other currencies and has an exchange rate [15]. Cryptocurrency markets are one of the most enticing markets that can investors make money by making financial speculation. As a result, they've received unparalleled social media attention. Aside from serious debates and legitimate investment opportunities, several misleading practices have arisen [16]. DC has both positive and negative aspects some of these are highly secure currency and indestructible.

The most serious risks come from this area, particularly cybercriminal actions, which are growing easier to carry out [17]. Cryptocurrency market behavior is very closely similar to the stock market. Therefore, the rule and influences of the stock market may be can help to understand the behavior of cryptocurrency markets. Market patterns in such marketplaces can be generated using historical stock market data, as well as a variety of other factors. According to Efficient Market Hypothesis (EMH), prices include all relevant market information [18].

Regulatory policy and the legal framework are critical in light of the expanding public interest in cryptocurrency circulation, stock market movements, economic rewards, and prospects, as well as the accompanying dangers and problems. Without identifying the legal status of a cryptocurrency, it is difficult to tackle the problem of legal, accounting, and tax ramifications of cryptocurrency transactions.

2.4 Cryptocurrency

Various countries regard cryptocurrencies differently, ranging from classifying them as a form of payment or an asset to outright banning those [19], so the risk and return considerations associated with the usability and adaptability of cryptocurrencies have had a good and negative influence on consumers. There are many issues unique to the cryptocurrency sector; including mining, cybersecurity, anonymity, and privacy. The amount of bitcoin in circulation, the daily volume of cryptocurrency exchanged, demand, and supply is only a few factors from the list of factors that can influence and affect the price of the cryptocurrency [20].

2.4.1 Bitcoin

Because of the increase in market capitalization, bitcoin has grown increasingly popular in 2017 [21]. The actual platform for Bitcoin transactions came into being through the release of the first open-source Bitcoin Client and the concomitant issuing of Bitcoins. Nakamoto mined the first block of Bitcoins with a reward of 50 Bitcoins. This block is commonly referred to as the “genesis block.”. The number of Bitcoin wallet users had surpassed 65 million by January 2021 as reported in [22].

2.4.2 Ethereum

Ethereum sought to be more than straightforward money, while Bitcoin transformed the way cryptocurrencies were implemented by using the blockchain as a mechanism for distributed consensus. The inventor of Ethereum, Vitalik Buterin, saw the value of a blockchain with a built-in Turing-complete programming language that enables its users to create smart contracts that may interact with the underlying blockchain in a variety of more complex ways. We must first examine the disadvantages of scripting in Bitcoin to comprehend the advantages Ethereum has over Bitcoin. There are numerous issues with the Bitcoin scripting protocol [23].

2.4.3 Litecoin

Litecoin (LTC) was introduced in 2011 and is regarded as the "silver" to bitcoin's "gold" due to its larger total quantity of 84 million LTC. It inherits bitcoin's fundamental concepts but modifies several critical elements (The mining algorithm, for example, is based on Scrypt rather than the SHA-265 method used by bitcoin) [24].

2.5 Blockchain (BC)

Blockchain is a technology that has been used in several applications not only for cryptocurrency it's also used for transfers and payments; the payment procedures are more flexible, liquidated, and automated in different circumstances. This is resulting in a paradigm shift in the payment sector [25]. This technology is considered the underlying technology of cryptocurrencies, and it, has gotten a lot of interest from individuals, institutions, and academics, Blockchain's key features include decentralization, consensus, transparency, open-source, identity, data security, autonomy, immutability, and anonymity. All these are significant components of a blockchain, it is built on the principles of security and privacy.

The transactional data in a block is protected from internal, peripheral, and unintended risks [26]. Security may be assured by the detection and prevention of risks, as well as the use of security technologies, rules, and IT services to provide an adequate reaction to threats, also it is the most modern technology for storing and distributing valuable transaction data that has been created in recent years. Furthermore, this technology employs a variety of data storage methods that are distinct from those used in the past, and it represents a technology that is a great example of how to keep your data private and secure.

Because of the importance of blockchain, this pushes researchers and business leaders to need to know where and when blockchain applications will be used and researched [27]. The current uses that have made highlighted the blockchain's latent potential include multiple domains such as the food industry, healthcare, and logistics of supply chains [28]. Efficiencies, transparency, and traceability, as well as billions in corporate financial savings, might all help with supply chain and sustainability concerns [29].

Many people believe blockchain is only about cryptocurrencies like bitcoin, but it is much more. It is a technology that creates immutable and distributed data records that are transferred by persons' in-network database systems and record digital events that cannot be modified or remembered before they arrive at their destinations [30]. Also, it is considered the heart of most cryptocurrencies, but it's also being investigated in other significant domains, such as decentralized autonomous groups that can run without human intervention and new governance structures that allow for more participatory decision-making.

This robust technology also invades many life domains it also can be used in important and critical fields like tracing sensor data and preventing duplication in the internet of things (IoT) applications, healthcare business, and e-voting, and it may be used for any interconnected data chunks [31]. Moreover, it's users in multiple sectors such as financial transactions, user identification, and cybersecurity technology development.

The blockchain is undoubtedly one of those technologies that, in terms of potential application fields, it is a new technology that provides an essential framework for the use of different

cryptocurrencies like bitcoin [32] and is now creating high expectations. It is a global ledger that efficiently and permanently stores transactions on a chain of blocks.

Figure 2.1 depicts a typical blockchain database with two customers, the sender, and the receiver, each of whom wants to transfer money to the other. Each transaction between the client's sender and receiver may be authenticated using distributed ledger or blockchain, which is maintained by all participants. It verifies the transaction's legitimacy at the sender, adds transaction data to global storage at the receiver end, and finally closes the transaction.

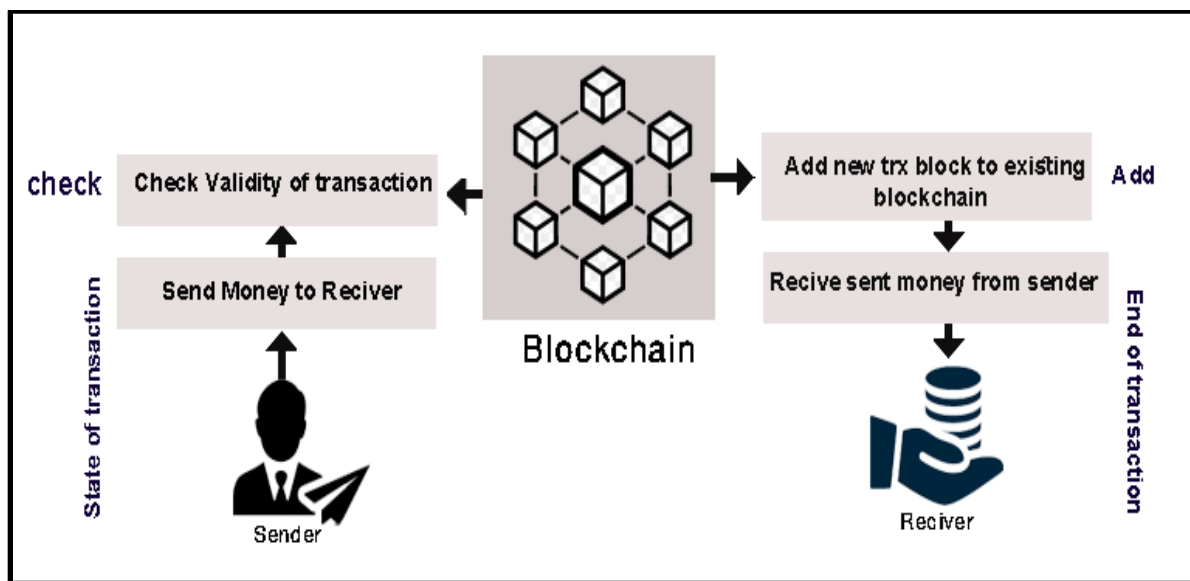


Figure 2.1. Cryptocurrency workflow through blockchain.

Blockchain requirements can be divided into four groups [33], Figure 2.2 illustrated the Blockchain environment requirements.

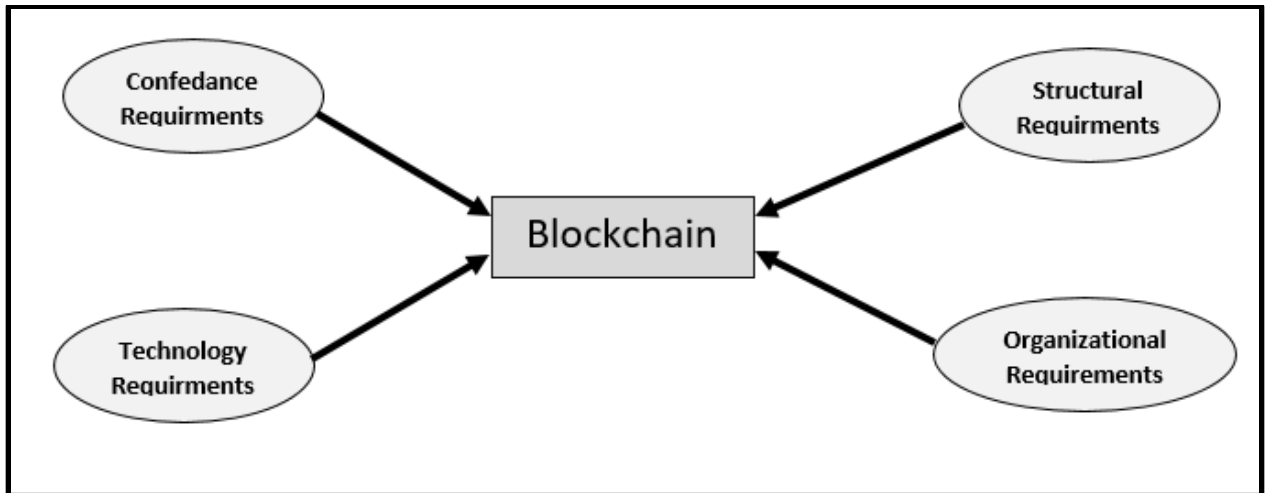


Figure 2.2: Blockchain environment requirements.

- **Structural requirements:** For the described technologies to work, specified types of atoms must be available in blockchain linkages. There may also be a demand for global blockchain cohesion to understand the general structure or a specific subset of blockchain, one or more linkages must exist [34].
- **Organizational requirements:** This is based on national or international cryptography legislation requiring the use of national, recommend, or approved cryptographic tools to create and process blockchain atoms. In addition, requirements relating to national or institutional standards in application domains such as taxes, voting technology, in-house document workflow, and so on may be included in this category [35].
- **Technological requirements:** which are related to blocking link storage reliability, which must aid in the maintenance of storage reliability and availability characteristics for these links. Regulatory organizations in the blockchain application industry establish the parameters. Furthermore, the technological requirements must specify the capability of operations with linkages, as well as the high amounts of data that must be accumulated and stored [36].

- **Confidence requirements:** That provide a high degree of confidence, all applicable interfaces must be published with source codes and have a defined blockchain structure, controlled link processing technologies, and an interface for link operations. Mathematical modeling can also be used to formally verify the technology [36].

2.6 Cryptocurrency price prediction

By focusing on the nature and behavior of cryptocurrency trading and investing, the price oscillations could be influenced by investors' interest in other cryptocurrencies, and social media usage as well as other factors. Cryptocurrency price prediction is one of the hottest topics among researchers because of the popularity of cryptocurrency and it takes the focus of attention from the investor side. Cryptocurrency has no seasonal influences, and this makes it difficult to be forecasted by using statistical methods.

The goal of investing in a financial market is to increase profits; this complex market is impacted by a great number of events, making forecasting future market dynamics difficult. Investors' attitudes toward the stock market may need the research of numerous related elements and the extraction of important information for accurate forecasting. To achieve this, researchers employ multiple techniques to go through this market and go the profits safely as possible they can. They used Traditional statistical approaches, while simple to use and interpret, rely on several statistical assumptions that may be impractical, leaving machine learning as the finest technology in this sector, as it is capable of forecasting pricing based on prior experience. Bayesian Regression, Logistic Regression (LR), Linear regression, Support Vector Machine (SVM), Artificial Neural

Network, Deep Learning, and Reinforcement Learning are some of the traditional statistical and machine-learning methodologies utilized in this field of research.

2.6.1 Cryptocurrency price prediction through twitter channel

Predicting cryptocurrency prices through social media especially Twitter considered the most popular channel that is used. Due to the cryptocurrency's deterministic character, most of the methods published in the literature for cryptocurrency price forecasting may not be suitable for real-time price prediction [37]. Academics are increasingly recognizing Twitter's predictive ability for a wide range of events, researchers found that there is significant evidence between cryptocurrency prices and return with social media, especially Twitter. The fundamental nature of microblog content, such as that seen on social media networks such as Twitter and Facebook, provides significant problems to practical sentiment analysis systems. Some of these issues originate from the sheer volume and rate of user-generated social content, as well as the contextual sparseness caused by the text's short length and a desire to express thoughts using shortened linguistic rules [38].

Using Twitter as a tool to predict the prices price of return relay on public Twitter sentiment, specific lexicon-based sentiment analysis techniques, financial data, and bilateral Granger-causality testing to prove that Twitter can predict [39]. Using Twitter as a cryptocurrency price prediction tool is a good choice but maybe it is not sufficient to face all factors that affect the direction of cryptocurrency price, some of the research interacted with tweets in multiple ways. The study in [40] determined that tweet volume instead of tweet sentiment which is always positive in direction of the price is a better predictor of price direction. Some of the traditional methods

such as statistical methods can be used to analyze tweets and provide an indication to the investors to make buy/sell decisions [41], but maybe the traditional methods do not give a good view because of the high fluctuation in the cryptocurrency market.

Authors in [42] suggest forecasting the price movement of the BTC, ETH, XRP, and LTC cryptocurrency markets using traditional machine learning methods and social media data and examine the use of Neural Networks (NN), SVM, and Random Forest (RF), the findings demonstrate that with the combination of machine learning and sentiment analysis of tweets, it is can be used to predict cryptocurrency markets, and NN outperforms the other models, on the same some opinions backs up a few writers' claims on the influence of social media on cryptocurrency pricing. The focus is initially on BTC, but a similar approach can eventually be applied to other cryptocurrencies [43]. Some directions support the relationship between social media and cryptocurrency prices and others refuse the concept. The study in [44] supports the concept by employing classical machine learning approaches such as SVM and Logistic Regression (LR) to predict the irregular price movement of cryptocurrencies by monitoring social media sentiment and identifying a correlation between the two.

Some studies indicate that the audience opinion affected cryptocurrency prices if they used suitable machine learning algorithms and sentiment analysis. The study in [45] used a lexicon-based approach to analyze reviews on a wide range of cryptocurrencies over tweets to measure sentiment scores that include positive, negative, or neutral sentiments, the statistics also show that more than half of the population has a favorable opinion of cryptocurrencies. Furthermore, if marketers employ suitable categorization algorithms like SVM, they may accurately anticipate the mood of

tweets regarding these cryptocurrencies, moreover study in [46] undertook empirical research that compares the Bayesian neural network BNN to other linear and non-linear benchmark models to predict cryptocurrency price.

Data preprocessing for the text and removing non-required text like punctuations, and stop words have a positive effect on the results when analyzing the text and extracting the sentiment. An example of this is the study in [47]. This study examines several text preparation techniques for connecting the emotional ratings of Twitter content with BTC prices, then investigates the impact of various preprocessing algorithms, characteristics, and data lengths on correlation findings. The main finding is that dividing words, eliminating Twitter-specific tags, or combining the two improves sentiment and volume polarity scores' association with BTC prices in terms of anticipating and modeling the BTC process.

Using non-classical machine learning algorithms like NN algorithms may be a good choice to enforce the fluctuations in time series [48], the empirical tests indicate that BNN is effective at forecasting BTC price time series to explain the current situation of BTC price volatility. Researchers believe that emotion may be a better predictor of Bitcoin's [49] value than standard models. Authors in [50] provide a dataset of Tweets that spans over a year and has been annotated for Sentiment. Furthermore, they present preliminary data from studies that confirm the premise that sentiment information is strongly predictive of value creation. Moreover, authors in [51] describe how they used common AI gadgets and publicly available internet networking. Data to forecast the value of BTC, ETH, XRP, and LTC digital currency display advancements. They also

utilized neural networks, SVMs, and RF to analyze the Twitter components. The findings show that by combining (AI) and hypothesis testing with Twitter data, it is feasible to forecast the direction of cryptocurrency markets. In [52] authors implement deep learning, LR, and Support Vector Regression (SVR) are utilized to forecast Cryptocurrency values, some mega-event emotion is also employed to improve the performance of these models. The findings demonstrate that business intelligence models like deep learning and SVR produce superior results than traditional ones. During the Covid-19 era, each model was assessed on its ability to forecast public mood states as indicated by tweets from Twitter using different machine learning algorithms with Twitter data and historical data, the highlighted result was that tweets data can improve the performance of models [53].

Analysis of social media text depends on text analysis to analyze the text. The study in [54] used different natural language modeling processes such as N-Gram to investigate the relationship between the Twitter feed on BTC and its sentiment analysis, regarding that the results demonstrate that in some datasets, the manual technique produces substantially better results and that Multilayer Perceptron (MLP), WiSARD, and decision tree approaches perform significantly worse. On the other side, the study in [55] examines how well public opinion and emotion from Twitter can be used to estimate the BTC returns. The Twitter sentiment was confirmed to have predictive power for BTC results using a sentiment analyzer on BTC-related tweets with financial data. The findings indicate the existence of a link between them once more. When making predictions based on bitcoin-related tweets, and found a 62.48 % accuracy rate.

By looking at the idea of leveraging data from social media [56] sites to make BTC price forecasts based on sentiment analysis. The data collected from these social media sites is analyzed with the VADER sentiment analysis tools and compared to the historical BTC price, the findings essentially match previous studies on the predictability of Bitcoin fluctuations based on social media attitudes. The study in [57] evaluates the collection, modification, and interpretation of Tweets to forecast early bitcoin market moves. Sentiment analysis and text mining approaches such as LR, SVM, and NB were taken into consideration. Other techniques used in [58] focused on anticipating the Bitcoin price by studying user views on social media platforms like Twitter. RNNs, LSTMs, and CNNs are utilized as deep learning architectures, while Word2Vec, GloVe, and FastText are used as word embedding models in the tests to estimate Bitcoin's direction. Experiment findings reveal that using the FastText model as a word embedding model outperforms other models in predicting the direction of BTC price with an accuracy value of 89.13%.

Authors in [59] aim to forecast the fluctuating price of BTC by monitoring Twitter's mood and establishing a link between the two, the positive and negative feelings are assigned to BTC tweets gathered from various news account sources. The acquired percentage of sentiment score that includes positive and negative tweets, with historical cryptocurrency price, are fed into an RNN model to forecast the new price for the following period. The total price prediction accuracy was found to be 77.62%, while the accuracy for sentiment categorization of tweets in two classes, positive and negative, is found to be 81.39%. Authors in [60] collected a dataset combining prices and social media activity over 71 days to prove variations in Altcoins. Sentiment analysis was then used to estimate the containing public mood. Researchers used 45 days of data to conduct a linear regression study to forecast Altcoin returns. Also, activity and sentiments on Twitter can be used

to anticipate short-term returns. The study in [61] backs up a few writers' claims on the influence of social media on cryptocurrency pricing. The focus is initially on BTC, but a similar approach can eventually be applied to other cryptocurrencies. To forecast prices, sentiment ratings from tweets and news feeds are combined with past prices and volume. The findings of the experiments demonstrate that sentiment scores have little influence unless they are not skewed toward one class.

A study in [62] Got the last five years of daily tweets containing NEO hashtags from Twitter. The tweets were then filtered to exclusively include or reference NEO. Then they manually assigned positive, negative, and neutral sentiment categories to a subset of the tweets. The Random Forest classifier achieved an accuracy of 77%. The second phase of the research [63] looked at whether the daily sentiment of tweets was connected to the price of NEO. The research outcome illustrates that the number of tweets and daily prices, as well as the prices of other cryptocurrencies, had favorable relationships.

Depending on the previous studies there are different directions some of them support Twitter as a prediction channel and others refuse. Other trends support that the cryptocurrency price prediction can be better using the tweets volume rather than the sentiment analysis of the tweets. Many techniques can be used to analyze the text of the tweets in previous studies like text analysis using VADER sentiment analysis. But it is very important to do some preprocessing in the text like removing punctuation and stop words. Using non-classical methods like NN has gotten better results by comparing with classical machine learning algorithms.

2.6.2 Cryptocurrency price prediction through news agency and financial forum channel

Recent research has demonstrated that the huge amount of publicly available internet information, including social media debates, Wikipedia usage patterns, and news stories from the major media, can have an observable impact on investors' perceptions of the financial markets [64].

Some researchers take news agencies as a tool to predict cryptocurrency news. [65] Focusing on the ability and capacity of news and social media data to anticipate the changing in the price of three cryptocurrencies that are BTC, LTC, and ETH. The Text-based sentiment categorization was done using traditional supervised learning algorithms instead of positive or negative sentiment, daily news and social media data were tagged depending on actual price movements one day in the future for each currency instead of having to estimate sentiment first, the algorithm may directly anticipate price changes using this technique. Authors in [66] examine the ability of news data to forecast price changes for Ethereum, which is counted as the second most valuable cryptocurrency by market capitalization. By signaling whether to purchase, sell, or hold, the model can forecast price direction immediately. The use of a collection of Natural Language Processing algorithms to determine public emotions for bitcoin volatility is a critical component of our approach. The impact of news of a cryptocurrency heist on the price of cryptocurrencies is investigated by [67] this involves robberies from big exchanges, thefts of bitcoin technology, or thefts of significant cryptocurrency holders. When news of bitcoin thefts was revealed, an event analysis was done on normalized values of cryptocurrencies for three days.

Other tools that are used to predict cryptocurrency prices are a financial forums study in [68]. The study investigates the interaction between various cryptocurrency forums and volatility in

cryptocurrency, the most popular cryptocurrency such as BTC as well as a related busy discussion forum Bitcointalk. This study demonstrates that the activity of the Bitcointalk forum is directly related to the trend in BTC values; thus, an examination of this relationship would be an excellent foundation for personal investments in an unregulated market, as well as to see if cryptocurrency forums can detect suspicious BTC value activity. The experiment demonstrates how forum data may be used to explain specific financial occurrences.

2.6.3 Cryptocurrency price prediction using google trends

Google trends are considered a good tool to predict cryptocurrency prices. Researchers in [69] investigate the relationship between BTC price and search trends on Twitter and Google using different ML algorithms and those are LR, polynomial regression, Recurrent Neural Network (RNN), and LSTM-based analysis. The study concludes that there is a substantial association between Google Trends and Tweet volume statistics and the price of BTC, but no meaningful correlation with tweet emotions. On the other hand, authors in [70] linked cryptocurrency price volatility and investor interest. The researchers showed that investor interest, as measured by the number of tweets, retweets, and likes correlates with increased cryptocurrency price volatility.

The influence of Google search on the number of tweets, and active addresses on the blockchain on BTC and ETH values over time is evaluated using a VAR GARCH-BEKK model; researchers have identified solid evidence that the number of active addresses is the most crucial factor influencing price swings in Bitcoin and Ethereum, among other cryptocurrencies. Based on spillover effects and GIRFs [71]. Google searches and tweets have some impact on Bitcoin and ETH values. However, the effects are smaller and less significant than active addresses in terms

of size and significance. Moreover, the study in [72] employs lexicon-based sentiment analysis algorithms coupled with Harvard Psychosocial and financial sector specialized dictionaries. Inspection of response patterns is used to develop sentiment-based trading techniques.

Topic modeling and public opinion mining concerning cryptocurrencies, blockchain networks, BTC, LTC, and ETH are used. Study [73] suggested a technique to identify the top areas where bitcoin is frequently utilized, and then investigate the users' worries and sentiment assessments in those regions. Top destinations include Australia, Denmark, the Netherlands, and the United States, among others. Almost 83.7 % of tweets in Sweden express positive emotions toward bitcoin investment, making Sweden the country with the most favorable cryptocurrency investment climate. Even though these so-called Altcoins have lower trade volumes, they can garner a lot of interest on social media.

There are more than 1500 cryptocurrencies in the markets most of the research focused on the most popular cryptocurrency markets and others, on the other hand, the others focused on the cryptocurrency in the down markets. The study in [74] presents a deep learning model for forecasting Dogecoin (DOGE) cryptocurrency values, the suggested model is based on DOGE cryptocurrency historical market price data as well as sociological tendencies. DOGE market data is gathered daily from Kaggle, and verified tweets with the hashtags "Doge-coin" and "Doge" are also collected for the same period. The suggested model produces a promising forecast of Dogecoin's future price, according to experimental data. The proposed method performs well in tests.

2.6.4 Cryptocurrency price prediction using price historical data

Cryptocurrency price prediction for investors is essential [75] to reduce the risk in the investment process, some data scientists and business analysts predict the cryptocurrency price depending on the price itself. Study [76] predicts the most popular cryptocurrency prices which are BTC, ETH, and LTC by depending on historical cryptocurrency prices by using multiple deep learning algorithms and those are LSTM, Bi-LSTM, and GRU. The researchers found that the GRU model is considered the best model overall targeted cryptocurrencies with a close value between the prediction and current price.

The study in [77] tested the predictability of 12 different cryptocurrencies depending on classification machine learning algorithms that is SVM, LR, ANN, and RF, the model feature was both historical prices and technical indicators, and the finding indicated that SVM presented the best accuracy than the other models, with taking into consideration that the average accuracy of predictive algorithms about 55–65%. Looking for other methods instead of classical ML models. The work in [78] Proposes a hybrid cryptocurrency prediction system based on both LSTM and GRU that focuses on only two cryptocurrencies; LTC and XRM. The findings show that the suggested technique correctly predicts prices with high accuracy, indicating that it may be used to anticipate prices for a variety of cryptocurrencies. The work in [79] Used a supervised machine learning (ML) strategy to solve the difficulties of short-term forecasting of cryptocurrency time series. They employ RF and SGBM as features and prior price data and technical indicators as features of moving average for the most popular and capitalized cryptocurrencies namely; BTC, ETH, and XRP, the accuracy of the short-term forecast daily close prices achieved by the SGBM and RF with Mean Absolute Percentage Error (MAPE) in the range of 0.92-2.61%.

A Study in [80] shows that, regardless of whether a re-sampling approach is used or not, all learning algorithms beat the Buy and Hold (BH) strategy in the vast majority of the 100 markets examined. Nonetheless, the unweighted average, out of all the learning algorithms, produces the greatest overall performance, with accuracies of up to 59.26% for time-resampled series. Most notably, the alternative resampling approaches evaluated are capable of producing considerably bigger returns and with reduced risk when compared to time-resampled data. The results of the SVM, ANN, NB, and RF were compared to the results of the LR using the t-test. The empirical studies show that the RF has the best-predicting performance in the continuous dataset, whereas the NB has the worst. In the discrete dataset, however, the ANN has the best performance and the NB has the worst. Furthermore, in all algorithms (models) calculated, the discrete dataset enhances overall predicting performance [81].

Some machine learning methods are used in [82] to discover the best method for forecasting BTC prices based on three other well-known currencies. In this study, a new approach is being created to forecast Bitcoin's value, which includes considering the values of various cryptocurrencies (Ethereum, Zcash, and Litecoin). The results showed that among these three cryptocurrencies, Zcash had the greatest performance in projecting BTC prices. [83] Focusing on developing forecasting models that can provide reliable cryptocurrency price forecasts. BTC, ETH, and LTC prices are forecasted using the ARIMA time series forecasting model, the Prophet Model, and the LSTM deep learning algorithm.

The proposed advanced forecasting model CNN [84] takes various cryptocurrency data as inputs and processes them separately to extract valuable information from each coin. The empirical analysis was conducted using three years of cryptocurrency data from the three most valuable

cryptocurrencies, namely BTC, ETH, and XRP. In comparison to standard fully-connected deep neural networks, the suggested model can efficiently utilize mixed cryptocurrency data, eliminate overfitting, and reduce computing costs.

By gathering real data about bitcoin from [85], which included 8 characteristics. Then examining the performance of 11 regression methods in terms of prediction, it was discovered that Lasso regression with a combination of generalized linear regression beat other regression methods by 9%. [86] investigates the influence of global currencies, such as the U.S dollar, and foreign exchange rates on BTC values, as well as whether BTC has the stability to dethrone global currencies and replace them as the only means of exchange; the findings achieved from predicting BTC prices using a machine learning-based model are sufficient to aid with price prediction. The NN achieved an accuracy of 94.89% under all conditions of technical trade indication. The study [87] uses a mix of machine learning and deep learning methods to build a price prediction model for cryptocurrencies based on transaction data from three separate marketplaces and the amount and substance of user comments and answers from online forums. It revealed that the trading price premium rate in different markets has an impact on the price to be anticipated and that integrating social media comment elements can increase prediction accuracy greatly. This post is helpful for investors who want to make more scientific judgments by encrypting their money.

The machine learning industry has progressed significantly and will continue to do so; this progress has led to a larger problem-solving approach, namely data prediction or trend analysis in any format. The structure of [88] is a time-series study of digital currency's daily high-low-close. For this investigation, the algorithms employed include LSTM and evaluating the model using MAE (Mean Absolute Error).

A Study in [89] describes a reinforcement machine learning-based model for active trading and applies it to five of the most popular cryptocurrencies in circulation. This model generates higher risk-adjusted returns and lowers downside risk. [90] Uses KNN, eXtreme Gradient Boosting, and RF classifiers to solve the challenge of detecting the direction of three cryptocurrency markets in this study. Price data and technical indicators are used among the data sources. These classifiers are used to create a trading strategy for certain markets. The test findings on unseen data demonstrate that this strategy has a lot of potential for assisting investors with an expert system to profit from the market. The largest gain over 66 days is 60 dollars on an investment of 1800 USD. And also go into the limits of these methods and how they could affect efficient market theories.

2.6.5 Cryptocurrency price prediction and machine learning and statistical methods

Traditional statistical approaches, while simple to use and interpret, rely on several statistical assumptions that may be impractical, leaving machine learning the finest technology in this sector, capable of forecasting pricing based on prior experience [91]. Because of the vast quantity of market data and trends, it is feasible to anticipate market behavior using sentiment analysis and machine learning techniques for the difficult challenge of predicting Cryptocurrency trends.

Bitcoin has been the most successful cryptocurrency of all. Using sentiment analysis and the XG-Boost algorithm, and suggested a two-step machine learning strategy to forecast bitcoin price. Using supervised machine learning to create a prediction model; [92] employs many machine learning such as linear regression and RNN to predict BTC prices, the experiment outcome shows that BTC prices can be predicted with high accuracy using machine learning algorithms. [93] Uses

bivariate time series to forecast the three major and popular cryptocurrency prices: BTC, XRP, and ETH using three nonlinear algorithms: Decision Tree Regressor (DTR) and KNN models. The performance matrix is based on MAE, MSE, RMSE, and R-Squared Error (RSE). The findings show that (DTR) surpasses the K-Nearest Neighbor (KNN).

The proposed model by [94] takes a time series of the prices of many major cryptocurrencies and learns their interdependencies by carefully recalibrating the weights of each sequence. To establish memory and extract spatial and temporal data; components from Convolutional Neural Networks CNN and LSTM are used. Furthermore, regularization methods such as kernel regularize and bias regularizes, as well as the Dropout method, are used to increase the suggested model's generalization capabilities. In this study, a series of tests are carried out. The results show that the WMCR model outperforms existing baseline models and reaches state-of-the-art performance.

The study in [95] addresses the issue by forecasting BTC, ETH, and XRP using three clever algorithms: KNN, SVM, and LSTM. This forecast was made based on the up and down value of cryptocurrencies using the RMSE to assess financial stability. The accuracy system using LSTM has proven that the price of cryptocurrencies will fit with actual data using LSTM, with an accuracy rate of roughly 80%.

The holy grail of investors is accurate cryptocurrency price prediction, but the task is extremely difficult and time-consuming because cryptocurrencies have high volatility. As a result, traditional statistical approaches fail to capture the complexities of bitcoin exchange rates; leading researchers

to rely on sophisticated machine-learning techniques. They also provide a framework for developing deep learning models to forecast cryptocurrency prices, which they use to forecast the price of ETH; resulting in both short and long-term forecasts with up to 84.2 % accuracy [96].

The study in [97] examines the impact of cryptocurrency trends on predictive models; these models will then be enhanced through hyperparameter tuning, and their performance will be compared across the two currencies. The datasets for each cryptocurrency are found to be different because of the varying permutation significance. However, this does not alter the overall predictability of the models, with the same models being the best performers for both short and medium-term forecasts. The findings of the studies in [98] show that while chaotic data can yield favorable results for short and medium-term forecasting; technical analysis-based forecasting is insufficient for long-term forecasting.

The study in [99] compares the results of two traditional time series methods, the ARMA, GARCH, and KNN approaches were used in predicting market liquidity in crypto and fiat currencies. Three cryptocurrencies were tested namely; BTC, ETH, and XRP. In addition, 16 main fiat currencies were used, and the finding reveals that the KNN technique is good for capturing market liquidity in a cryptocurrency in the near run than the ARMA and GARCH models.

2.6.6 Cryptocurrency price prediction using block-chain information

The link between inherent ETH Blockchain information and ETH pricing is investigated by [100] they also look at how Ethereum values are linked to Blockchain information about other publicly

accessible currencies on the market. The findings imply that macroeconomic factors, Ethereum-specific Blockchain data, and Blockchain data from other cryptocurrencies all play a role in Ethereum price prediction, with the ANN beating other models. A study in [101] revealed that the risk factors like beta and idiosyncratic volatility are connected to cross-sectional performance, implying that trading rules predominantly capture market risk premiums. In general, trading rules in cryptocurrency exchanges do not appear to give any significant benefits when compared to regular financial markets.

Research in [102] looks at realized volatility, skewness, and jump under the area of volatility risk factors. Also, the study investigates liquidity issues including bid-ask spreads, volume growth, and Roll's measure. Other crypto-unique variables are gathered, and they illustrate how machine learning models that can handle 30 different factors explain the majority of the cryptocurrency excess return. The LSTM network, the Stacked LSTM network, the Bidirectional LSTM network, and the GRU network are used to evaluate the price data collected by the Poloniex exchange [103], these deep learning models are compared and benchmarked based on a variety of measures. The research findings show that some of the aforementioned models may be used to predict the ETH closing price in real time with reasonable accuracy and profitability.

2.7 Data fusion

The complexity levels of economic markets can be represented by the time-series nature of economic markets, substantial cross-correlation between diverse entities, and collective market response under extreme influencing events. Various data sources, information that has been analyzed, and qualities that have been deduced and/or prediction algorithms are integrated into fusion approaches under appropriate conditions [104]. Fusion can help extract information from a

wide group of data sources for a hazy cryptocurrency market; it can extend the strength of a model to overcome the weakness of another model and it can leverage the search space to develop effective solutions. Fusion may also be used to investigate the reliance on certain market notions. As a result, this thesis gives a concentrated study on recent improvements in fusion-based cryptocurrency market prediction to address and illustrate the importance of fusion in cryptocurrency market prediction.

2.8 Conclusion

There are multiple methods used to predict the price of cryptocurrency depending on audience opinions through social media like Twitter or depending on the trends. On the other hand, some research depends on the historical data for cryptocurrency itself like the closing price. Under the right circumstances, different data sources, processed information, derived features, and/or prediction algorithms are combined in fusion techniques. Fusion techniques that depend on the fusion of data from multiple sources (channels) can assist a cloudy cryptocurrency market to acquire information from a variety of data sources, extend the strength of one model to overcome the weakness of another, and harness the search space to generate successful solutions.

Fusion may also be used to look into how reliant people are on particular market assumptions. As a result, this thesis focuses on current advancements in fusion-based cryptocurrency market prediction to address and demonstrate the significance of fusion in cryptocurrency market prediction. The goal is to figure out how one or more entities' restrictions are overcome by working together, as well as to emphasize the relevance of fusion in real-world cryptocurrency market

applications. To the best of our knowledge, our research is the first to focus on fusion in cryptocurrency market forecasting.

The next chapter includes the data description and the sources of data and the methodology that I used to collect the data which is the first step in the research journey.

Chapter Three

The data and data collection methods

3.1 Introduction

The chapter shows in detail the data description used in the thesis, the collection methods used, and the benefits of using the methods rather than the others. Because of using multiple data sources, multiple methods are required. The data collection method is a starting point for creating a prediction model. Multiple techniques can the model creator follow to reach the goals. The most popular methods use python libraries like snsrape and beautifulsoup. Using such libraries depends on how, when, and where they can be used.

In this thesis, I have used different data sources those are Twitter, news agencies, cryptocurrency market data, and blockchain information data. Twitter considers one social media giant. Over the last decade, the popularity of social media services has increased as it allows users to exchange material including text, video, and photographs. Hundreds of services now fall under the wide tent of social media and may be accessed via a web browser or a mobile application. Tweets allow its users to collect data for free, in this research we used scrapping techniques to achieve the goal of this work.

3.2 Scrapping tools

Scrapped tweets contain different types of languages the dominant language is English. The scrapped data presents more than (20) different languages. By applying insights gained from previous data, statistical analysis may help data mining by filtering out noisy data. Noisy data can be caused by hardware issues, programming errors, or meaningless input from voice or optical

character recognition (OCR) systems. Industry abbreviations, slang, and spelling errors can all make machine reading difficult.

Snsrape: is a scraper for Social Media Platforms (SNS); it scrapes data such as user profiles, posts, tweets, and Twitter data since 2006 [105], and other data related by using search keys and returns the results such as relevant postings. It's a great tool that works well without using Twitter's Application Programming Interface (API). On the other hand, it's a legal tool.

If the scraped data is shared in an unauthorized way; Twitter can disable any API access you have and potentially act against the Twitter account. Many other tools can be used to collect the data from Twitter, but practically proven it is the best and less limited than the others.

Tweepy: is one of the most widely used Python modules for connecting to Twitter. You sign up for a Twitter developer account, acquire your credentials, and begin scraping. It's a python library package [106] tool for simple automation, Twitter bot creation, and little school projects. Tweepy, on the other hand, has a scraping limit of 3200 tweets and a week-long limit. There is no way to obtain historical information. GetOldTweets3 is no longer useful because Twitter has removed the endpoint it uses.

Twint: is a sophisticated Python tool used to scrap data from Twitter [107], but Twitter appears to have a stricter device + IP-ban policy after a specific number of queries. It's a hit-or-miss situation. Many individuals have trouble installing it. While the solution becomes evident, the author of the library proposes temporarily utilizing a Dockerfile. Here's where you can learn more about TWINT. Octoparse is a commercial program that lets you acquire data from the web without having to code by extracting content from web pages automatically.

The main purpose for using this criterion and multiple search keywords is not to lose any single tweets that can give us a piece of important information, the outcome of this stage is to get millions of tweets. Snscape gives a lot of features shown in Figure 3.1:

Attributes Available Through snscape Tweet Object:

Attribute description left blank if purpose is unknown

- url: Permalink pointing to tweet location
- date: Date tweet was created
- content: Text content of tweet
- renderedContent: Appears to also be text content of tweet
- id: Id of tweet
- user: User object containing the following data: username, displayname, id, description, descriptionUrls, verified, created, followersCount, friendsCount, statusesCount, favouritesCount, listedCount, mediaCount, location, protected, linkUrl, profileImageUrl, profileBannerUrl
- outlinks
- tcooutlinks
- replyCount: Count of replies
- retweetCount: Count of retweets
- likeCount: Count of likes
- quoteCount: Count of users that quoted the tweet and replied
- conversationId: Appears to be the same as tweet id
- lang: Machine generated, assumed language of tweet
- source: Where tweet was posted from, ex: iPhone, Android, etc.
- media: Media object, containing previewUrl, fullUrl, and type
- retweetedTweet: If is a retweet, id of original tweet
- quotedTweet: If is a quoted tweet, id of original tweet
- mentionedUsers: User objects of any mentioned user in tweet

Figure 3.1. Attribute available through snscape.

3.3 Data collection from twitter

The first stage of this research is data collecting. On the internet, several data sources may be employed in the research process. Web scraping is the technique of extracting data or information from websites on the internet [108]. Figure 3.2 illustrates the scrapping methodology and data preparation processes used in this research for data collection from Twitter.

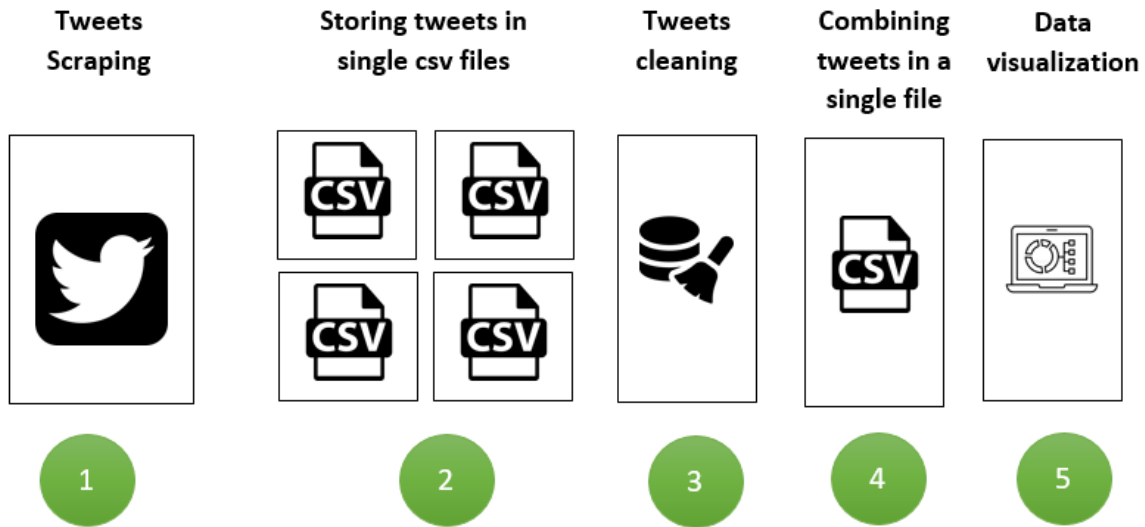


Figure 3.2. Data collection method.

The data collection process starts from scrapping tweets daily and is stored in single CSV files, in this stage multiple keywords were used to run snsrape tools to collect the data. Table 3.1 illustrates the keyword that was used. Many keywords are used to not lose any single records.

Table 3.1: The search keys are used in the Twitter scrapping process.

	Cryptocurrency	Search Key
1.	Bitcoin	BTC, #BTC, \$BTC, BTCUSD, #BTCUSD, \$BTCUSD, BITCOIN, #BITCOIN, \$BITCOIN.
2.	Ethereum	ETH, #ETH, \$ETH, ETHUSD, #ETHUSD, \$ETHUSD, ETHEREUM, # ETHEREUM, \$ ETHEREUM.
3.	Litecoin	#LTC, #LTC, \$LTC, LTCUSD, \$LTCUSD, #LTCUSD, LITECOIN, \$LITECOIN, #LITECOIN.
4.	Dogecoin	#DOGE, \$DOGE, DOGE, DOGEUSD, \$DOGEUSD, #DOGEUSD, DOGECOIN, #DOGECOIN, \$DOGECOIN.

Clearing the scraped data is represented in the third stage this includes, removing duplicated records, and noisy data that does not represent useful information. In this stage, we removed the

duplicate records, and noisy data, which represented meaningless data, a synonym for corrupt data. In general, data scientists and data miners consider this data type of data as corrupted data. Removing this type of data help in execution time while training and testing the proposed model. The fourth stage is combining CSV files into one file and processing the data, this makes it easy to interact with data.

Data visualization represented in stage 5 can give a general overview of the data like data shape and data distribution. Table 3.2 – 3.5 illustrates the number of scrapped tweets and the net number of tweets after processing duplicate records and noisy data.

Table 3.2 Number of scrapped tweets vs clean tweets for Ethereum.

Month	Total scrapped	English Tweets	Non-English Tweets	Total Clean Tweets
Jan	1,114,123.00	522,948.00	160,028.00	682,976.00
Feb	1,769,235.00	651,267.00	184,916.00	836,183.00
Mar	2,821,403.00	694,635.00	188,298.00	882,933.00
Apr	2,638,873.00	817,405.00	247,607.00	1,065,012.00
May	3,514,457.00	1,161,877.00	348,662.00	1,510,539.00
Jun	3,573,531.00	761,569.00	209,596.00	971,165.00
Total	15,431,622.00	4,609,701.00	1,339,107.00	5,948,808.00

Table 3.3 Number of scrapped tweets vs clean tweets for Bitcoin.

Month	Total scrapped	English Tweets	Non-English Tweets	Total Clean Tweets
Jan	32,677,305.00	2,109,783.00	664,795.00	2,774,578.00
Feb	25,396,924.00	2,254,368.00	753,364.00	3,007,732.00
Mar	42,829,832.00	2,037,406.00	701,665.00	2,739,071.00
Apr	55,777,286.00	1,896,353.00	789,445.00	2,685,798.00
May	39,590,145.00	2,934,004.00	1,128,031.00	4,062,035.00
Jun	15,632,093.00	2,607,473.00	1,003,805.00	3,611,278.00
Total	211,903,585.00	13,839,387.00	5,041,105.00	18,880,492.00

Table 3.4 Number of scrapped tweets vs clean tweets for Litecoin.

Month	Total scrapped	English Tweets	Non-English Tweets	Total Clean Tweets
Jan	314,143	142,047	34,031	176,078
Feb	276,774	125,163	39,061	164,224
Mar	250,777	105,165	34,772	139,937
Apr	308,899	130,540	48,389	178,929
May	335,192	148,235	49,446	197,681
Jun	175,062	78,125	25,718	103,843
	1,660,847	729,275	231,417	960,692

Table 3.5 Number of scrapped tweets vs clean tweets for DOGE.

Month	Total scrapped	English Tweets	Non-English Tweets	Total Clean Tweets
Jan	1,839,282	1,043,610	150,831	1,194,441
Feb	2,471,510	1,168,494	217,881	1,386,375
Mar	930,703	387,129	108,459	495,588
Apr	3,475,563	1,636,634	375,154	2,011,788
May	5,289,890	2,668,936	657,858	3,326,794
Jun	2,087,103	851,771	202,826	1,054,597
	16,094,051	7,756,574	1,713,009	9,469,583

Figure 3.3 illustrates the distribution of tweets regarding the initial data preprocessing and the net number of tweets per targeted cryptocurrency. BTC represents the largest number of tweets with the initial raw data of more than 220 million tweets and approximately 18.8 million clean ones. On the other hand, Litecoin represents the smallest number of tweets with the amount of 960,692.00 clean tweets from 1.6 million raw data. The total number of clean tweets for all five targeted cryptocurrencies is more than 220 million tweets representing raw data, approximately 41 million of the record that represents the clean tweets.

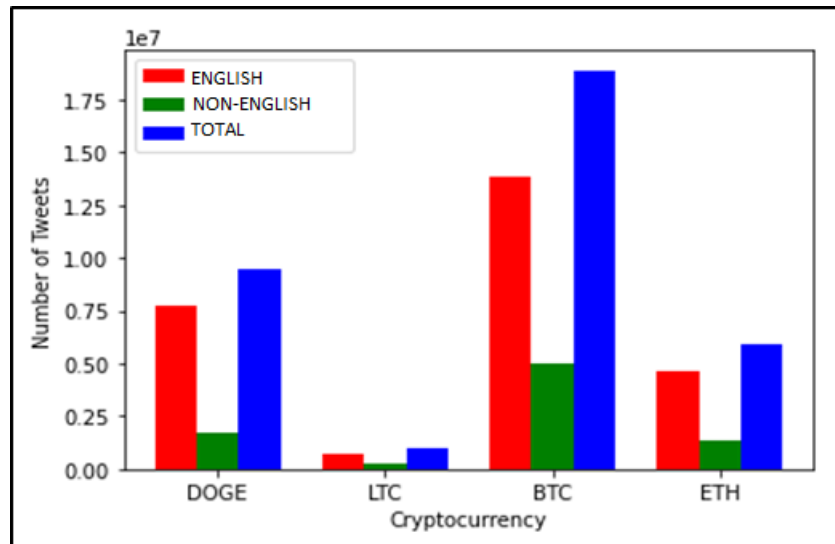


Figure 3.3. Scrapped Tweets distribution per cryptocurrency

In general, the data has been collected daily using snsrape tools this leads to collect hundreds of CSV files that need to be combined into a single CSV file this is done in stage 3.

Multiple Tweets language may include important information that can help in the targeted model and the goal of this work, interacting with different languages is very difficult in direction of cost, time, and complexity. To avoid these difficulties, we translate all non-English tweets into English.

Table 3.6 Tweets distribution per language for the top 10 languages.

Cryptocurrency	Language Code	Language Description	Number of tweets
Bitcoin	EN	English	15,644,790
	TR	Turkish	1,249,650
	ES	Spanish	1,026,677
	JA	Japanese	716,313
	FR	French	272,572
	PT	Portuguese	254,211
	HI	Hindi	228,458
	DE	German	163,538
	NL	Dutch	101,417
	AR	Arabic	78,924

Ethereum	EN	English	4,981,749
	TR	Turkish	255,028
	JA	Japanese	181,327
	ES	Spanish	150,109
	FR	French	112,803
	HI	Hindi	60,622
	CY	Welsh	43,372
	PT	Portuguese	40,593
	DE	German	34,359
	TL	Tagalog	28,665
Litecoin	EN	English	510,740
	TR	Turkish	29,654
	ES	Spanish	10,924
	RO	Romanian	9,921
	HI	Hindi	9,672
	JA	Japanese	9,298
	DE	German	8,085
	FR	French	5,807
	HT	Haitian	4,570
	PT	Portuguese	4,361
Dogecoin	EN	English	7,761,120
	TR	Turkish	286,163
	ES	Spanish	132,050
	HI	Hindi	122,344
	FR	French	57,466
	JA	Japanese	53,827
	PT	Portuguese	42,268
	TL	Tagalog	29,946
	DE	German	27,805
	NL	Dutch	20,110

In stage 4, data preprocessing has done to prepare the data for the visualization stage, this stage included removing duplication in tweets, and removing unnecessary data that does not give useful information, regarding BTC there are about 1,019,026 tweets classified as an undefined language which means not meaning data can be extracted from these tweets that removed through preprocessing step.

Figure 3.4-A illustrates the distribution of the tweets per language it shows that English tweets have the largest number of tweets and can be considered the dominant language. The English language is considered the dominant language with 15,644,790 tweets a percentage of 79.50%. For Ethereum cryptocurrency Figure 3.6 illustrate tweet distribution per language. Again, the English language is considered the dominant language as it represents 81.75 % of the total tweets. The undefined languages (noisy data or meaningful tweets) approximately represent 6% of tweets. On the other hand, For Litecoin cryptocurrency figure 3.4-C illustrates tweet distribution per language, also the English language is considered the dominant language in another hand it represents 79.10 % of the total tweets. The undefined languages (noisy data or meaningful tweets) approximately represent 6% of tweets. For the Dogecoin cryptocurrency, Figure 3.4-D illustrate tweet distribution per language. The English language is considered the dominant language as it represents 66.36% of the total tweets. The undefined languages (noisy data or meaningful tweets) approximately represent 6.4% of tweets. On the other hand, figure 3.5, 3.6, 3.7, and 3.8 illustrate a screen shoot from undefined tweets for BTC, ETH, LTC, and DOGE respectively.

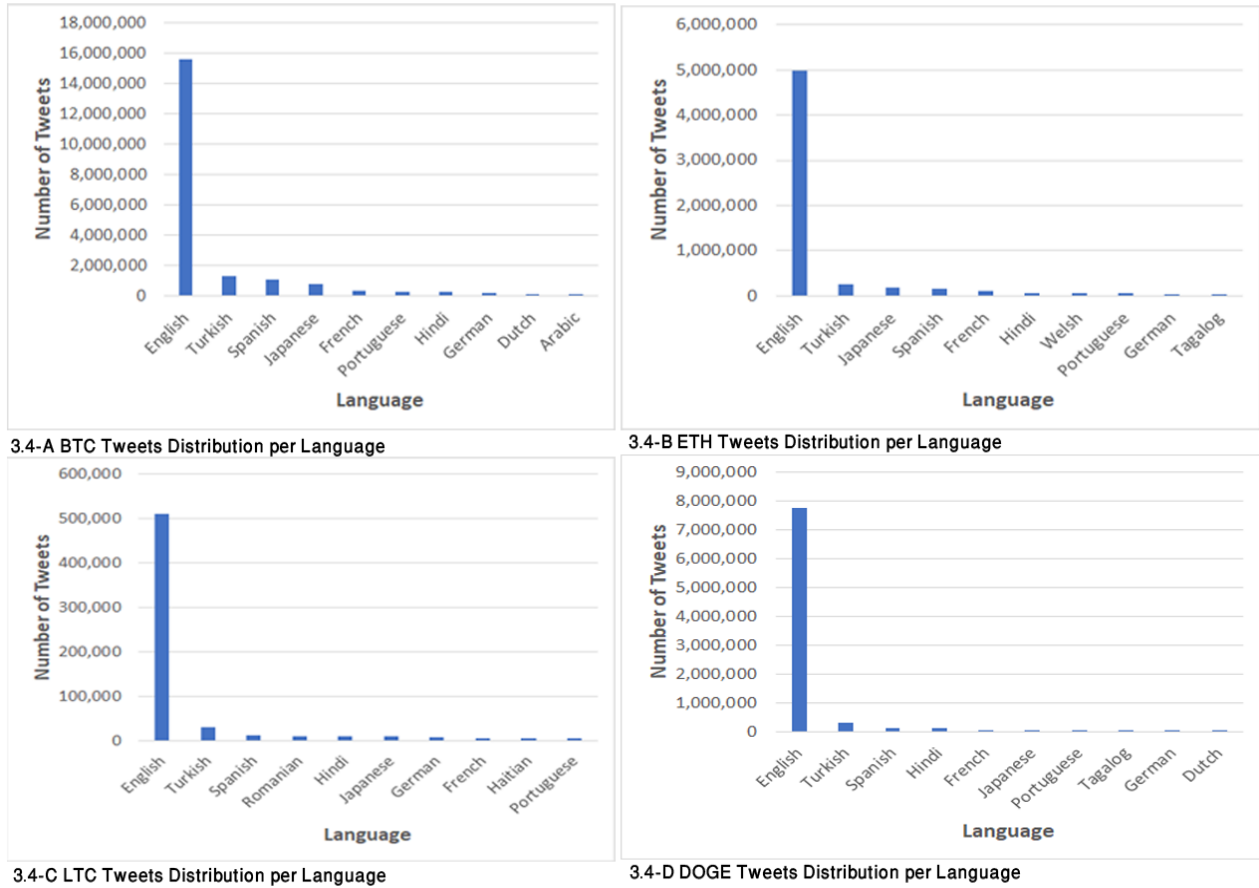


Figure 3.4. BTC, ETH, LTC, and DOGE Tweets Distribution per Language

Tweet Id	Text	Username	cashtags	user_location	hashtags	user_description	source	language
377771847974969345	#CRYPTO #BITKUP #BTC https://t.co/n561KUI8JB	funtech_stock	NaN	NaN	['CRYPTO', 'BITKUP', 'BTC']	อัพเดทข่าวสารและแลกเปลี่ยนความเห็นเกี่ยวกับกร...	<a href="http://twitter.com/download/iphone" r...	und
377771312152723456	#business #businessgoals #businessgrowth #busi...	bmurphypointman	NaN	Bay Area, CA	['business', 'businessgoals', 'businessgrowth']	#affiliate #affiliatemarketing #bitcoin #blogg...	<a href="https://contentstudio.io" rel="nofollow..."	und
377770535594024960	#greenscreen #crypto #macguyvertech #fyp #prog...	macguyvermedia	NaN	Philadelphia, PA	['greenscreen', 'crypto', 'macguyvertech', 'fy...']	CEO of @MacguyverTech01. White hat for transpa...	<a href="https://mobile.twitter.com" rel="nofo...	und
377770437623607301	@sgrmidolov #btc	Skaterbating	NaN	Ottawa, Ontario	['btc']	👉👉	<a href="http://twitter.com/download/iphone" r...	und
377770425581776897	https://t.co/pgHozlRi8F #cryptotab\n\n#xrp #BT...	NonnyMyles	NaN	NaN	['cryptotab', 'xrp', 'BTC', 'cryptocurrency', ...]	Giving up is illegal....	<a href="http://twitter.com/download/android" ...	und

Figure 3.5. BTC screenshot for undefined language.

Datetime	Tweet Id	Text	Username	cashtags	user_location	hashtags	user_description	source	language
2021-04-01 17:12+00:00	1377761996951580673	@cryptocom @wanderingstan #FET #RAVENCOIN #LTC...	MarkOfTheBEAS14	NaN	#BEASTISH #BEASTNODES	'[FET]', 'RAVENCOIN', 'LTC', 'ETHEREUM', 'MANA' ...	I walk by faith! Esoteric knowledge! My heart ...	<a href="http://twitter.com/download/android" ...	und
2021-04-01 14:45+00:00	1377761381169164292	👉🏻👉🏻\n.\n.\n.\n.\n.\n.\n.\n.#c...	5_0Trading	NaN	Lyon, France	['crypto', 'btc', 'eth', 'xrp', 'dot', 'bnb', ...	Mainly swing trading ALTS 📈 Yes a woman is o...		und
2021-04-01 09:30+00:00	1377760058105335815	👉🏻👉🏻\n.\n.\n.\n.\n.\n.\n.\n.#c...	5_0Trading	NaN	Lyon, France	['crypto', 'btc', 'eth', 'xrp', 'dot', 'bnb', ...	Mainly swing trading ALTS 📈 Yes a woman is o...		und
2021-04-01 02:28+00:00	1377758288520081408	👉🏻👉🏻\n.\n.\n.\n.\n.\n.\n.\n.#c...	5_0Trading	NaN	Lyon, France	['crypto', 'btc', 'eth', 'xrp', 'dot', 'bnb', ...	Mainly swing trading ALTS 📈 Yes a woman is o...		und
2021-04-01 07:12+00:00	1377741865689178115	#ltc \$lrc 👉 https://t.co/ /Sp4cLgbfYH	Bitcoinmerakisi	['ltc']	NaN	['ltc']	Sosyal medya hesaplarına indirimi borsa re...		und

Tweet Id	Text	Username	cashtags	user_location	hashtags	user_description	source	language
71265357826	#DogeCoinRise #dogeday #n#dogeN https://t....	MakeFamousDoge	NaN	Los Angeles, CA	['DogeCoinRise', 'dogeday', 'doge']	In Christ alone I place my trust, and find my ...	<a href="http://twitter.com /download/android" ...	und
18553231360	#doge 🐶🐶🐶 #dogearmy #DogeCoinToTheMoon #dogec...	ColaciJarred	NaN	New York	['doge', 'dogearmy', 'DogeCoinToTheMoon', 'doge...]	LI NY living crypto lovin HODL #btc #eth #ltc ...	<a href="http://twitter.com /download/iphone" r...	und
06859388928	#dogeday #doge https://t.co /dynQfCSap5	bay_aura	NaN	Bay Area	['dogeday', 'doge']		<a href="http://twitter.com /download/iphone" r...	und
089070223360	#doge	avegaofficial	NaN	NaN	['doge']	Creator and Host of The Independent Hustle Pod...	<a href="http://twitter.com /download/iphone" r...	und
40884834307	#doge #DogeCoinToTheMoon https://t.co /pJthUgzrnwG	SendmeDoge	NaN	NaN	['doge', 'DogeCoinToTheMoon']	Tired of this Doge- eat-dog world. Going off gri...	<a href="https://mobile.twitter.com" rel="nofo...	und

Figure 3.8: DOGE screenshot for undefined language.

3.4 Data collection from news agencies

It is self-evident that media news, which has surpassed newspapers is the most important source of daily living information. It is a decision-making tool for traders, investors, and markets. Because of their low transaction fees, and peer-to-peer architecture cryptocurrencies take social media attention. This increases the volume, volatility, and price, as well as the regular mention of cryptocurrencies in the mainstream media. One possible social aspect that may have influenced the pricing is this news item [109].

The arrival of news influences the view or emotion toward cryptocurrencies at any given time. Traders and investors now have regular access to updated news, owing to the wonders of the internet, and the news constantly molds their thoughts and influences their decision to invest in cryptocurrency.

For the news channel, the data was scrapped from [110]. This is a service created by cryptocurrency lovers called Crypto News. The owner intends to develop the most beneficial digital platform for anyone who wants to learn about and profit from the cryptocurrency market. It includes a website and app for Crypto News, as well as a portfolio manager, crypto market tracking service, and news and article aggregator. A news aggregator and editorial office-provided news and articles that can be found on the website. The goal of Crypto News is to educate the public on cryptography and how to effectively use it in everyday life, freedom, honesty, and decentralization of information are ideals of Crypto News. To scrap news, we employ a Python module that is used to extract data from HTML and XML files for web scraping purposes. From the page's source code, it generates a parse tree that may be used to extract data in a hierarchically and legibly way. Table 3.7 illustrates the number of scraped news.

Table 3.7: Number of scrapped news

Cryptocurrency	Number of records
BTC	11,437
ETH	2,644
LTC	324
DOGE	1,012

Figure 3.9 illustrates the number of scrapped news for every targeted cryptocurrency.

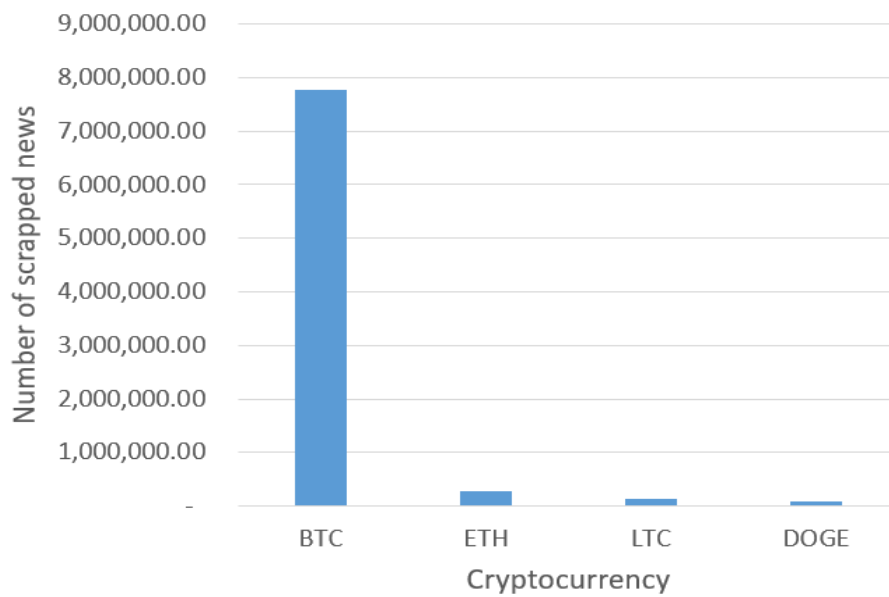


Figure 3.9. The scrapped news distribution per Cryptocurrency.

3.5 Data collection for prices and trading information

For the data that is related to cryptocurrency itself, the data has been collected from [111] such as cryptocurrency price and trading volume. Yahoo finance provides hours of live, daily market coverage are provided by Yahoo Finance, with professional analysis and up-to-the-minute market information. Investors, financial experts, and corporate executives who take their money seriously go there because they are insight-driven.

The data collected for the gold price is obtained from [112]. This website offers stock market data, business news, analysis, and financial information. It is a division of Dow Jones & Company, a News Corp-owned company, along with The Wall Street Journal and Barron's. Both websites provide free access to the data and allow it to download there as CSV files. For the data that are related to cryptocurrency production, this type of data has been collected from [113], the main goal of this is to present cryptocurrency data points in the most straightforward

3.6 Conclusion

The data collection method is the starting point to create prediction models. Multiple methods can be used but researchers should focus on more effective ones. Not all collected data is useful that should take a lot of attention in the data preparation stage. Python makes data collection easy by providing multiple and robust free libraries. The collected data used to achieve the purpose of the research has been scrapped daily.

In the next chapter, I will explore the collected data by using data visualization tools for all channels that the data was collected from to take an overview of the data and also study the relations between the different variables.

Chapter Four

Exploratory and explanatory data analysis

4.1 Introduction

In this chapter I will use the data visualization technique using different types of colored charts, to give an overview of the data, and the relationships between the variables. The chapter also includes the visualization of the different channels that we rely on which are: Twitter channels, news agencies channels, cryptocurrency historical data, and blockchain information. Data visualization tools came to simplify the process by creating and spreading meaningful color charts that tell us a brief of the story. In this section, we will explore the collected data mentioned in chapter 3 that was used to build our proposed model.

4.2 Cryptocurrency financial information

Financial information includes the closing price which represents the latest price that the cryptocurrency reaches at the end of the trading day. High and low price represents the highest cryptocurrency price and the lowest price on a specific trading date; however, the volume represents the number of sales and buy on a specific trading date. The used data that represented the previous information has been collected from [114].

4.2.1 Closing price

The closing price for every targeted cryptocurrency namely: Bitcoin, Ethereum, Litecoin, and Dogecoin in the interval between 01 January 2021 and 30 June 2021 is illustrated in Figure 4.1. For all cryptocurrencies. The minimum values at the beginning of the targeted interval increased at the middle of the same interval, and down at the end.

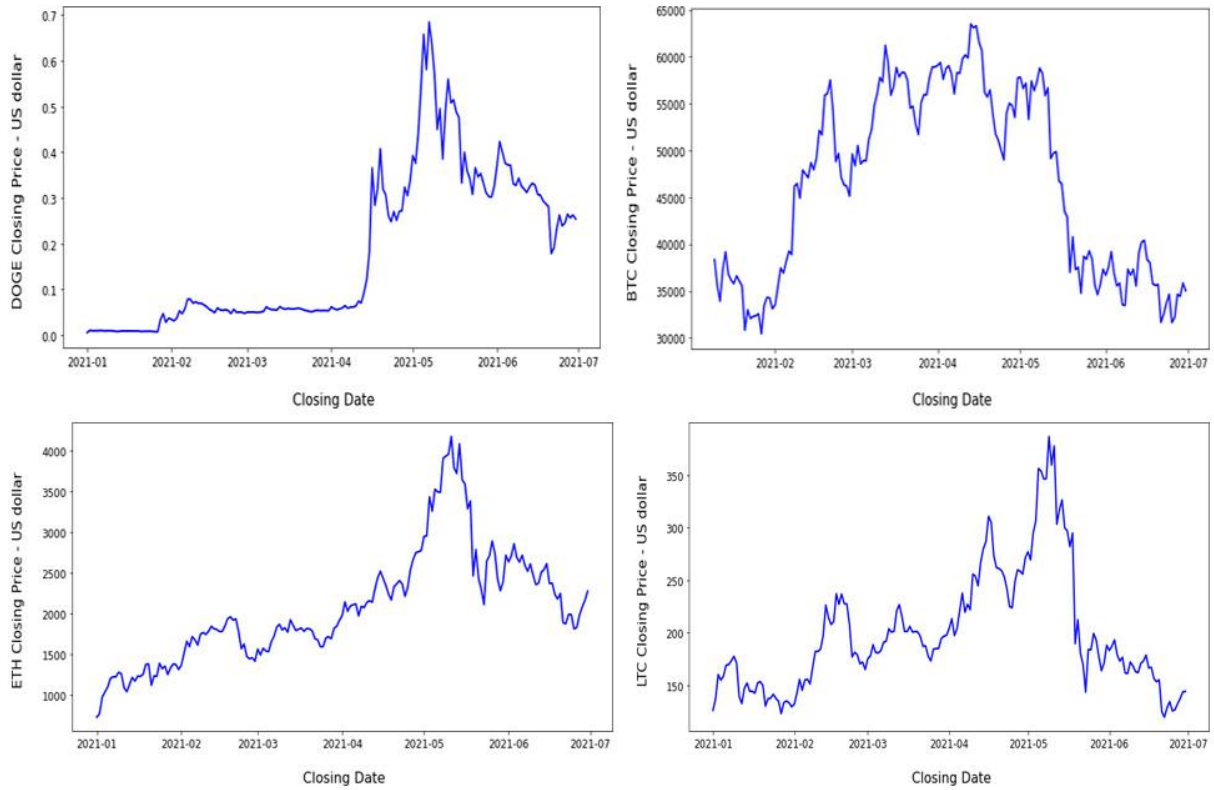


Figure 4.1. Closing Price for BTC, ETH, LTC, and DOGE.

The mean value is 45884.31 USD and the standard deviation is 10053.27 USD which is considered a high standard deviation, on the other hand, the maximum amount of Bitcoin is 63523.75 USD on 13 April 2021, and the minimum amount is 28994.00 USD at the beginning of the targeted interval of 01 January 2021. For Ethereum, the mean value is 2063.78 USD and the standard deviation is 702.94 USD which is considered a middle standard deviation, on the other hand, the maximum amount of Ethereum is 4174.63 USD on 15 May 2021, and the minimum amount is 730.40 USD at the beginning of the targeted interval of 01 January 2021. For Litecoin, the mean value is 199.20 USD and the standard deviation is 57.46 USD which is considered a middle standard deviation, on the other hand, the maximum amount of Litecoin is 387.86 USD on 9 May

2021, and the minimum amount is 119.67 USD on 22 Jun 2021. For Dogecoin, the mean value is 0.175 USD and the standard deviation is 0.171 USD which is considered a low standard deviation, on the other hand, the maximum amount of Dogecoin is 0.687 USD on 01 January 2021, and the minimum amount is 119.67 USD on 07 May 2021.

Figure 4.2 illustrates the correlation coefficient of targeted cryptocurrencies.



Figure 4.2. Correlation matrix for the BTC, ETH, LTC, and DOGE.

It shows a moderate positive correlation between BTC and DOGE and a high positive correlation between BTC and ETH and also LTC. On the other hand, the correlation is very strong between DOGE and ETH and low positive between LTC and ETH, and negligible between DOGE and LTC.

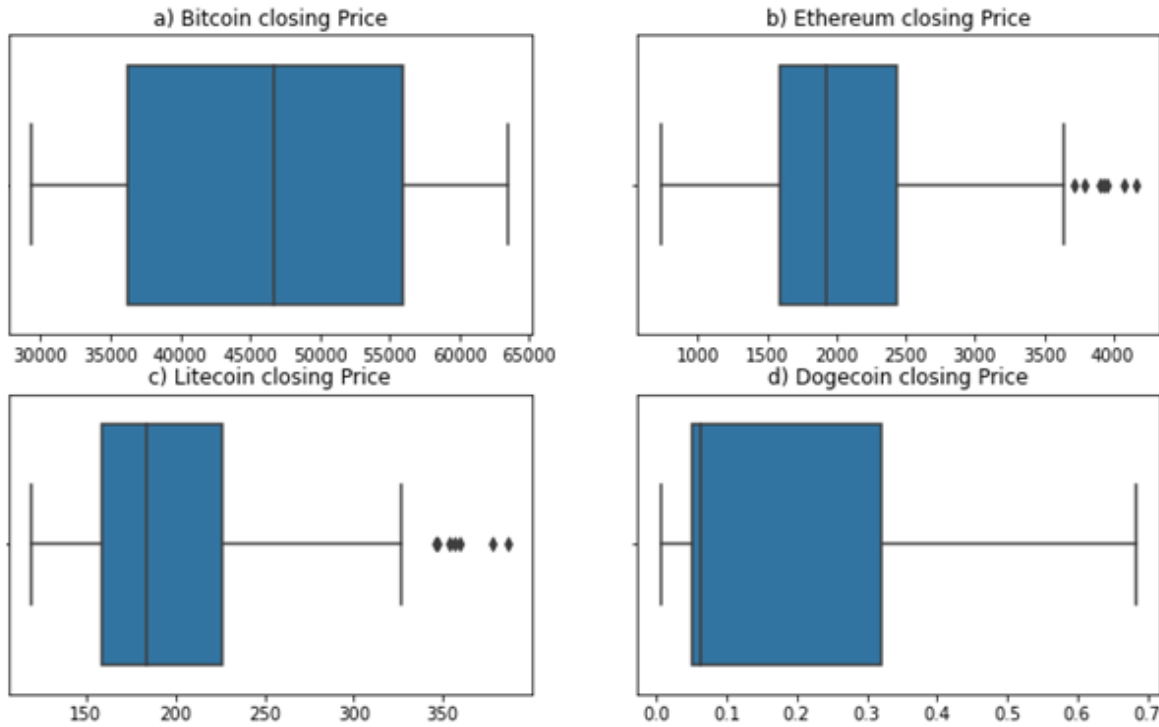


Figure 4.3. Boxplot for closing Price for BTC, ETH, LTC, and DOGE.

Figure 4.3. (a) shows that the closing price for BTC is between 36178.14 USD and 55973.51 USD. Any value over 64,000 USD is considered an outlier in which no outliers are detected for BTC within the targeted interval. Figure 4.4. (b) illustrates the closing price of ETH, it shows that the closing price is in the range of 1595 USD and 2431 USD; any value over 2700 USD is considered an outlier. Figure 4.4 (c) shows the boxplot for LTC, it shows that the closing price is between 159.59 USD and 226.22 USD, and any value over 325 USD is considered an outlier. Figure 4.4 (D) shows the boxplot of the closing price of DOGE, it shows that the closing price is between 0.05 USD and 0.26 USD, and any value over 0.32 USD is considered as an outlier, and for this plot, there is no outlier detected.

4.2.2 Open, high, and closed prices

Open price refers to the first price that occurs at the beginning of a specific date, on the other hand, the high price represents the highest price that occurs on the same date, which is shown in Figures 4.4 - 4.7.



Figure 4.4. Open, high, and close prices for BTC.

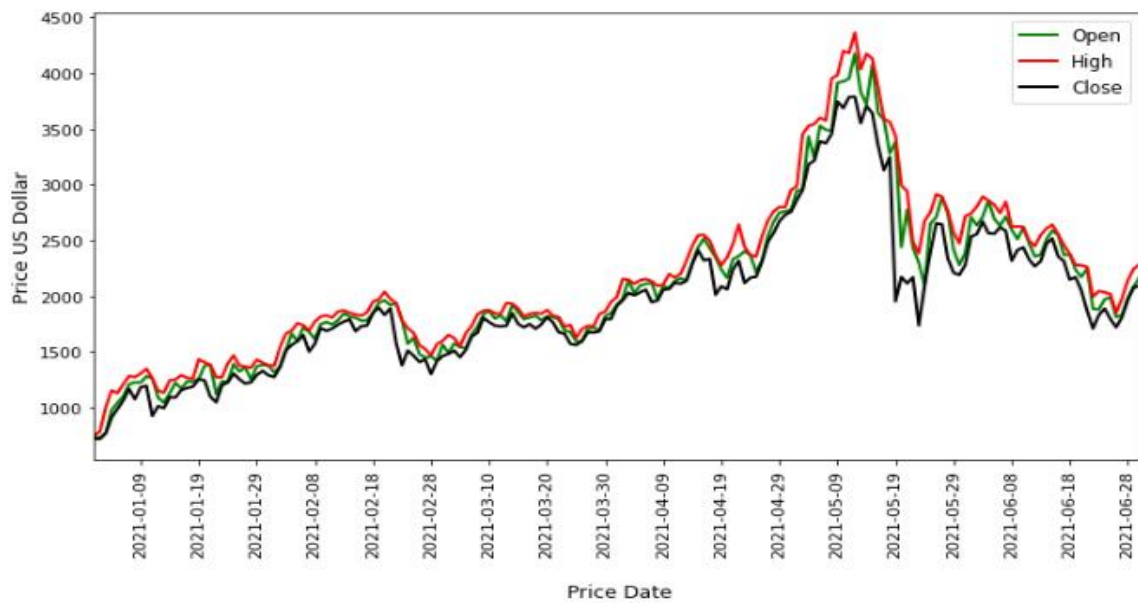


Figure 4.5. Open, high, and close prices for ETH.

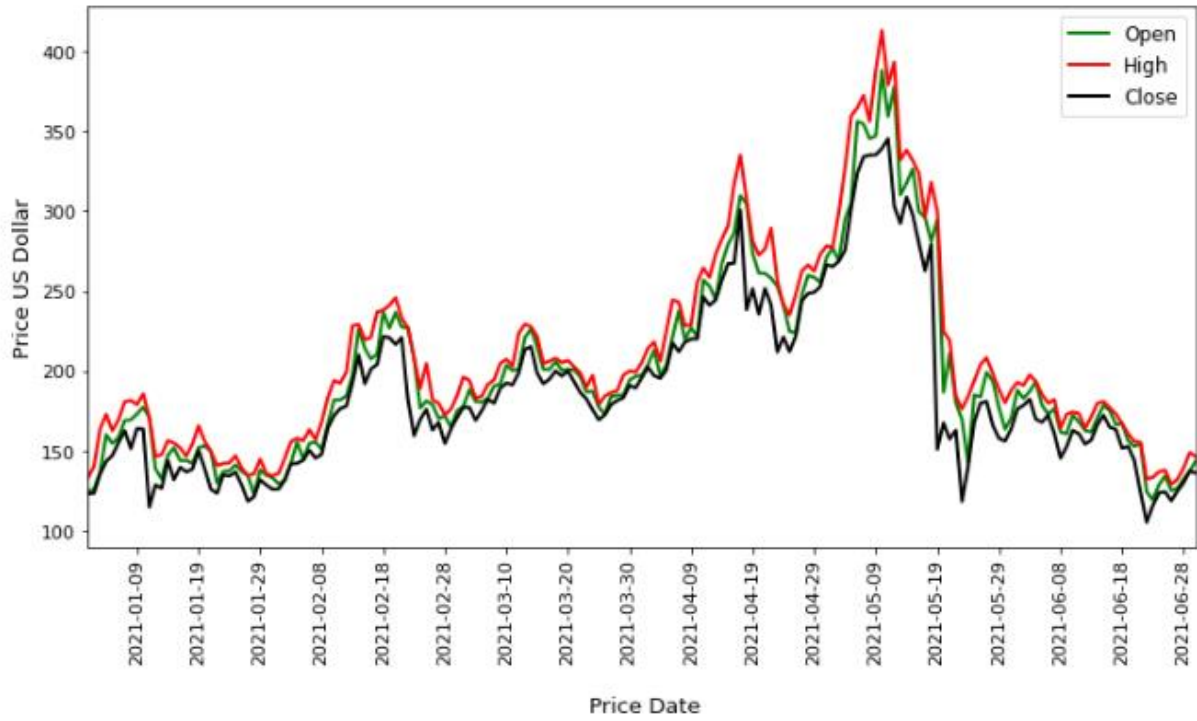


Figure 4.6. Open, high, and close prices for LTC.

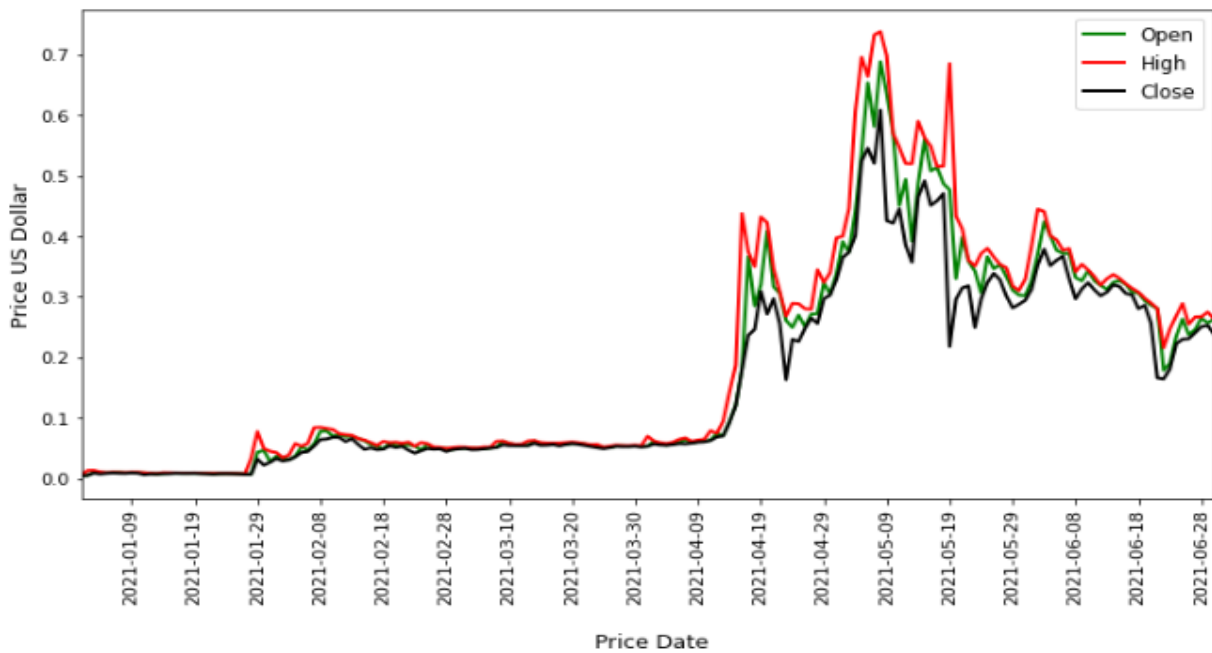


Figure 4.7. Open, high, and close prices for DOGE.

4.2.3 Volume

The volume is representing the total of actual trades taking place, how many Bitcoins are being bought and sold on particular exchanges is shown by the Bitcoin trade volume. More on-chain activity, such as when customers deposit and withdraw money, is probably caused by high trade volumes. Additionally, it is a reliable sign of public interest in the cryptocurrency market. Trading volume is a key characteristic that traders use to evaluate a coin's possible trajectory in cryptocurrency. The number of times a coin changes hands in a certain time frame is referred to as crypto trade volume. Investors look at crypto volume based on trading on a single crypto exchange or all crypto exchanges combined. Figures 4.8 – 4.11 illustrates the trading volume for targeted cryptocurrencies BTC, ETH, LTC, and DOGE respectively.

Results in Figure 4.8 shows that the mean volume trading of BTC within the targeted interval is more than 61 billion USD and the maximum is 360 billion USD on 26 February 2021 which represents the beginning of the price increase for BTC price, and the minimum about 28 billion on 6 Jun 2021.

Figure 4.9 shows the volume trading for ETH, which presents about 34 billion USD mean value, and the maximum value is 84 billion USD on 19 May 2021. The amount of 13 billion represents a minimum volume of trading at the beginning of the targeted interval. For the LTC, Figure 4.9 shows a mean volume of trading of about 6.4 million USD, a minimum volume of trading is about 1.7 million USD, and a maximum volume of trading is 17.9 billion USD on 11 Jan 2021 which represents the first price down. For DOGE, Figure 4.10 represents trading volume, the mean volume is about 6.1 billion USD, the maximum volume is 69 billion on 16 April 2021, and the minimum volume is 154 million USD on 23 Jan 2021. The maximum volume of BTC makes it the largest volume.

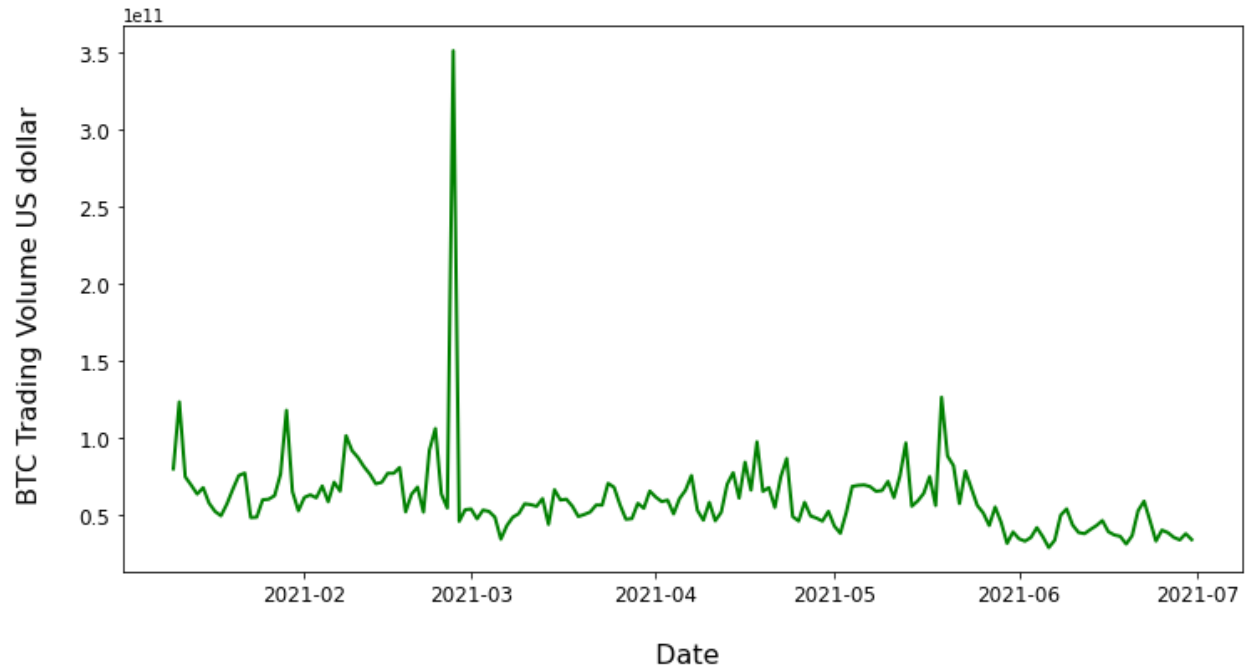


Figure 4.8. BTC Trading Volume.

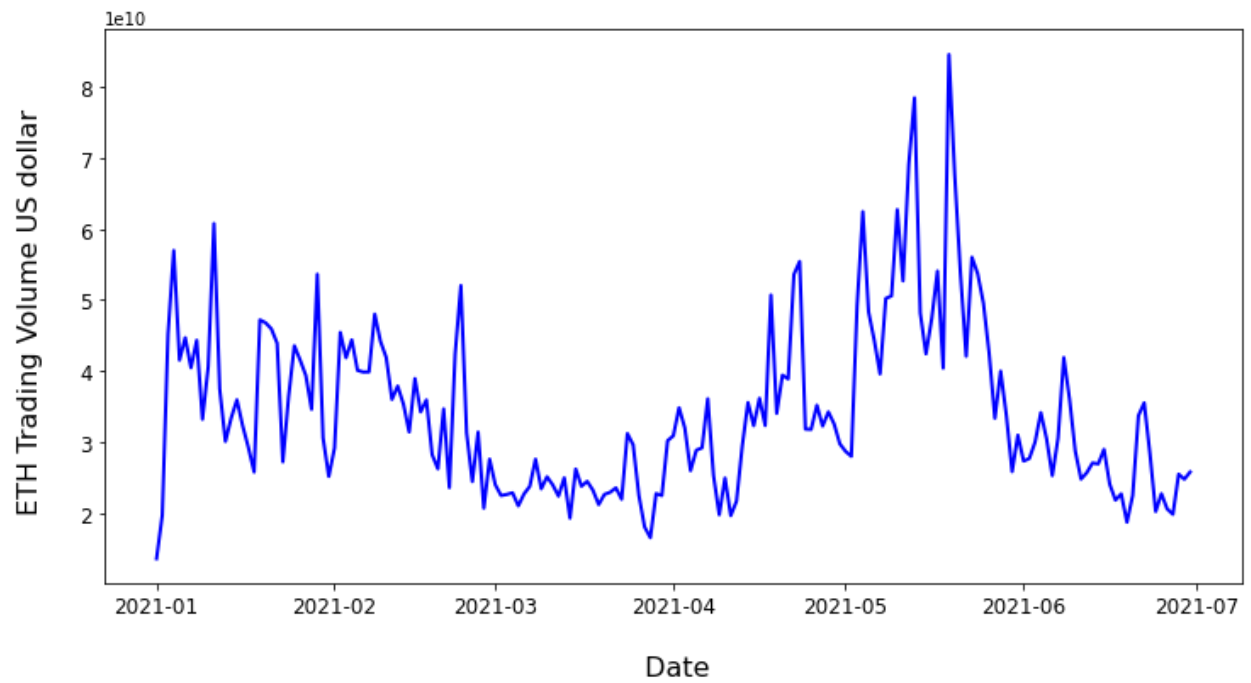


Figure 4.9. ETH Trading Volume.

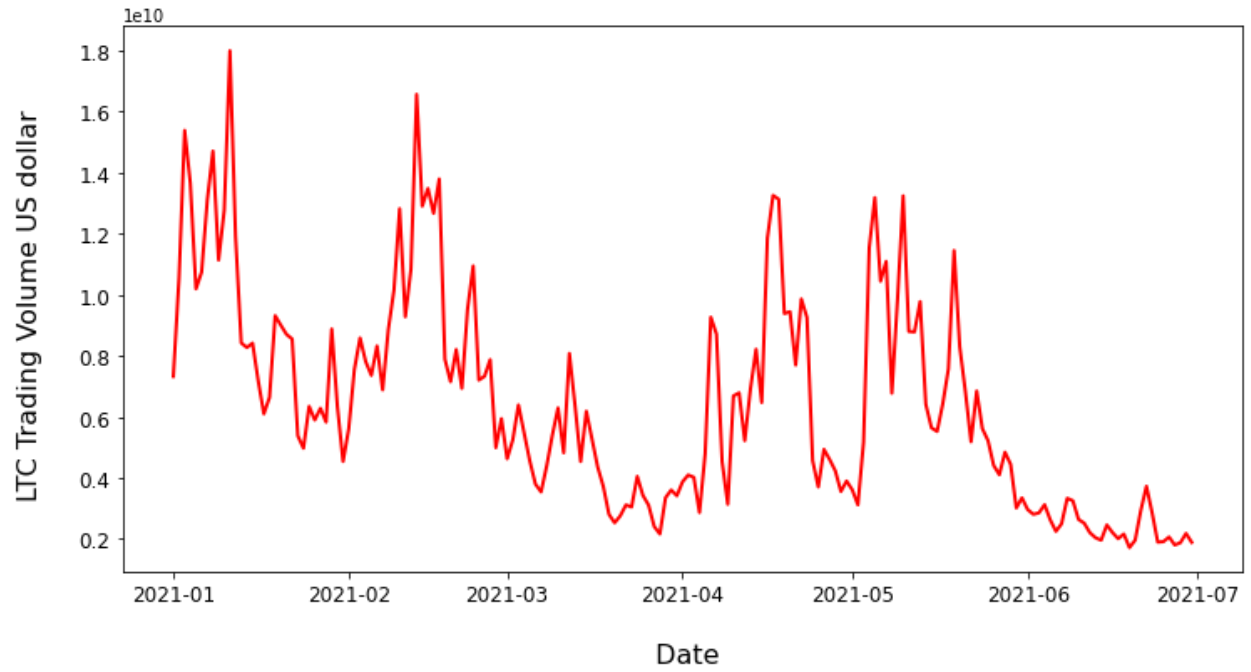


Figure 4.10. LTC Trading Volume.

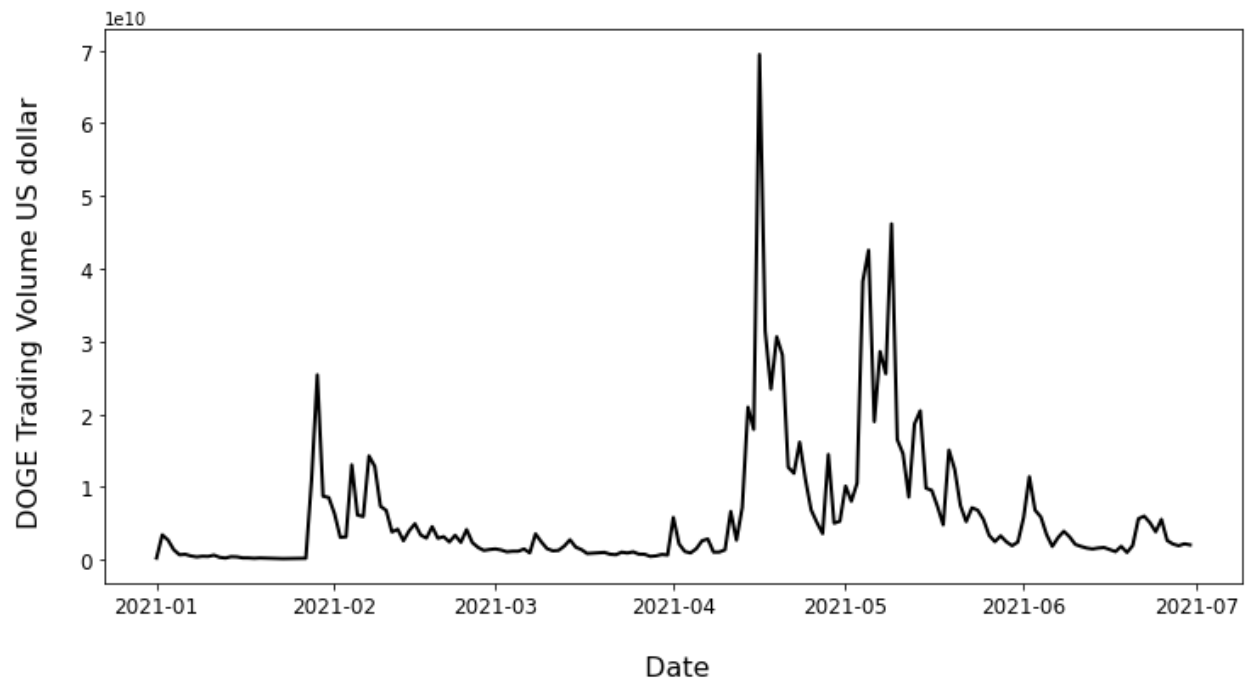


Figure 4.11 DOGE Trading Volume.

4.3 Blockchain and market information

In this section, I will show the other collected information related to the blockchain and market information which are: Average Hash Rate, Average Block Size, Mining Difficulty, Transaction fees, and Market capitalization.

4.3.1 Average hash rate

The entire network hash rate is estimated by comparing the average time between mined blocks to the network difficulty at a given point in time. The number of calculations that can be performed per second and measured by KH/s, MH/s, GH/s, TH/s, PH/s, and EH/s is hashed per second in the thousands, millions, billions, trillions, quadrillions, and quintillions. Cryptocurrency miners place a premium on the hash rate. They want their machines to solve puzzles as rapidly as possible since they are continuously competing with other miners to be the first. This has an impact on the profitability of mines.

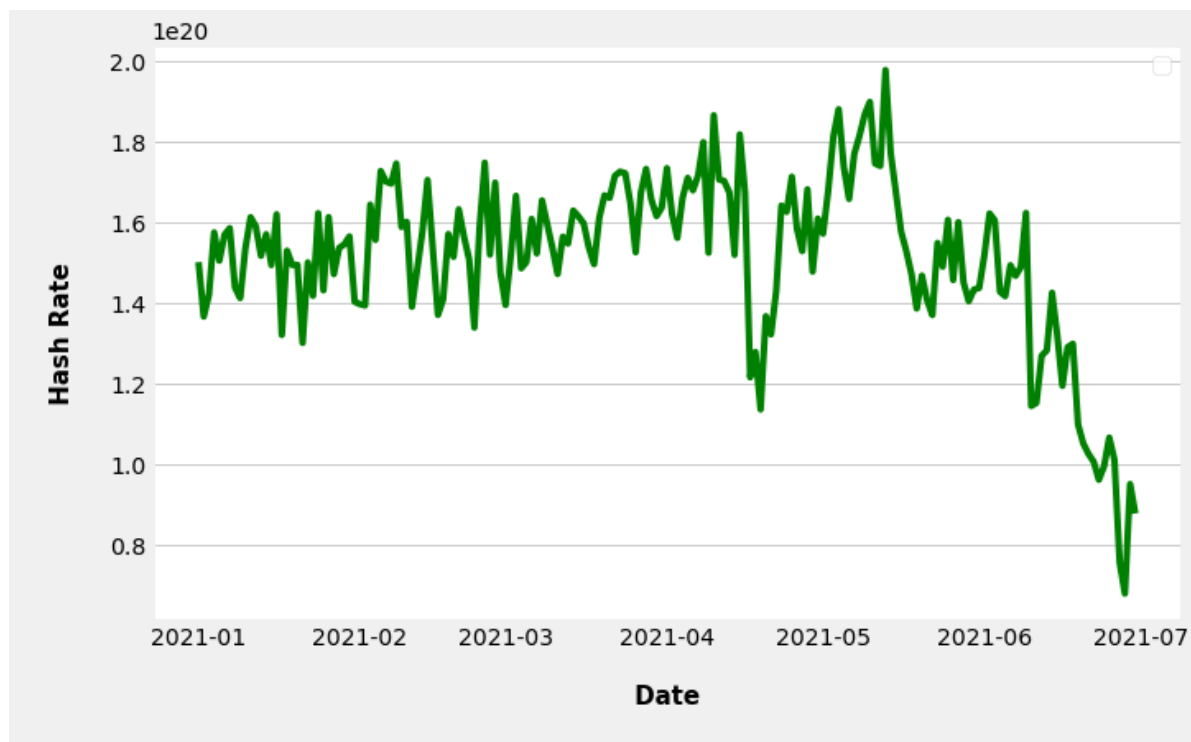


Figure 4.12. BTC average hash rate.



Figure 4.13. Correlation matrix between BTC closing price and average hash rate.

Figure 4.12 illustrates the BTC average hash rate. It shows that the frequency is normal at most of the targeted intervals, and it suffers from suddenly increased and decreased since the first quarter of April 2021. The correlation coefficient between the closing price and the average hash rate in Figure 4.13 indicates a weak positive correlation (0.23).

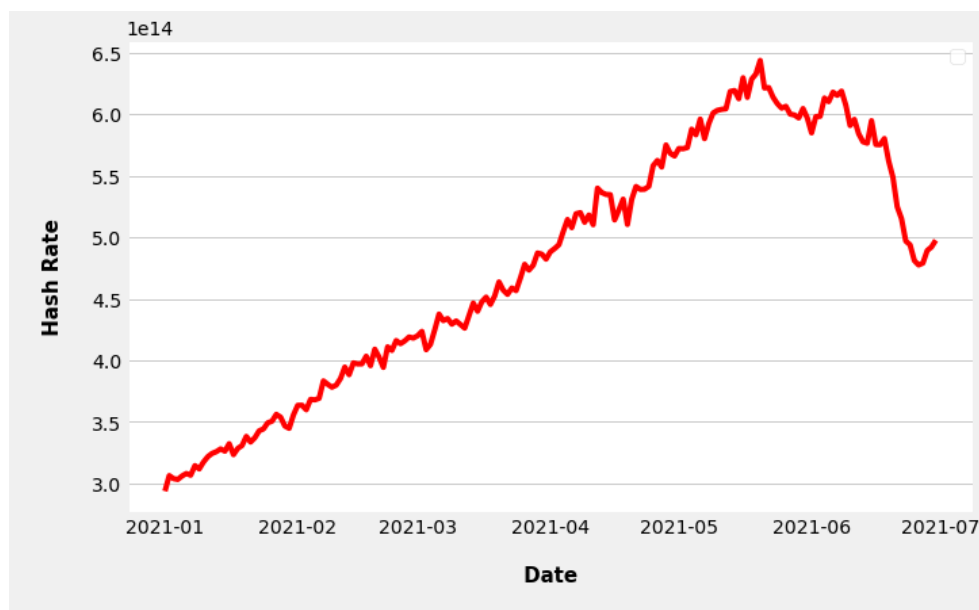


Figure 4.14. ETH average hash rate.

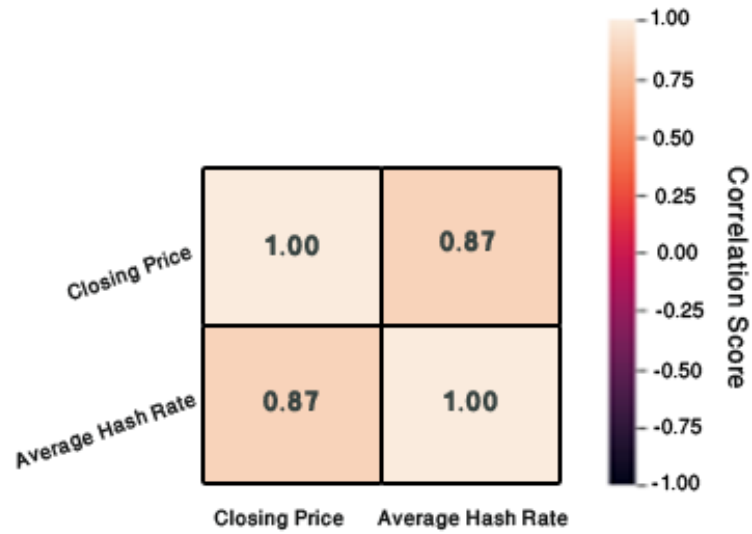


Figure 4.15. Correlation matrix between ETH closing price and average hash rate.

Figure 4.14 illustrates that the average hash rate for ETH increases fast until the second quarter of May 2021 and then fluctuated and decreased to the minimum value on 28 June 2021. Figure 4.15 shows that there is a strong positive correlation between the average hash rate and ETH closing price (0.87).

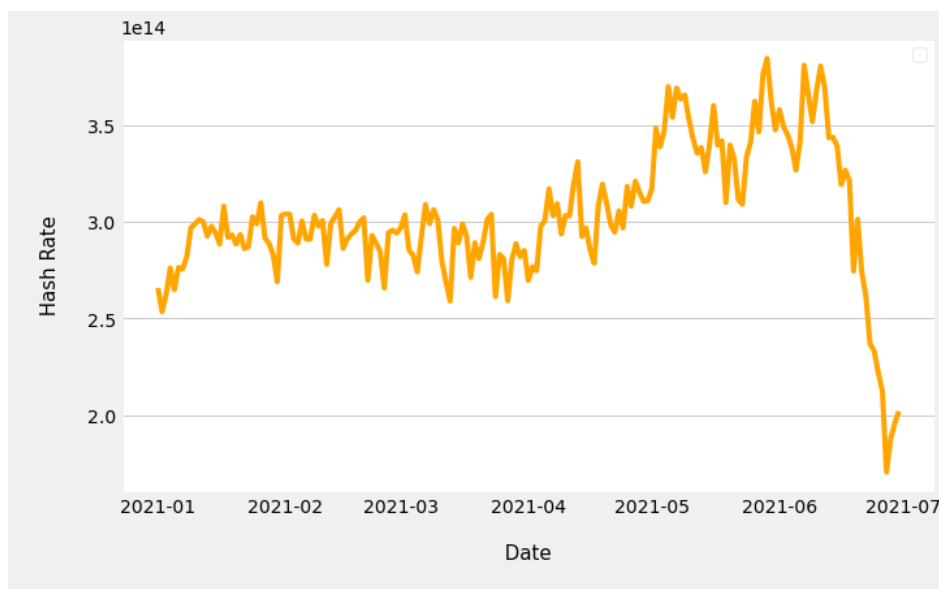


Figure 4.16. LTC average hash rate.

Figure 4.16 illustrates the average rate of LTC, it shows that the average hash rate steeply drops at the end of the interval. The average mean is 3.0×10^{14}

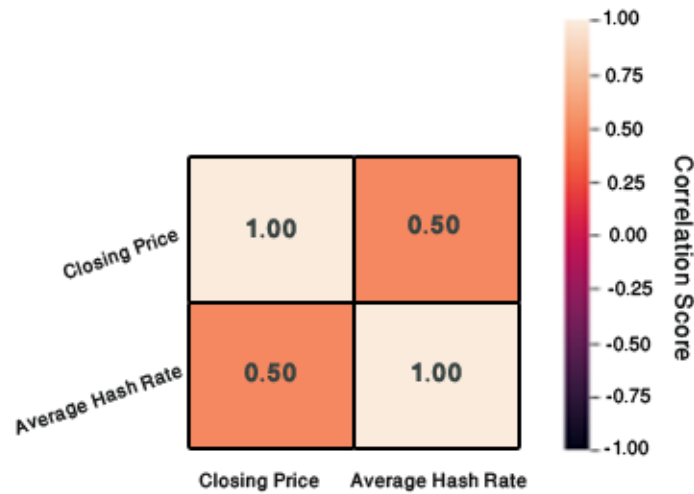


Figure 4.17. Correlation matrix between LTC closing price and average hash rate.

Figure 4.17 shows a moderated correlation between the average hash rate and closing price for LTC (0.5). In the same interval of decreasing average hash rate, we notice that the trading volume and closing price decreased also.

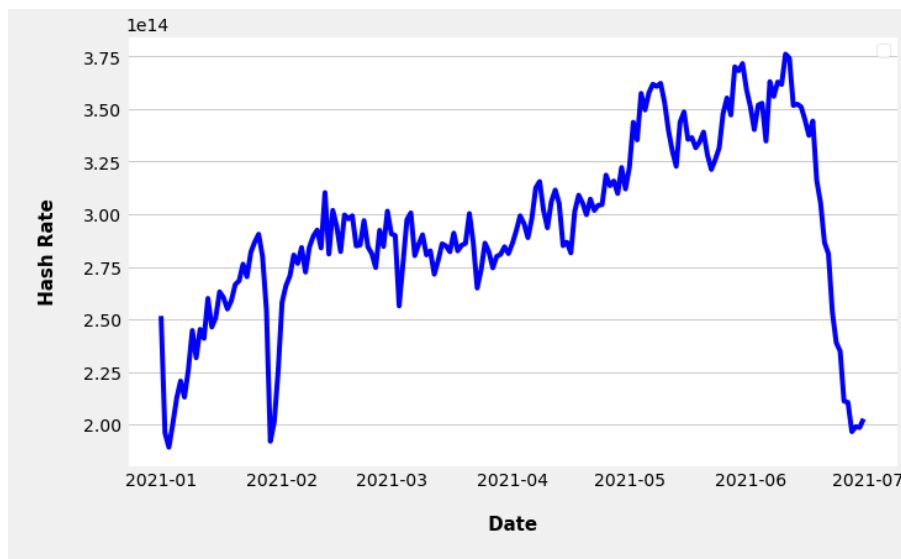


Figure 4.18. LTC average hash rate.

Figure 4.18 shows the average rate of BTC, the graph shows that the average hash rate steep drop at the end of the interval on 29 Jan 2021. At the same time, there is a smooth increase in the middle of the targeted interval.



Figure 4.19. Correlation matrix between Doge closing price and average hash rate.

Figure 4.19 shows a moderated positive correlation between the average hash rate and closing price for DOGE (0.67) which represents a strong correlation.

In general, the correlation coefficient between the average hash rate and the closing price for targeted cryptocurrencies, there is a variety of correlations there is a strong correlation between the average hash rate and the closing price for DOGE and ETH and a Moderate for LTC, and a weak for BTC. Moreover, the hash rate has a steep drop at the end of the interval.

4.3.2 Average block size

The amount of transactions in each block determines the appropriate or optimal block size for an application. Block size optimization is a critical issue for any blockchain-based application since it has a direct impact on the program's performance because scalability limitations can prevent increased throughput and generate congestion. In comparison to the lower block size, a bigger block size will require more transmission time [115].

A smaller block is more efficient, but clearing all of the transactions will take longer if the block is built too small. Both performance factors are incompatible with one another. An efficient blockchain network requires a block size that requires less transmission time and block composition time. Regarding the targeted four cryptocurrencies within the period from 01 Jan 2021 to 30 Jun 2021, Figure 4.20 illustrates the average block size for targeted cryptocurrencies.

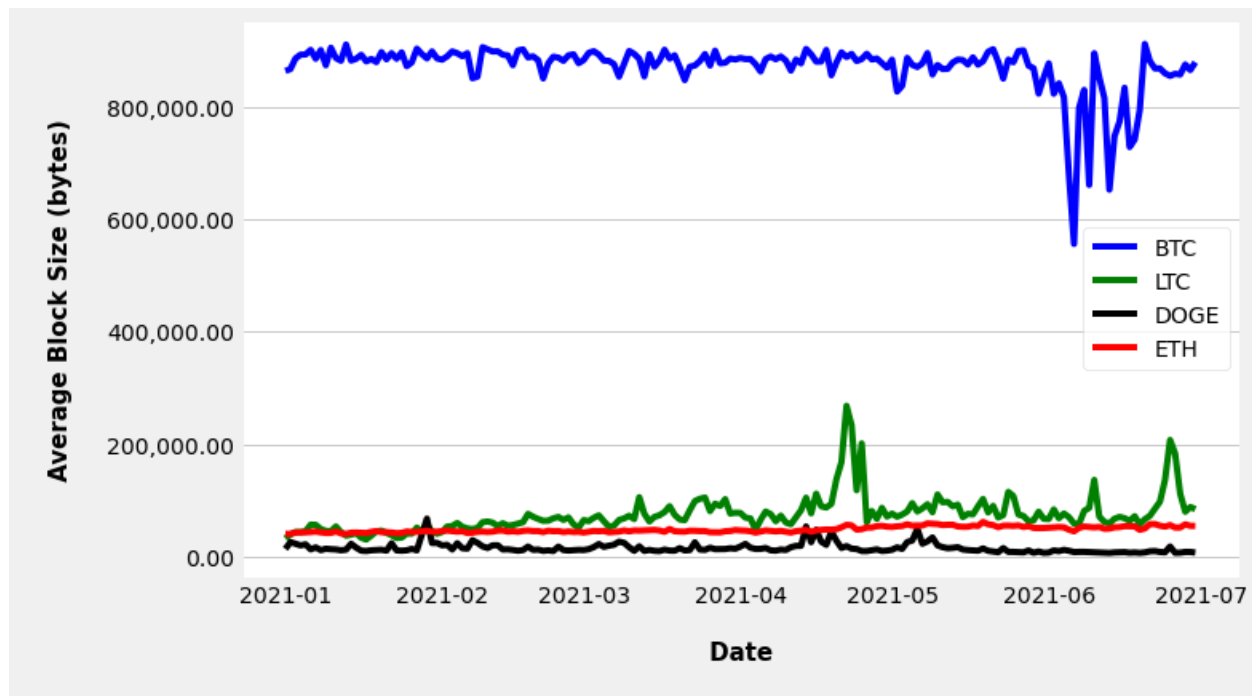


Figure 4.20. BTC, ETH, LTC, and DOGE average BLOCK size.

For BTC the average block size is 869885.12 bytes which represents the biggest average block size by comparing with another targeted cryptocurrency (48605.77 bytes for ETH and 75063.13 bytes for LTC which can be considered as the second average block size, and 15716.91 bytes for DOGE).

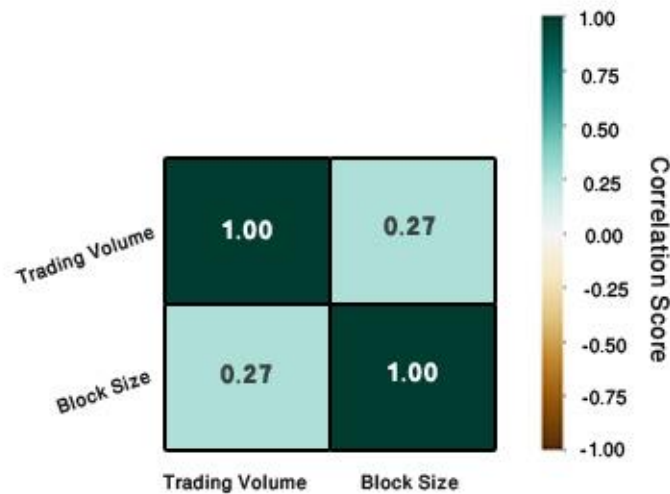


Figure 4.21-A - The correlation between trading volume and block size for BTC average block size.

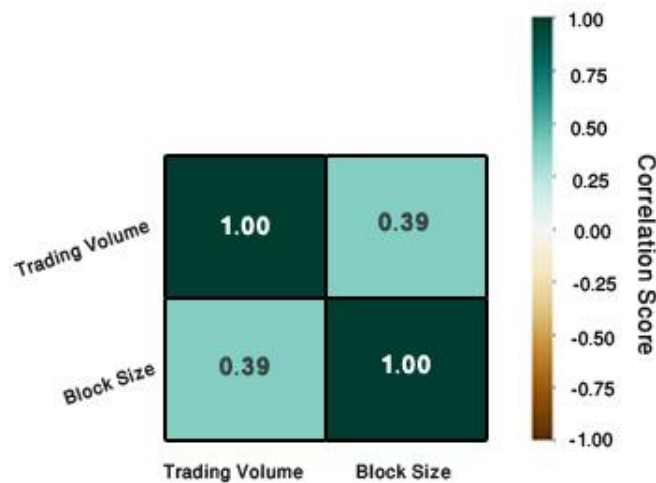


Figure 4.21-B. The correlation between trading volume and block size for ETH average block size.

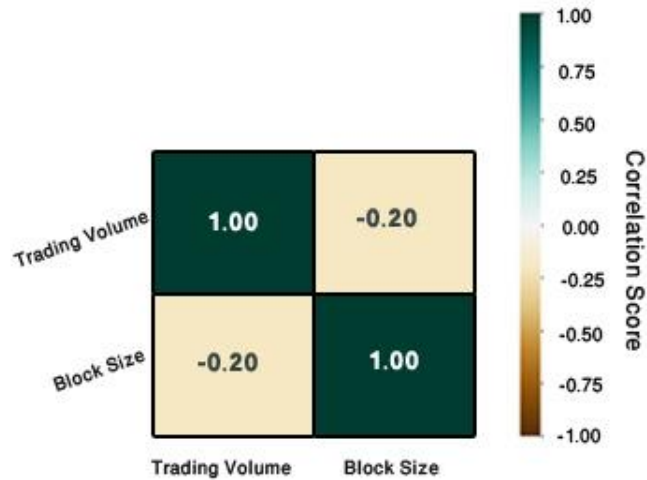


Figure 4.21-C. The correlation between trading volume and block size for LTC average block size.

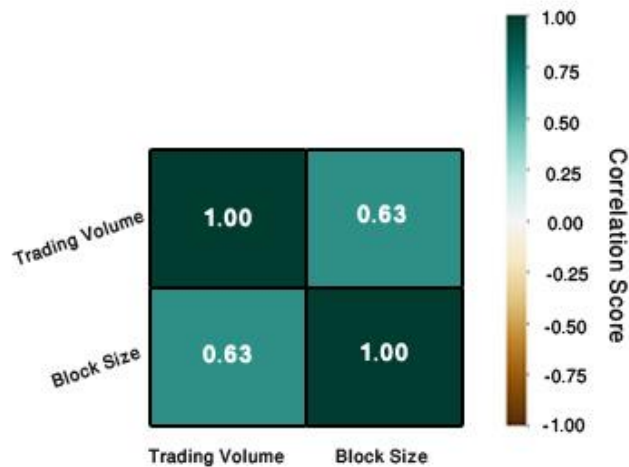


Figure 4.21-D. The correlation between trading volume and block size for DOGE average block size.

Figures 4.21-A, B, and D show there is a weak correlation between the volume of trading and average block size for BTC, ETH, and LTC with values of (0.27, 0.39, and -0.20 respectively). Figure 4.21-C illustrated the correlation between the average block sizes and the trading volume for DOGE, it shows that the correlation is strong with a value of 0.63.

4.3.3 Market capitalization

Market capitalization is significant in crypto since it is used to assess a cryptocurrency's overall potential and to compare its value to that of other cryptocurrencies. Market capitalization might reflect a cryptocurrency's development potential and whether it's a safe investment when compared to others [116]. Small-cap: cryptocurrencies have a market cap of less than \$1 billion and are the most vulnerable to market sentiment movements Mid-cap: Cryptocurrencies with market caps of \$1 billion to \$10 billion are considered mid-cap, they offer a bigger potential gain, but they also carry a higher risk. On the other hand, Large-cap cryptocurrencies, such as Bitcoin and Ethereum, have a market capitalization of more than \$10 billion. Because they have previously demonstrated a strong track record of growth, these are considered the safest investments. Large capitalization can also endure bigger amounts of investors cashing out without significantly affecting the price. Figure 4.22 illustrates the average market capitalization which shows that BTC is the biggest market capitalization and LTC is the smallest.

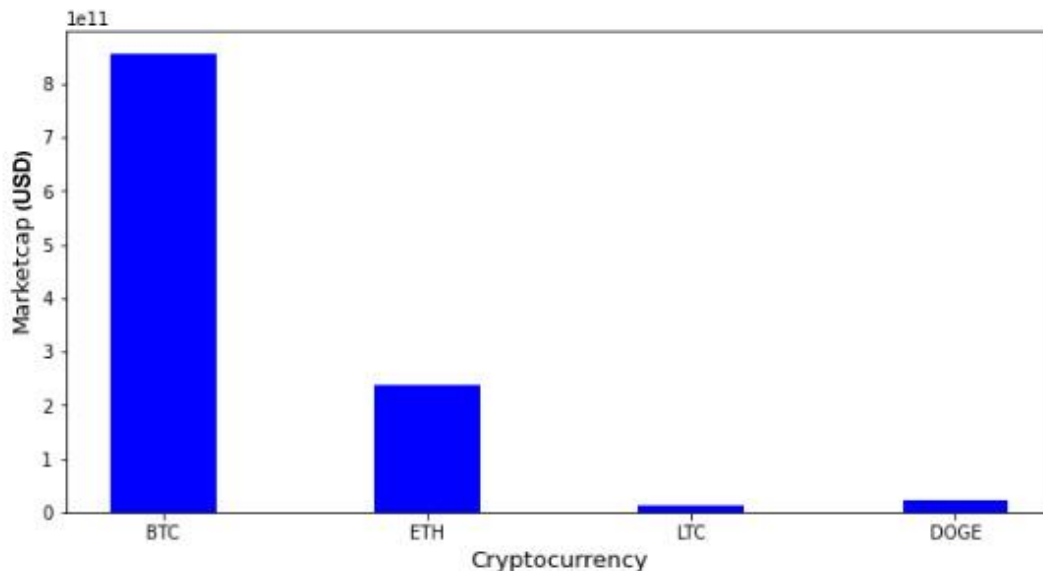


Figure 4.22. BTC, ETH, LTC, and DOGE market capitalization.

Figure 4.23 illustrates the correlation between trading volume, market capitalization, and closing price for BTC, ETH, LTC, and DOGE respectively.



Figure 4.23-A. The correlation between trading volume, market capitalization, and the trading volume for BTC.



Figure 4.23-B. The correlation between trading volume, market capitalization, and the trading volume for ETH.



Figure 4.23-C. The correlation between trading volume, market capitalization, and the trading volume for LTC.



Figure 4.23-D. The correlation between trading volume, market capitalization, and the trading volume for DOGE.

Figures 4.23 A-D illustrates the correlation between Market capitalization, closing price, and trading volume. It shows that there is a strong positive correlation between market capitalization and closing price for all of the targeted cryptocurrencies, but the correlation between market capitalization and trading volume and market capitalization and closing price is weak for BTC, ETH, and LTC. But the correlation between market capitalization and trading volume is moderate and the same for the correlation between trading volume and closing price with values of 0.55 and 0.56 respectively.

4.3.4 Mining difficulty

The constant growth of processing power may drastically lower the time it takes to find a new valid block as more individuals participate in Bitcoin mining. The Bitcoin network's difficulty will be changed every 2016 block to maintain the system's stability [117]. Figure 4.24 A-D illustrates the mining difficulty for BTC, ETH, LTC, and DOGE.

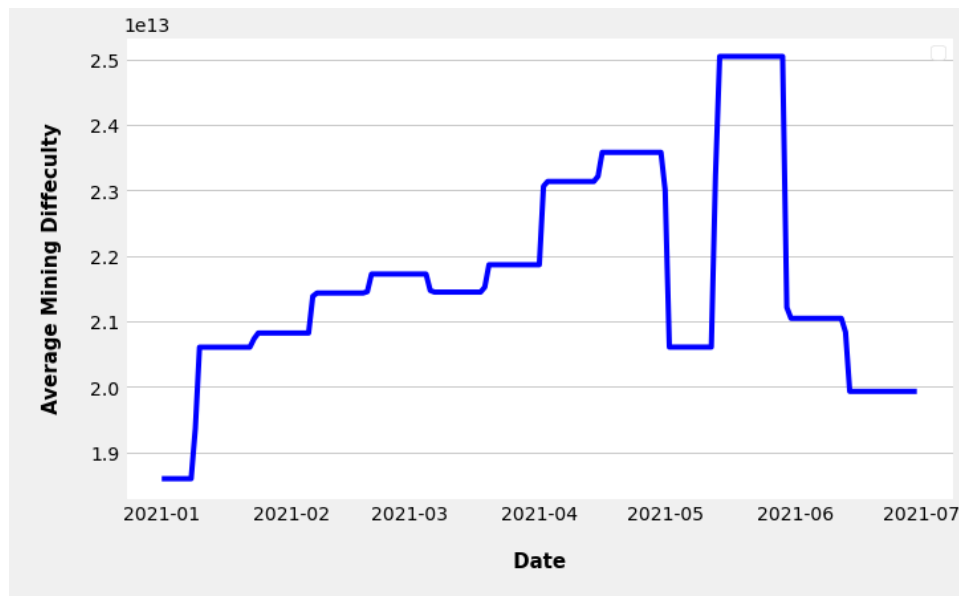


Figure 4.24-A. BTC average mining difficulty.

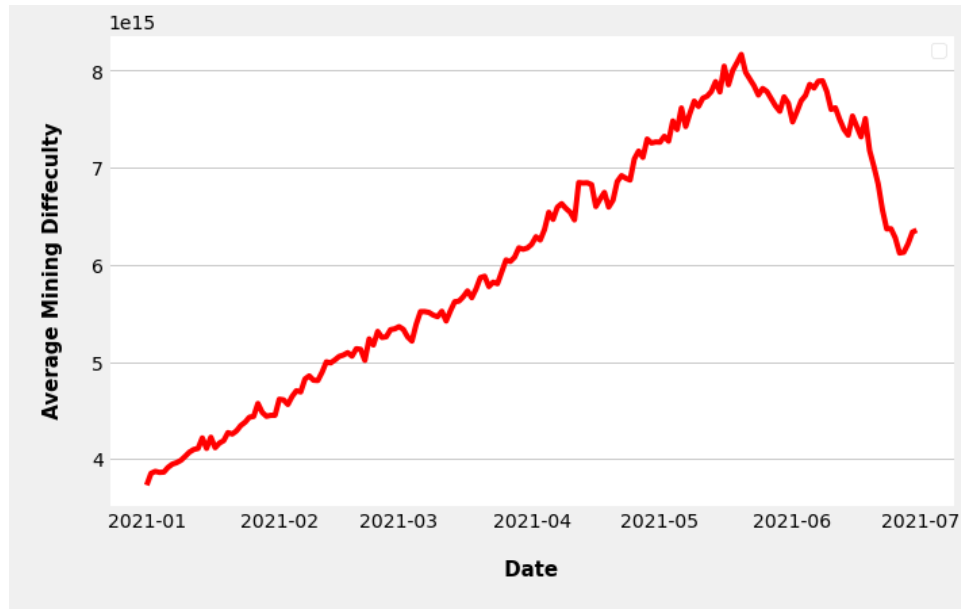


Figure 4.24-B. ETH average mining difficulty.

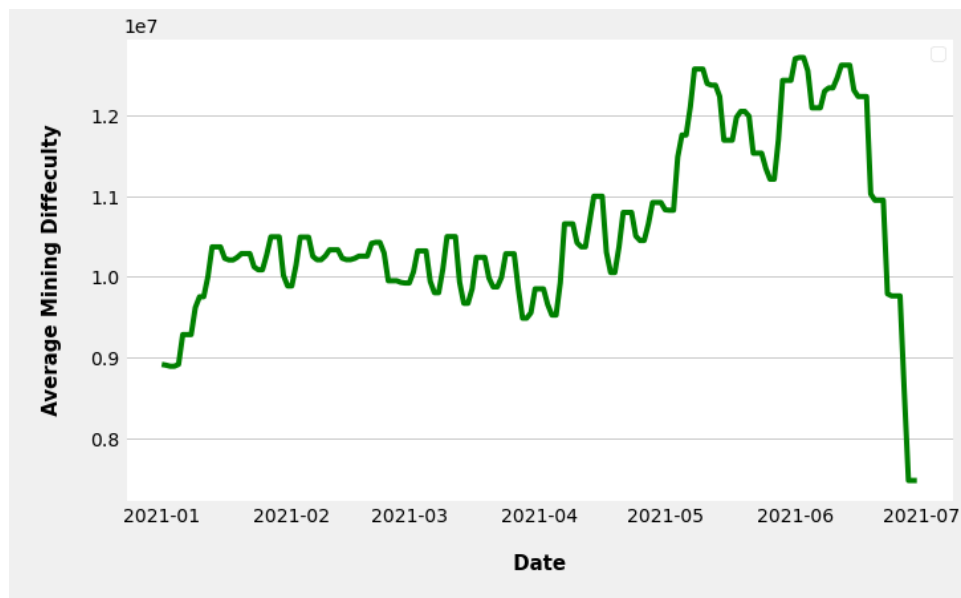


Figure 4.24-C. LTC average mining difficulty.

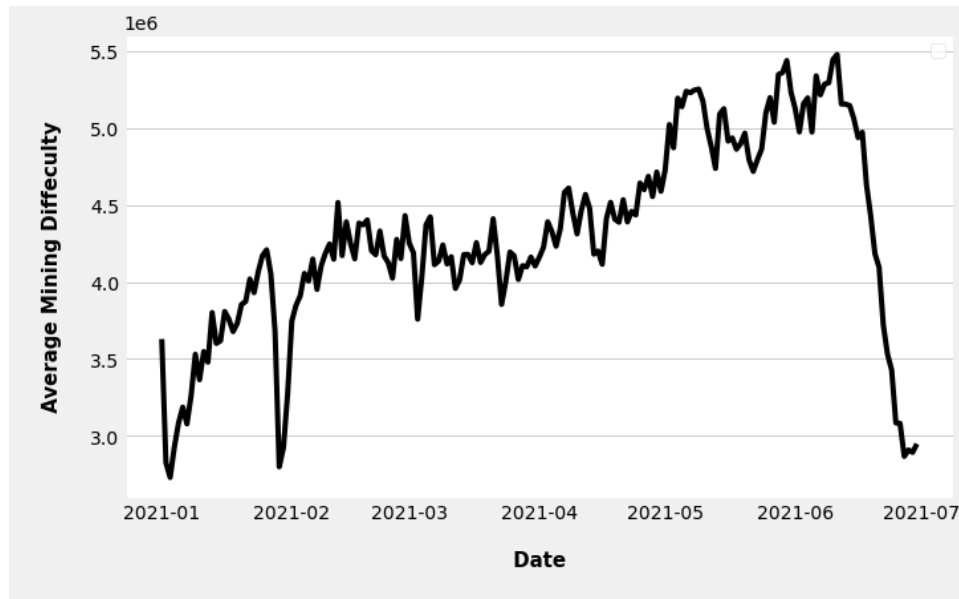


Figure 4.24-D. DOGE average mining difficulty.

Figure 4.23 shows that ETH mining difficulty has the largest and LTC was the smallest within the targeted interval. At the same time, the mining difficulty for the targeted cryptocurrency suffers from a strong decrease at the end of the targeted interval.

4.3.5 Transaction fees

A healthy blockchain should contain the following metrics: high honest mining power, low processing latency (short transaction wait times), and a consistent price of consumption (the minimum transaction fee necessary to complete a transaction) [118]. Demand for blockchains like Bitcoin and Ethereum outnumbers supply, necessitating a mechanism that selects a subset of transactions from a pool of all pending transactions to include "on-chain." EIP-1559 proposes several improvements to the Ethereum blockchain's transaction fee structure, including the addition of variable-size blocks and a burnt basic charge that grows and lowers in response to demand [119]. Figure 4.25 illustrates average transaction fees for BTC, ETH, LTC, and DOGE,

although the transaction fee is very close for both BTC and ETH, the average fee for BTC is bigger with a mean of 17 USD and 14 for ETH, on other hand LTC and DOGE show very close the mean values 0.037 USD and 0.5 USD for LTC and DOGE respectively.

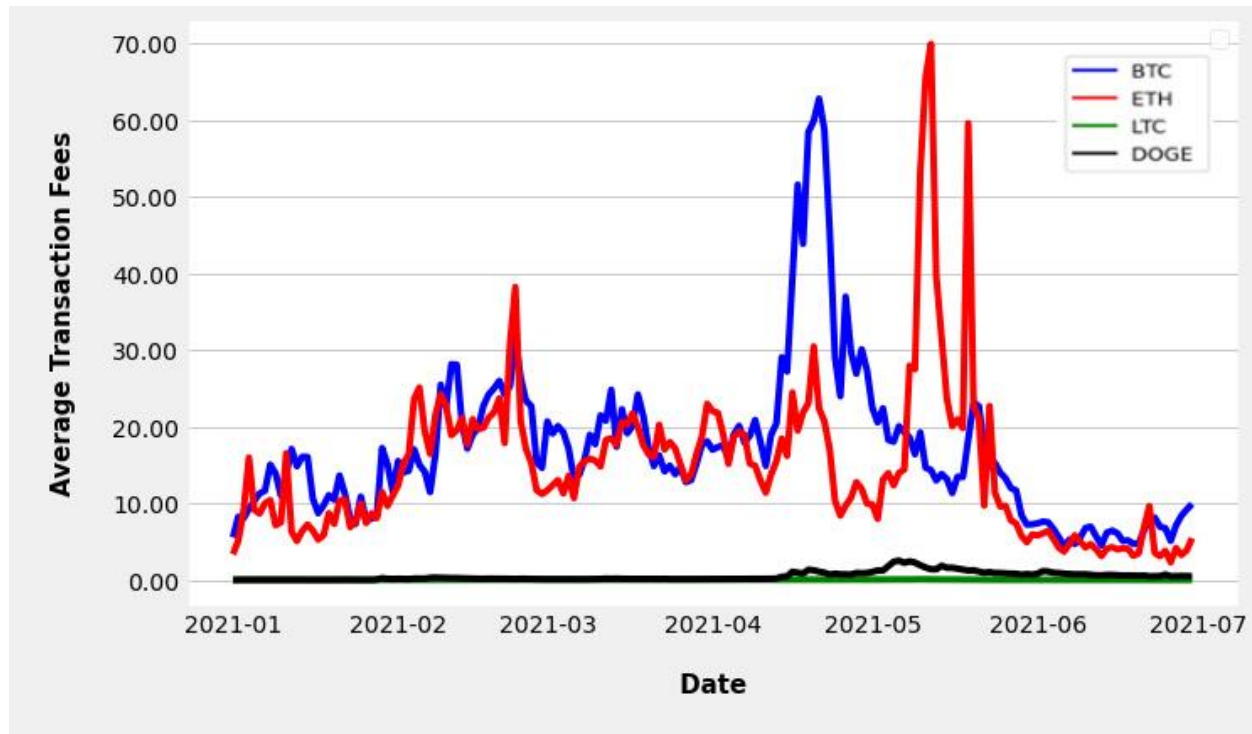


Figure 4.25. BTC, ETH, LTC, and DOGE average transaction fees.

Figure 4.26 illustrates the correlation between transaction fees and closing price and transaction volume, Figure 4.26-A and 4.26-B show that there is a moderate correlation between transaction fees and closing price, on the other hand, it indicates a weak correlation between volume and transaction fee for BTC and ETH. Figures 4.26-C and 4.26-D show that there is a strong correlation between both transaction fees and closing price and transaction fees and transaction volume.

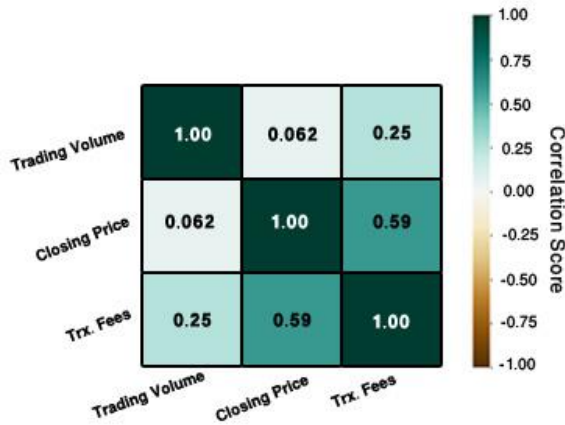


Figure 4.26-A. BTC correlation matrix for transaction fees, closing price, and transaction volume.

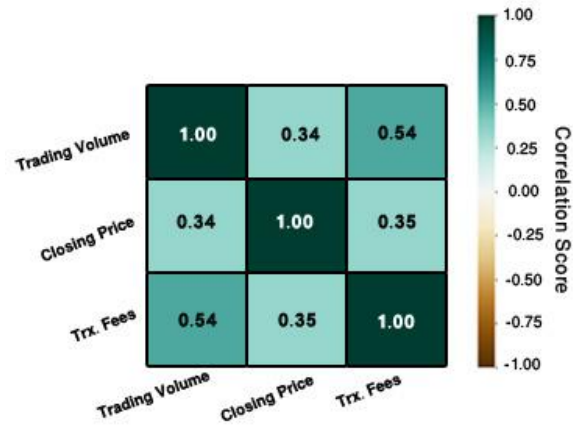


Figure 4.26-B. ETH correlation matrix for transaction fees, closing price, and transaction volume.

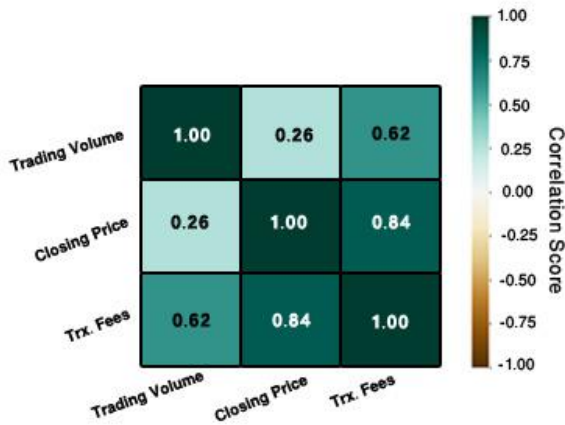


Figure 4.26-C. LTC correlation matrix for transaction fees, closing price, and transaction volume.

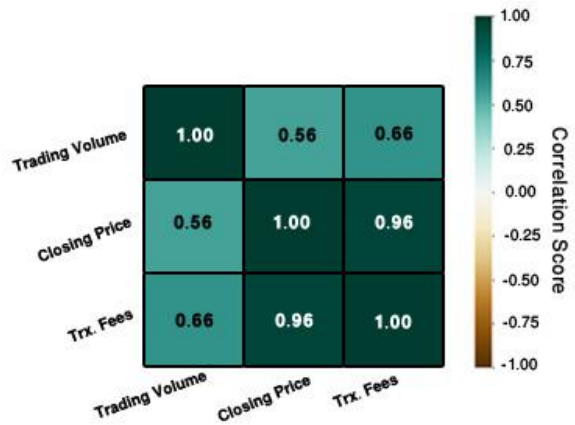


Figure 4.26-D. DOGE correlation matrix for transaction fees, closing price, and transaction volume.

Figure 4.26. BTC, ETH, LTC, and DOGE correlation matrix for transaction fees, closing price, and transaction volume.

4.4 Google trend

Previous research has suggested that Google Trends can be used as a good indicator of investor interest [120]. The economic activities are represented using Google trend data derived from keyword searches relating to economic activity. Figures 4.27 – 4.30 show the search interest for the given region and period at the highest point on the chart. A score of 100 indicates that the term is at its most popular, while a value of 50 indicates that it is half as common. Figures 4.27 – 4.30

illustrate trends in distribution over the world for BTC, ETH, LTC, and DOGE respectively. It shows that for BTC the most popular trend was in El Salvador in Figure 4.27, on another hand the most popular trend for ETH in Figure 4.28 was in Singapore. For LTC the most popular trend was in Nigeria shown in Figure 4.29, for DOGE the trend was in Turkey that shown in Figure 4.30.

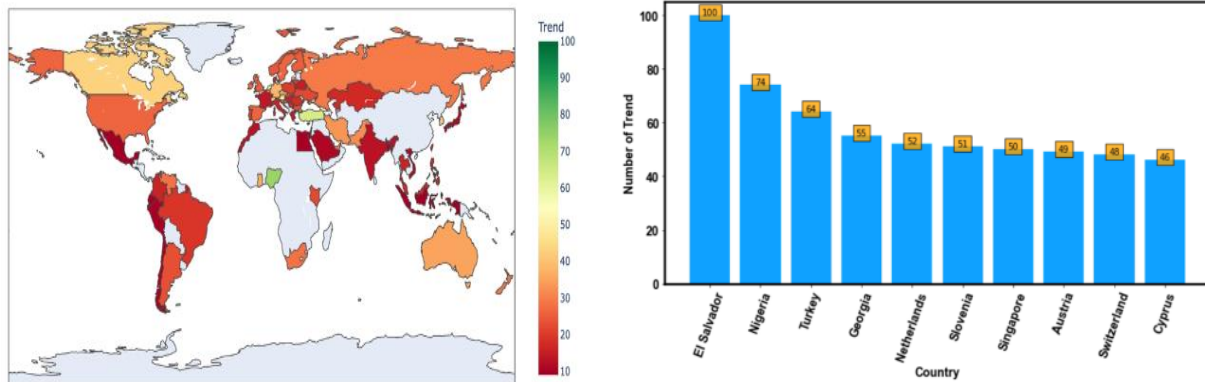


Figure 4.27. BTC trend distribution over the world and top 10 countries.

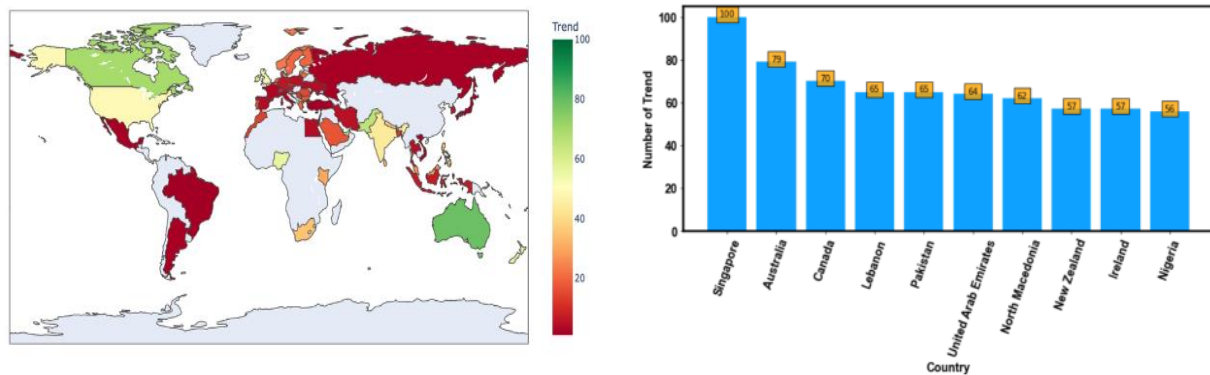


Figure 4.28. ETH trend distribution over the world and top 10 countries.

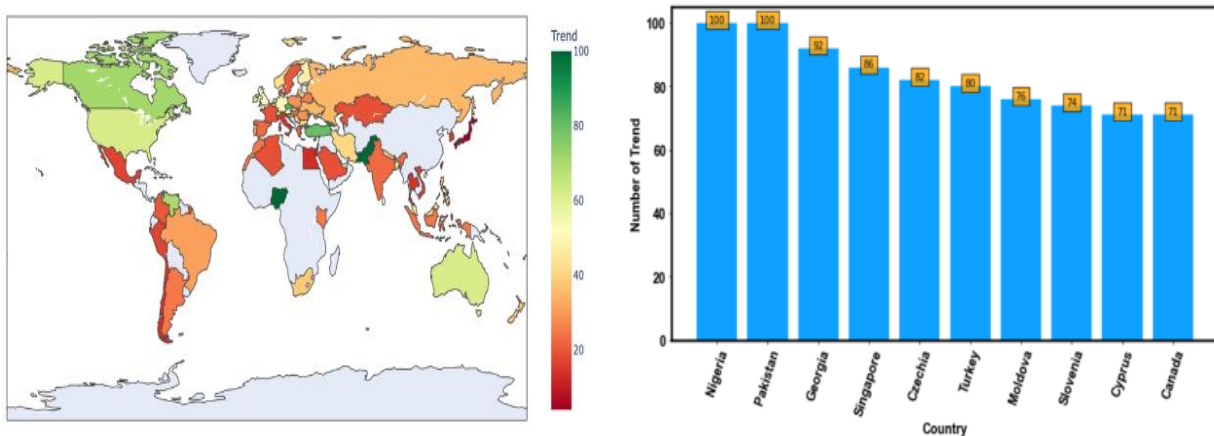


Figure 4.29. LTC trend distribution over the world and top 10 countries.

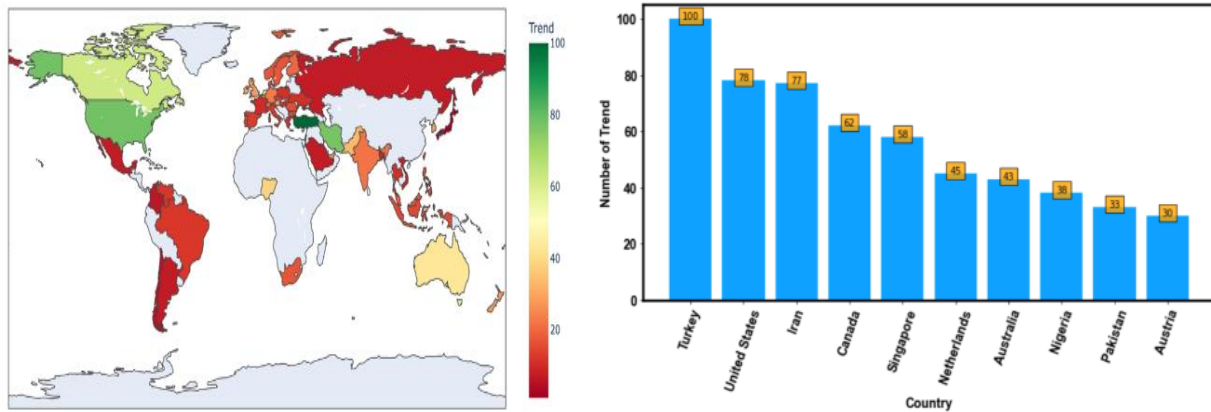


Figure 4.30. DOGE trend distribution over the world and top 10 countries.

4.5 News agencies

Traders and investors now have constant access to updated news, which molds their thoughts and influences their decision to invest in Bitcoin, owing to the wonders of the internet. The data was collected from [121] that collected the data from other sources. The number of the collected article was 11437, 380, 3384, and 1020 articles for BTC, ETH, LTC, and DOGE respectively, with a total number of 16,221 articles for the targeted cryptocurrencies. Figure 4.31 illustrate the number of daily published articles.

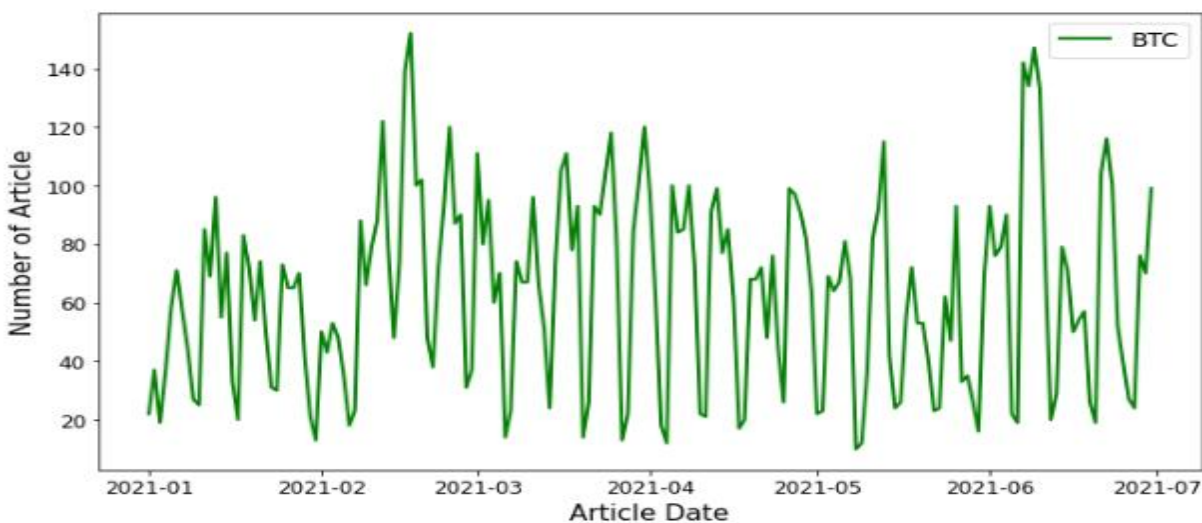


Figure 4.31. News Volume for BTC.

For BTC, the total number of collected news was about 11,437 articles distributed between 1 Jan 2021 and 30 Jun 2021, the maximum number of published articles was 152 articles on 17 Feb 2021 and the minimum number was 10 articles on 08 May 2021.

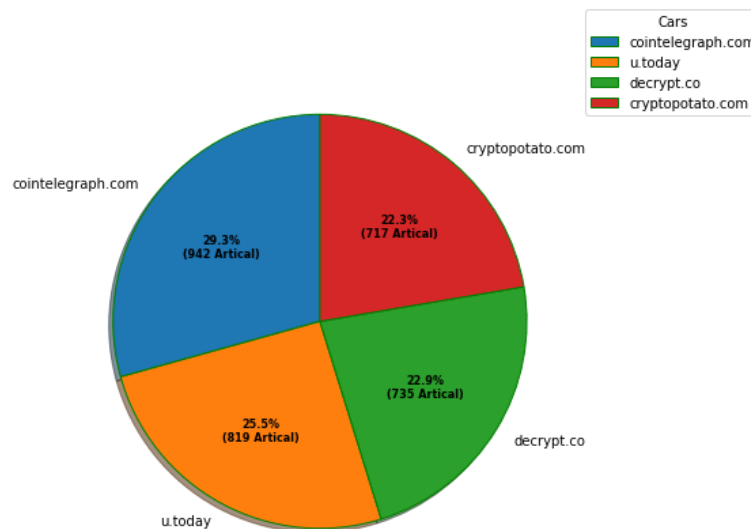


Figure 4.32. The top 4 BTC news sources.

Figure 4.32 illustrates the top 4 websites where the data was collected, most of the articles were collected from cointelgraph.com, and Cointelegraph which is a leading independent digital media site covering blockchain technology, crypto assets, and future fintech trends was founded in 2013.

For ETH, Figure 4.33 shows the total number of collected articles is 3,384 articles between 1 Jan 2021 and 30 Jun 2021, it shows a logical fluctuation over the period except on 04 May 2021.

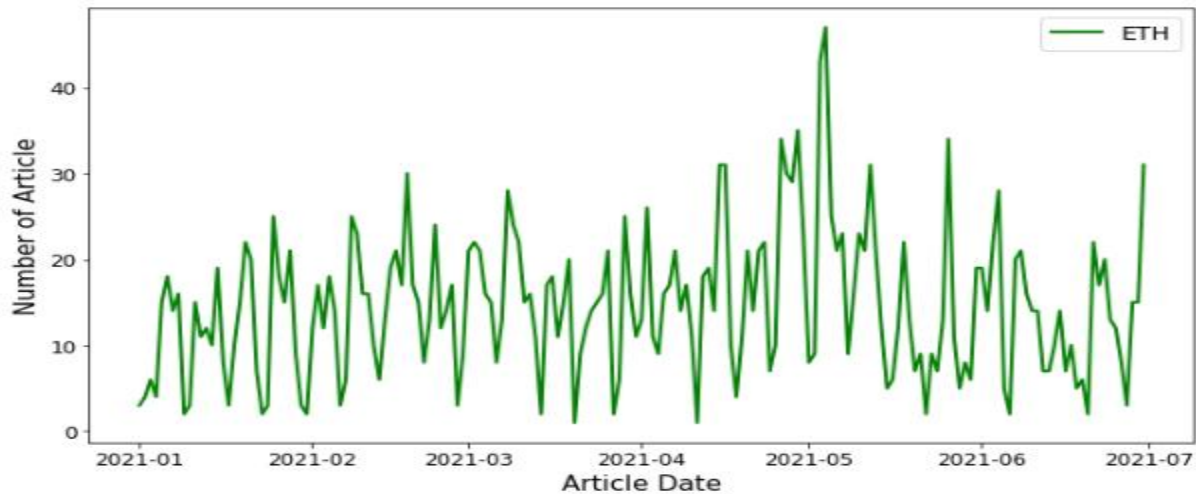


Figure 4.33. News Volume for ETH.

Figure 4.34 shows the top 4 news websites, it shows that the data that its source from decrypt.com represents the largest percentage rather than the top 4 sources with a percentage of 28%, on the other hand, the smallest percentage from the top 4 was from cointttelegraph.com with the percentage of 22%.

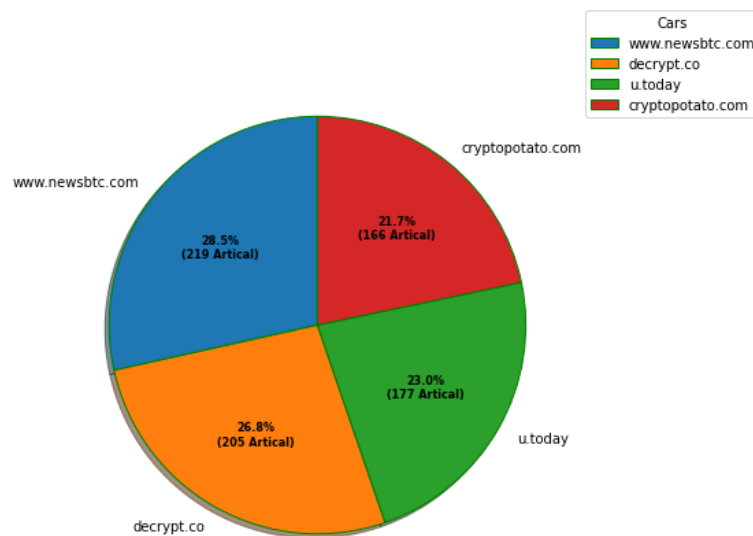


Figure 4.34. Top 4. Top 10 ETH news sources.

For LTC, the total number of collected articles is 380 articles between 1 Jan 2021 and 30 Jun 2021.

Figure 4.35 illustrates the distribution over the period, it shows that the maximum number of articles is 12 achieved on 06 April 2021.

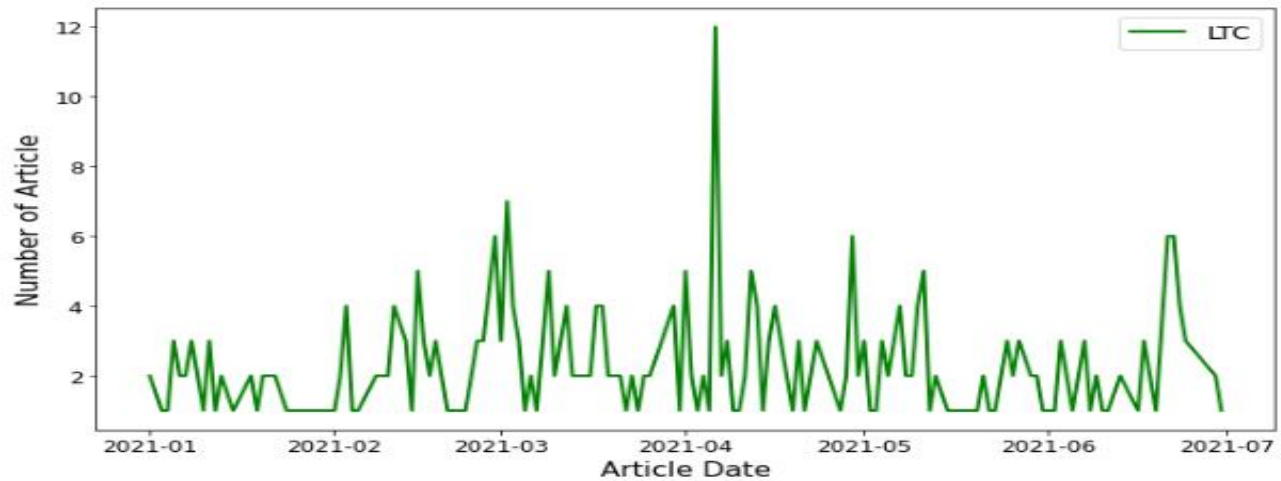


Figure 4.35. News Volume for LTC.

Most of the news was from ambcrypto.com, which was formed in 2018 to make blockchain and cryptocurrency news more accessible to readers. They have swiftly established as the digital news source for a new generation of cryptocurrency aficionados, with over a million monthly readers from all around the world. Figure 4.36 illustrates the top 4 news websites.

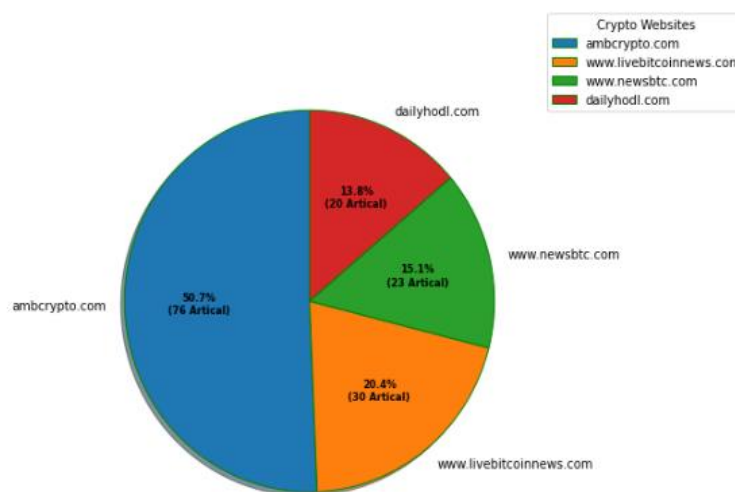


Figure 4.36. Top 4. LTC news sources.

Figure 4.37 illustrates the volume of news over the targeted period, it shows a high fluctuation in its volume, especially in the period from the end of April until the end of May. The total number of collected the article was about 1020 articles for DOGE.

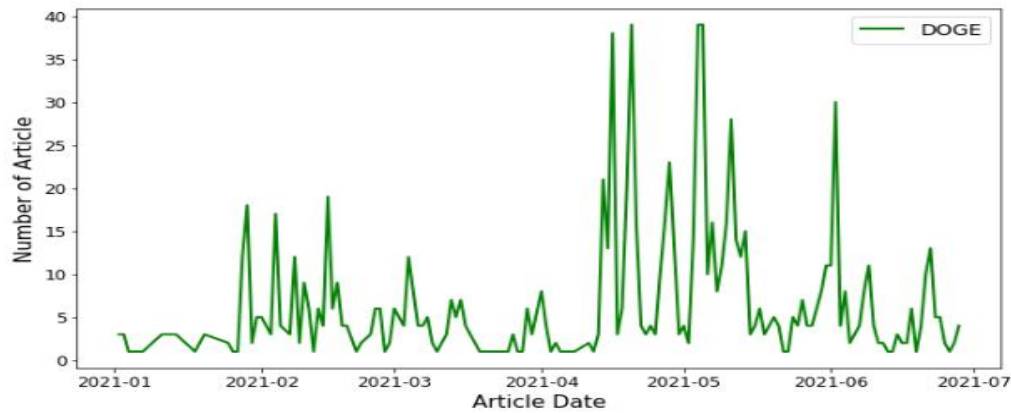


Figure 4.37. News Volume for DOGE.

The news was collected from multiple sources from a specialized website in cryptocurrency; Figure 4.38 illustrates the percentage of the top 4 websites in direction of collected data. The most data was from u.today, which is a global news organization dedicated to assisting you in comprehending new-generation technology and its implications for society. By adhering to the essential journalistic ideals of clear thinking, honesty, and independence.

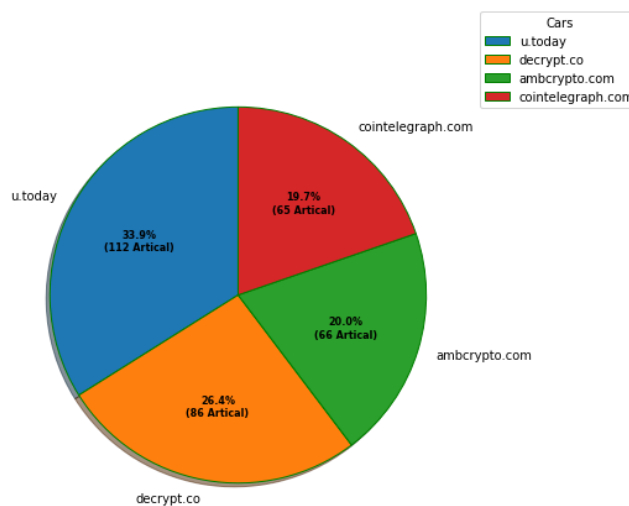


Figure 4.38. Top 4 Cryptocurrency news sources.

4.6 Social media (Twitter)

Twitter today has over 200 million unique daily users from all over the world, who produce over a billion pieces of content every 48 hours [122]. Figures 39-42 illustrate the daily tweets for the targeted interval for BTC, ETH, LTC, and DOGE respectively.

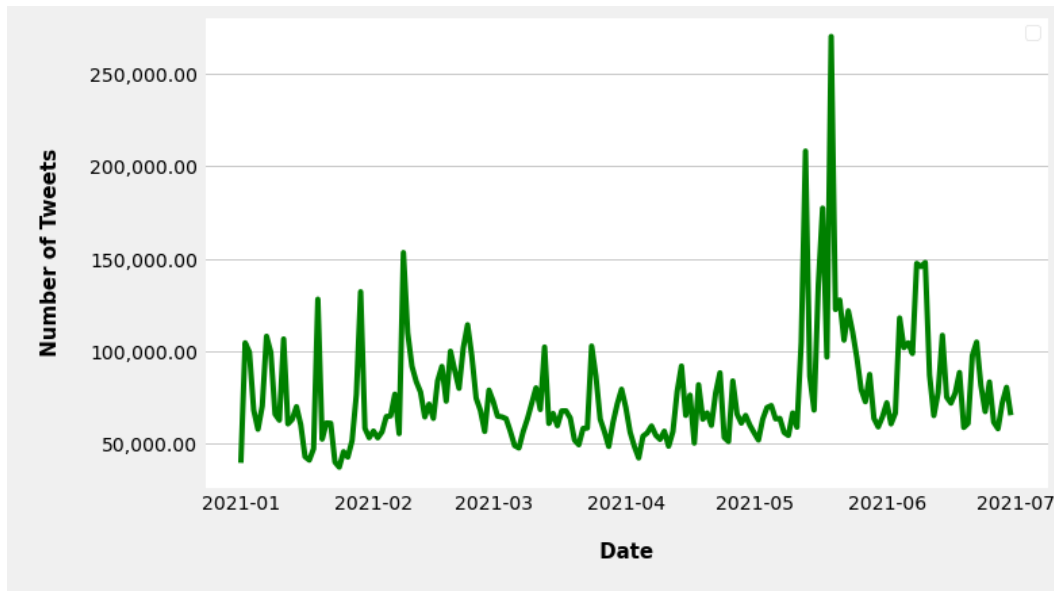


Figure 4.39. BTC daily tweets.

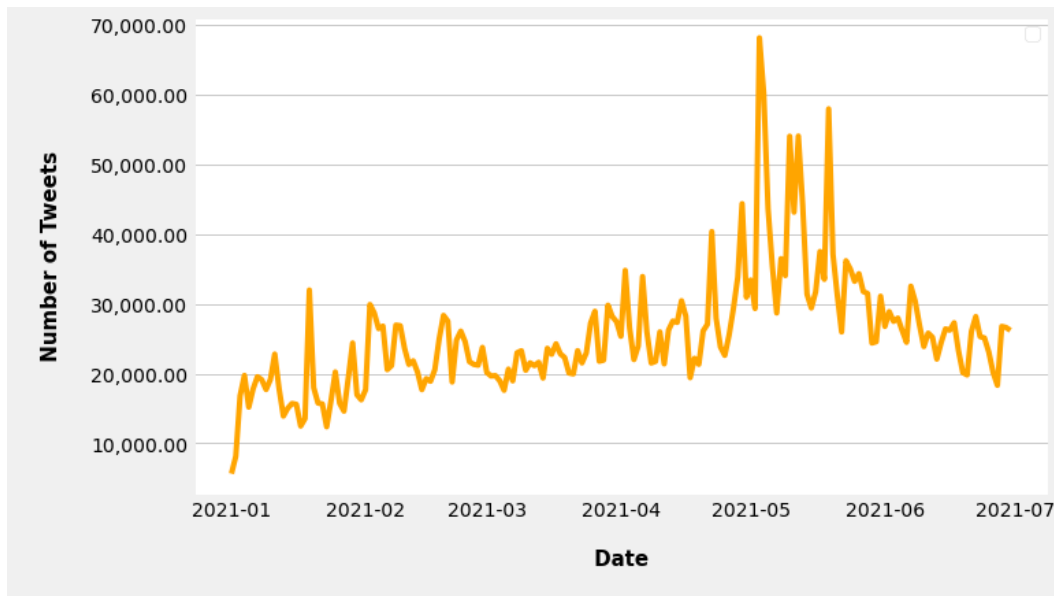


Figure 4.40. ETH daily tweets.

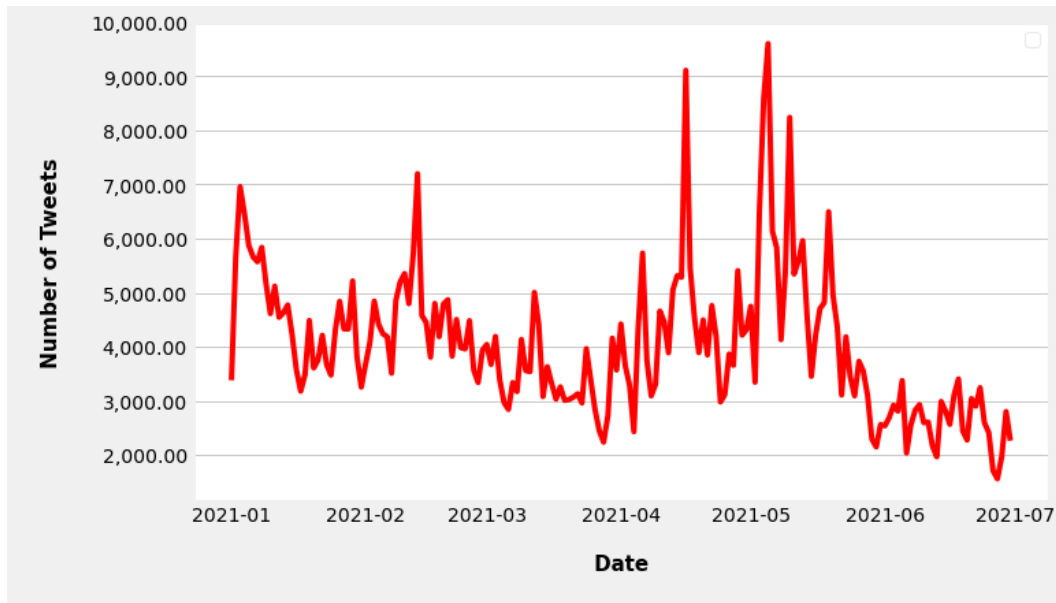


Figure 4.41. LTC daily tweets.

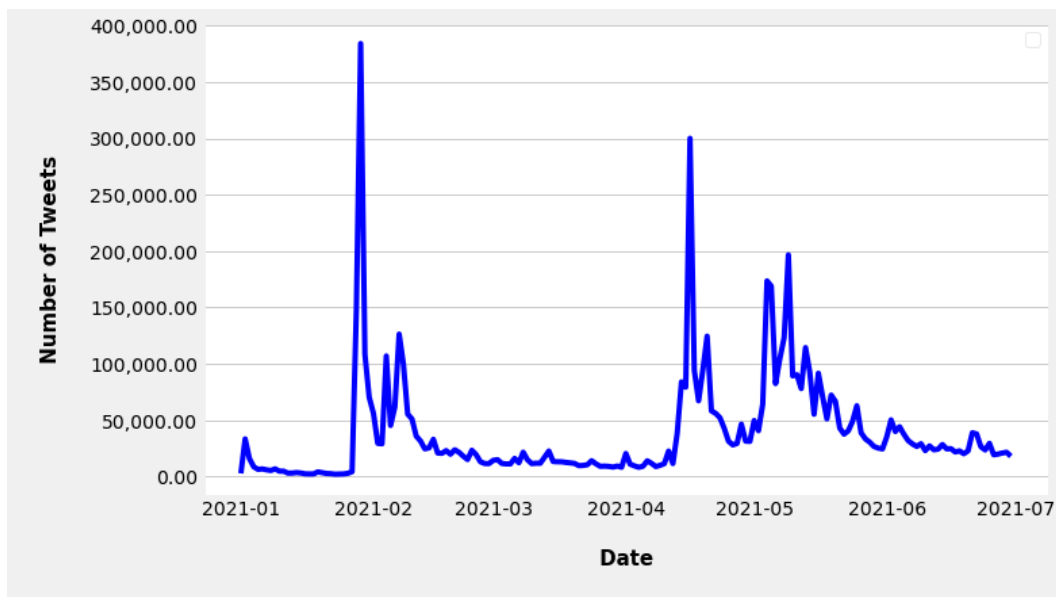


Figure 4.42. DOGE daily tweets.

Figure 4.39 illustrate the number of tweets for BTC per day between 1 Jan 2021 and 30 Jun 2021, the minimum number of tweets is 37,383 tweets, and the maximum number of tweets is 269,880

tweets, on the other hand, the average number of tweets is 76,460 tweets. Figure 4.40 illustrate the number of tweets for ETH per day in the same interval, the minimum number of tweets is 5,651 tweets, and the maximum number of tweets is 68,108 tweets, on the other hand, the average number of tweets is 25,467 tweets. For LTC Figure 4.41 show the tweets per day, the maximum number of tweets is 9,650 tweets and the minimum number of tweets is 1,559 with average tweets of 4029. Figure 4.42 illustrates the daily tweets for DOGE, the minimum number of tweets is 2,086, the maximum number of tweets is 384,161, and the average number of tweets is 37,936.

4.7 Conclusion

Data visualization is a process that can give an overview of the data, it simplifies and views the big data as a simple graph than can be understood by a human. This process can be highlighted or redirect the focus to important information. EDA significantly enhances an analyst's fundamental knowledge of several factors. Also, can assist analysts in locating any anomalies, large errors, or missing values in their dataset. Moreover, it can find important patterns.

In the next chapter, I will discuss the proposed models in detail, in this chapter we proposed multiple models to predict cryptocurrency prices for four different types of cryptocurrency those are: BTC, ETH, LTC, and DOGE using multiple channels.

Chapter Five

The Proposed methodology and models

5.1 Introduction

This chapter aims to present the proposed models to predict cryptocurrency prices depending on multiple data sources using the machine learning algorithm LSTM, the data that feed the models was collected from free access data sources like Twitter, google trends, news agencies, gold price, production cost information and other financial websites like yahoo finance. The second step of the methodology is the data preprocessing stage, in this stage, the data has been cleaned and prepared to be ready to be fed to the machine learning model. The third stage is data visualization and exploration (see chapter four), and the last stage is fitting the model and extracting the results.

Figure 5.1 illustrates the proposed system structure.

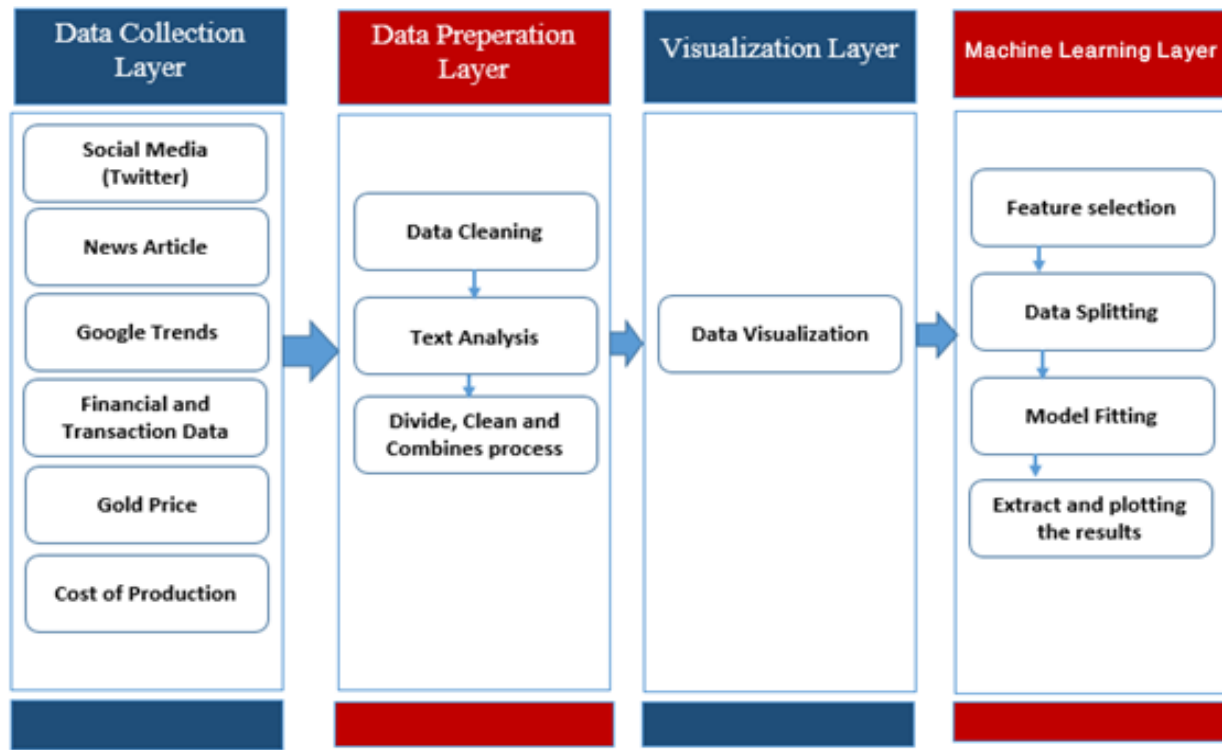


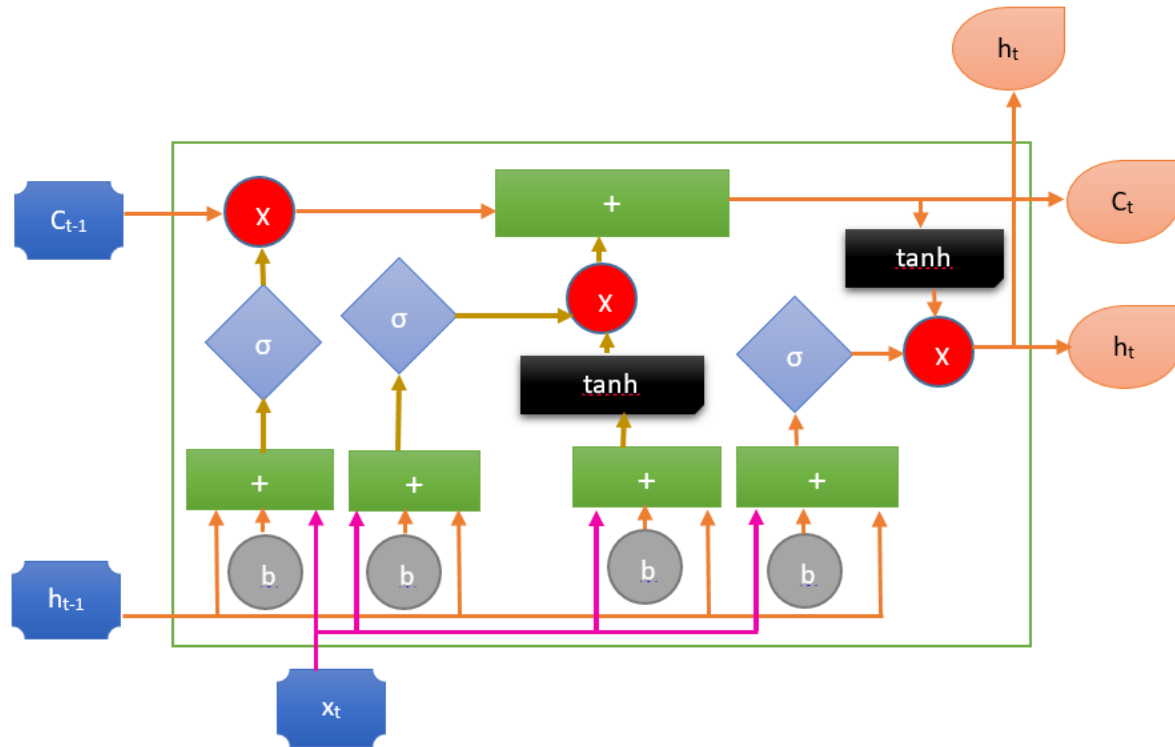
Figure 5.1 Proposed system to predict cryptocurrency prices.

5.2 Feature selection

Because the highly correlated characteristics will provide the same information; by it multiplying the value of that information and allowing it to dominate the model. Therefore, only one of the highly correlated features is used. Features with little correlation to the target are deleted as they will not improve the model performance as well as simplify the model or speed up the learning method. To isolate the highly correlated feature and the features with little correlation with the target; I used the filter method feature selection which depends on the correlation coefficient to apply a threshold and to remove the features with a high correlation.

5.3 Long Short-Term Memory (LSTM) Algorithm

Modeling multivariate time series and time-variant systems is particularly well suited to LSTM networks. As a result, the difference between real system outputs and expected outputs which are predicted by the network can be used to discover anomalies [123]. LSTM is one type of RNN that is considered an effective and scalable and robust technique for numerous learning problems using sequential data. Because its features are considered an excellent and amazing tool for capturing long-term temporal dependencies. The LSTM is an RNN-inspired design with gates that control information flow between its parts called cells. The input and forget gate structures can change the information that travels through the cell state, with the final output being a filtered version of the cell state dependent on the inputs' context [124]. The LSTM design has been chastised for being haphazard and for containing a large number of components whose function is unclear. As a result, it's hard to say if the LSTM is the greatest design. As a result, it's questionable whether the LSTM is the best design, and better ones are almost certainly available [125]. The framework of an LSTM algorithm is depicted in Figure 5.2.






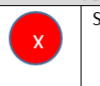





Inputs		Outputs		Nonlinearities		Vector Operation	
	Current Input		New Update Memory		Sigmoid Layer		Scaling of information
	The memory from the last LSTM unit		Current Output		Tanh Layer		Adding Information
	The output of the last LSTM unit						

Figure 5.2. LSTM working mechanism architecture.

The following equation is the description of the forwarding training process of the LSTM [126]:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \dots\dots\dots \text{equation (2)}$$

$$i_t = \sigma(w_f \cdot [h_{t-1}, x_t] + b_i) \dots\dots\dots \text{equation (3)}$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \dots\dots\dots \text{equation (4)}$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \dots\dots\dots \text{equation (5)}$$

$$h_t = o_t * \tanh(C_t) \dots\dots\dots \text{equation (6)}$$

Where:

i_t : represents the activation of the input gate.

O_t : represents the activation of the output gate.

f_t : represents the activation of the forget gate.

C_t : activation vector for each cell and memory block.

h_t : activation vector for each cell and memory block.

W: weighted matrix.

b: is a bias vector.

$\sigma(\cdot)$: sigmoid function

5.4 Data preprocessing and completion

Before using any machine learning algorithms, it's critical to prepare the data. Applying any machine learning algorithm to text data, like any other data, necessitates data preparation. Data preprocessing is used to prepare the data. Preprocessing text entails removing noise such as stop words, punctuation, and keywords that aren't important in the context of the text, among other things [127]. A supervised ML algorithm's generalization performance is frequently influenced by data preparation. One of the most difficult challenges in inductive ML is removing noise occurrences. The deleted instances usually have a lot of deviating instances with a lot of null feature values. The data preprocessing stage in this thesis for collected text includes; data cleansing, removing punctuation, and removing stop words.

5.4.1 Data cleaning

Data cleaning is a process of filtering the datasets to delete the null values and completing missing values by taking the previous observation. This stage is important and needs high care because any mistake may lead to losing important data. Also, deleting duplicates in data is very important for machine learning in the direction performance-wise and avoiding learning a machine with the same cases because of duplication.

5.4.2 Remove punctuation

Punctuation is a powerful indication of syntactic structure, and parsers trained in punctuated text frequently rely on it. Punctuation, on the other hand, is a distraction since human language processing does not use punctuation to the same level, and we commonly leave it out of casual writing [128].

5.4.3 Remove stop words

Using pre-compiled stop word lists or more advanced algorithms for dynamic stop word recognition, removing stop words from textual input is a common approach for reducing noise. However, in recent years, the efficacy of eliminating stop words in the context of social media mood categorization has been questioned [129].

5.5 Price prediction based on English twitter data

As a research area, sentiment analysis of Twitter data has gotten a high of attention and interest. Because of the brief language used in tweets, academics have been drawn to the opportunity to get

information about public opinions by studying Twitter data and automatically identifying its sentiment polarity [130].

5.5.1 Sentiment analysis based on VADER lexicon sentiment analysis for text analysis

VADER is a sentiment analysis tool that is exclusively geared to the sentiments expressed in social media and is based on vocabulary and laws, it is an open-source program that is free to access. VADER considers word order as well as degree qualifiers [131]. Figure 5.3 illustrates a flow chart that shows the systematic steps for tweets analysis using VADER and Harvard IV-4. The scraping process was done on daily basis using a snsrape package because of the size of scrapped data and stored in multiple .csv files by the date, the next stage is to clean the collected tweets and apply sentiment analysis techniques, and sore a result in a single file.

VADER proposed two types of measurement that measure the vigor of emotion the first is for single words the scale is between -4 and 4, and -4 represents the most negative emotion and 4 represents the most positive emotion, and the number 0 denotes a neutral feeling. Moreover, VADER represents another measurement scale for sentences; this scale is between -1 and 1 where -1 represents the most negative and +1 represents the most positive. Figure 5.4 illustrates the measurement concept.

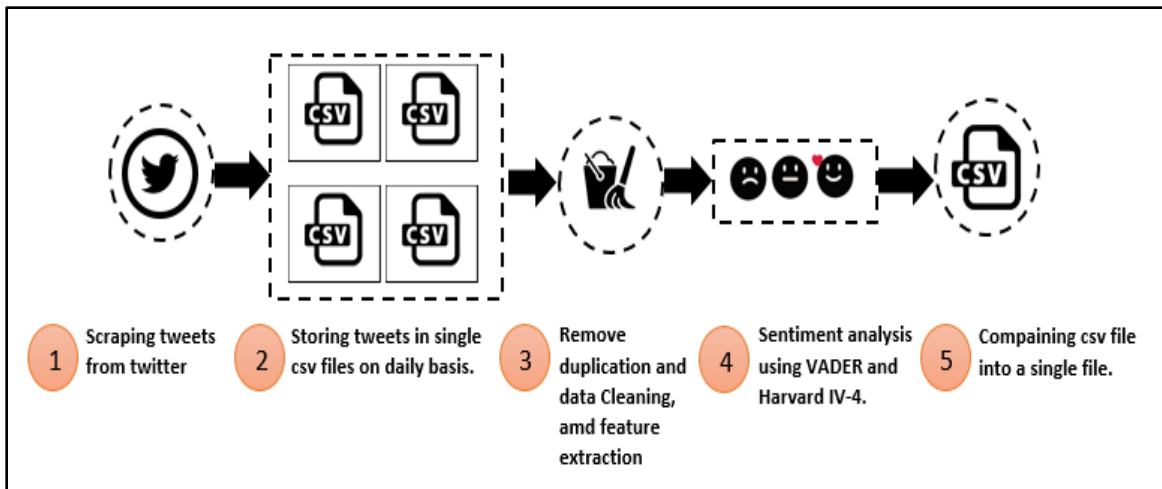


Figure 5.3. The systematic process of tweets sentiment analysis.

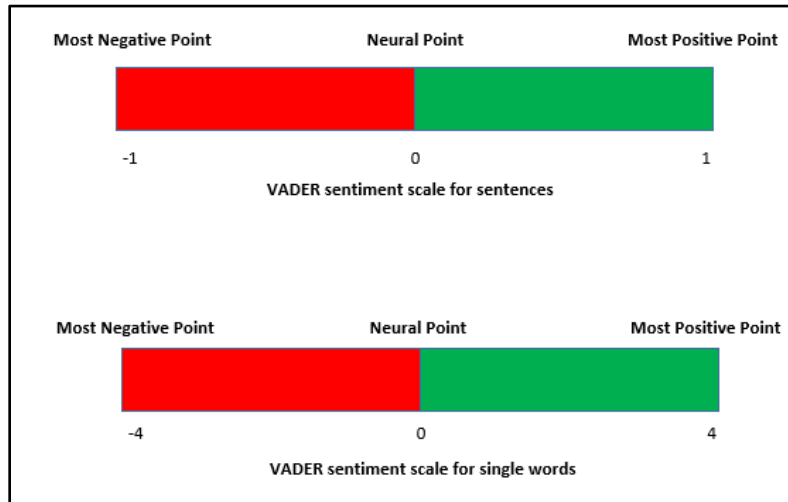


Figure 5.4. VADER sentiment analysis measurement concept.

The execution of Sentiment analysis is: -

- Negative (-).
- Neutral.
- Positive (+).
- Compound (computed by normalizing the above scores).

Data Normalization is the process of scaling the input variable to the range 0-1, this range is the range for floating-point values. Data normalization is a preprocessing technique in which data is measured to ensure that each feature shares and contributes equally to machine learning [132]. The closer the Compound score is to +1, the more positive the text is, and the Valence scores for the words are used to generate these scores.

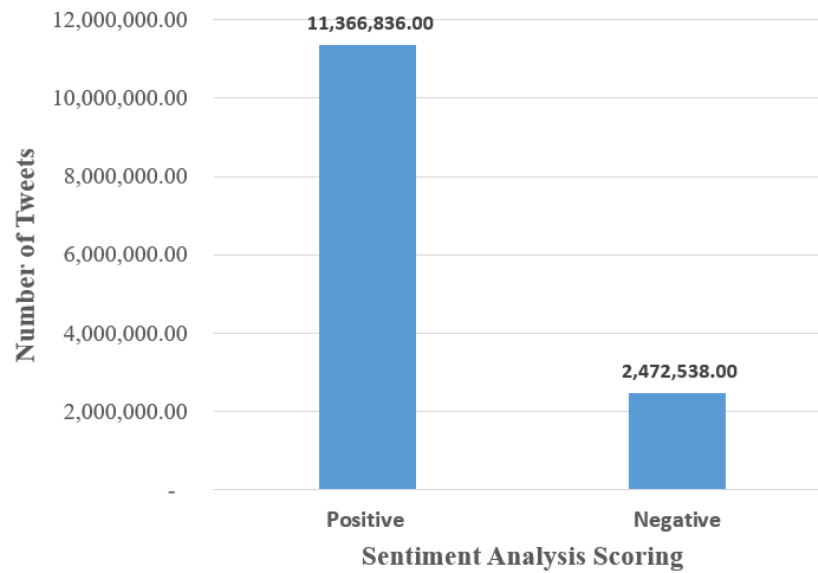


Figure 5.5. Positive and negative tweets distribution for BTC.

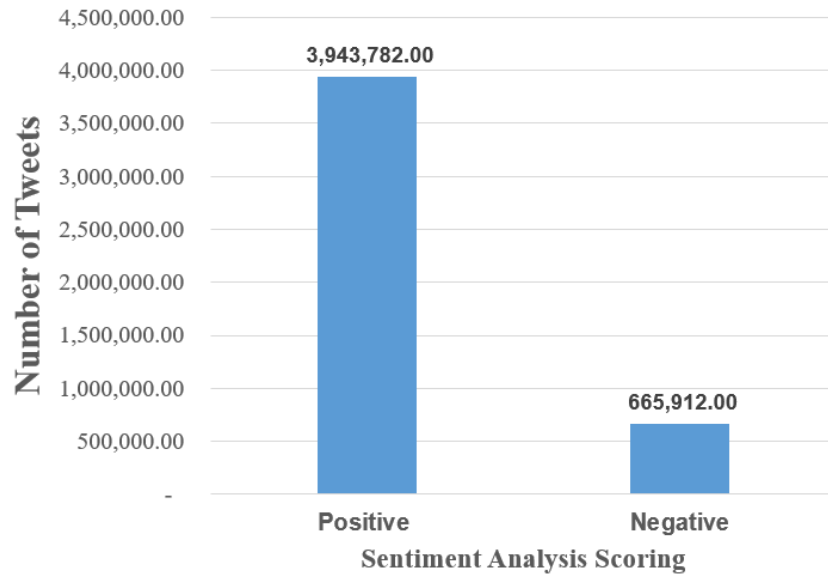


Figure 5.6. Positive and negative tweets distribution for ETH.

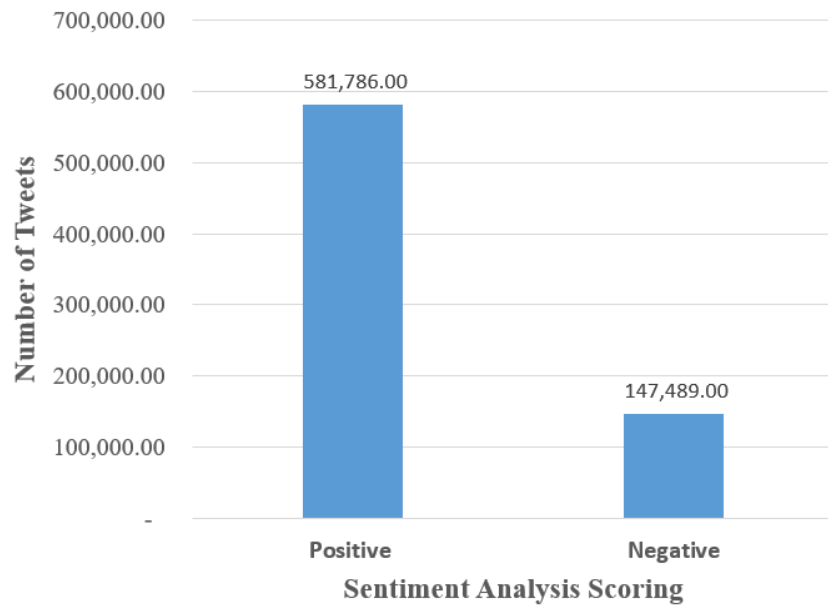


Figure 5.7. Positive and negative tweets distribution for LTC.



Figure 5.8. Positive and negative tweets distribution for DOGE.

Figure 5.5 illustrates the distribution of positive and negative tweets; It shows that approximately 10% of tweets represent negative tweets and 90% were positive tweets for BTC cryptocurrency for the targeted dataset with a total number of clean tweets of 13,839,374.00 tweets. For ETH Figure 5.6 illustrates the sentiment analysis classification it is shown that the positive tweets are more than the negative ones with a percentage of 85.50% and a negative represents 14.45% with a total number of 4,609,690 tweets. Figure 5.7 illustrates the distribution of positive and negative tweets for LTC; It shows that approximately 20% of tweets represent negative tweets and 80% were positive tweets for BTC cryptocurrency for the targeted dataset with a total number of clean tweets of 729,275 tweets. For DOGE there are 6,866,502.00 collected tweets there is 86% represent positive tweets and 24% represent negative ones.

Analyzing the text using VADER sentiment analysis represents three outputs that are positive, negative, and neutral, and with closing price, the figures below illustrate the relationship between those variables for BTC, ETH, LTC, and DOGE respectively using a correlation matrix.

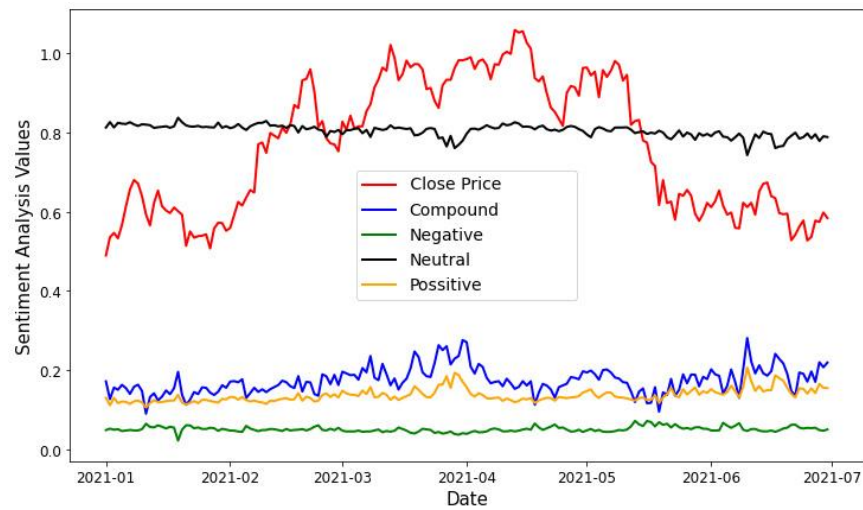


Figure 5.9-A. BTC VADER Sentiment analysis output parameters.

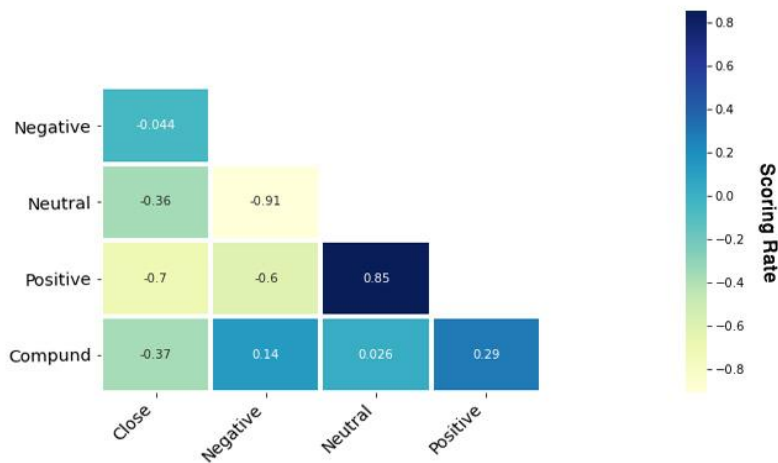


Figure 5.9-B. BTC VADER Sentiment analysis output parameters correlation matrix.

For BTC there is a strong positive relationship between neutral and positive. Regarding the other parameters the relation is weak, Figure 5.9-B illustrates the correlation between the parameters.

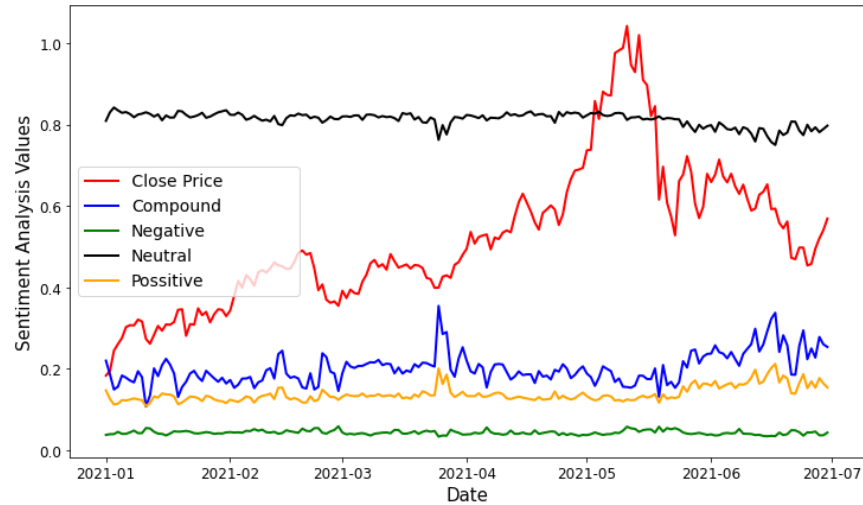


Figure 5.10-A. ETH VADER Sentiment analysis output parameters.

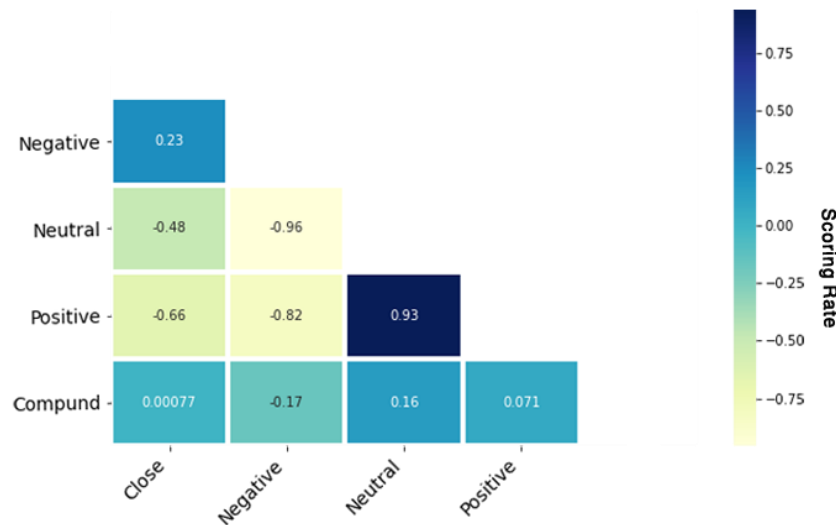


Figure 5.10-B. ETH VADER Sentiment analysis output parameters correlation matrix.

For ETH Figure 5.10-A there is a strong negative relationship between positive and negative and positive and a low positive correlation between close and neutral, a moderate negative correlation between the closing price and positive, and a strong positive between neutral and positive tweets. And the other has a weak correlation.

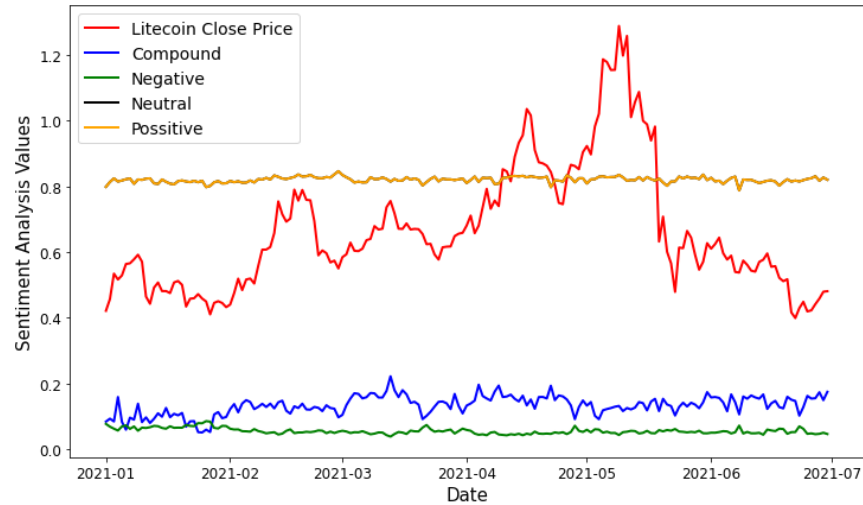


Figure 5.11-A. LTC VADER Sentiment analysis output parameters.

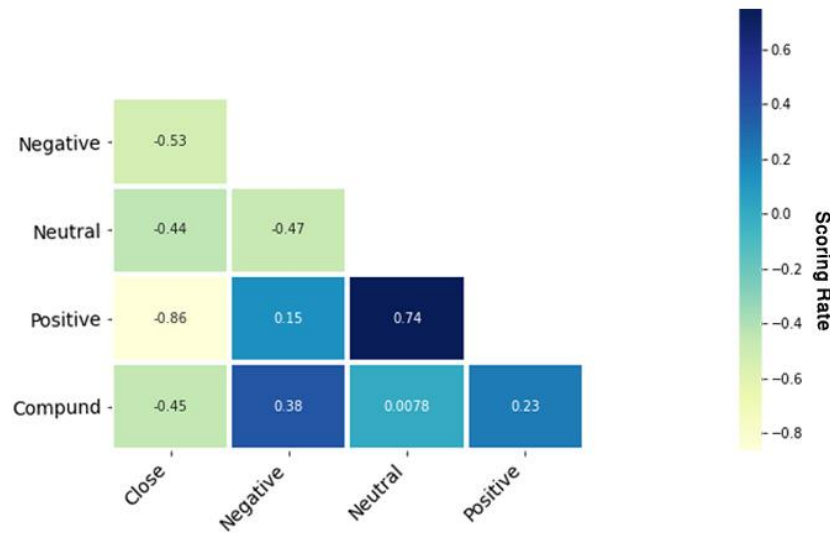


Figure 5.11-B. LTC VADER Sentiment analysis output parameters correlation matrix.

The output of VADER Sentiment result shows that there is a low positive relationship between compound and positive analysis, and a low negative relationship between compound and negative output, on the other hand, the other relations are weak and moderate positive and negative for LTC VADR sentiment analysis that shown in the correlation matrix of Figure 5.11-B. Figure 5.11-A

illustrates the fluctuation of the parameters with the targeted period, taking into account the negative, positive, neutral, and compound multiplied constantly to make a clear visualization.

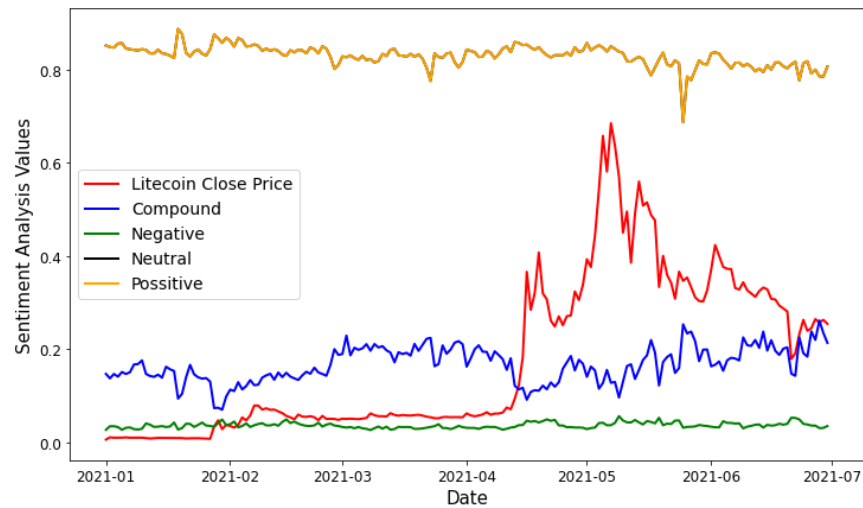


Figure 5.12-A. DOGE VADER Sentiment analysis output parameters.

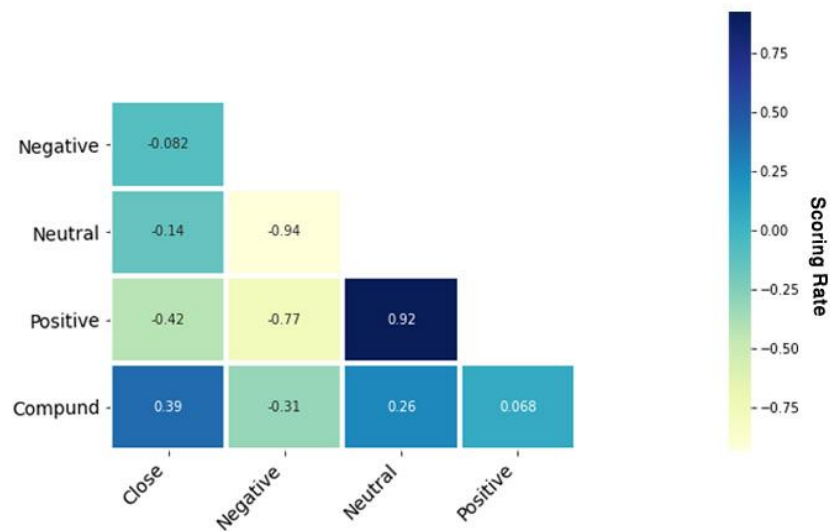


Figure 5.12-B. DOGE VADER Sentiment analysis output parameters correlation matrix.

The output of the VADER Sentiment result shows that there is a negligible relationship between compound and positive analysis, and a weak negative relationship between compound and neutral output, on the other hand, the other relations are between a weak and moderate positive and negative for DOGE VADR sentiment analysis that shown in the correlation matrix figure 5.12-B.

5.5.2 Sentiment analysis based on HARVARD IV-4 dictionary for text analysis

The Harvard-IV dictionary (GI) for psychology is a general-purpose dictionary created by Harvard University. According to the psychological Harvard-IV dictionary 2 utilized in the general inquirer program, the Harvard-IV lexicon contains a list of 1316 positive and 1746 negative words. The general inquirer-augmented spreadsheet includes a Harvard IV-4 dictionary with over 10,000 words and 182 sentiment dimensions [133].

Nearly three-quarters (73.8 %) of the negative word counts in the GI/Harvard list can be attributed to words that aren't often considered bad in a financial context. To address this issue, researchers have created dictionaries/word lists relevant to the finance sector to get more accurate and efficient sentiment scores [134]. Harvard IV-4 outputs are:

- Positive: the number of words (words count) in a positive manner.
- Negative: the number of words (words count) in a negative manner.
- Subjectivity: that is calculated depending on Equation (1).
- Polarity: that is calculated depending on Equation (2).

$$Subjectivity = \frac{Positive+Negative}{Count(*)} \dots\dots\dots \text{Equation (1)}$$

$$Polarity = \frac{Positive - Negative}{Positive + Negative} \dots\dots\dots \text{Equation (2)}$$

The output of the Harvard IV-4 for BTC Figure 5.13-B shows the relations between the Harvard IV-4 output are weak except the relation between the positive and negative is a high negative correlation.

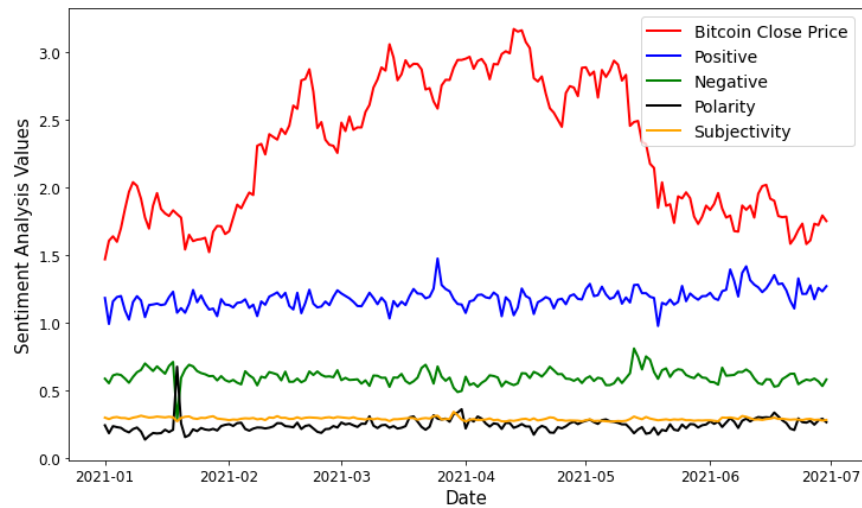


Figure 5.13-A. BTC Harvard IV-4 Sentiment analysis output parameters.

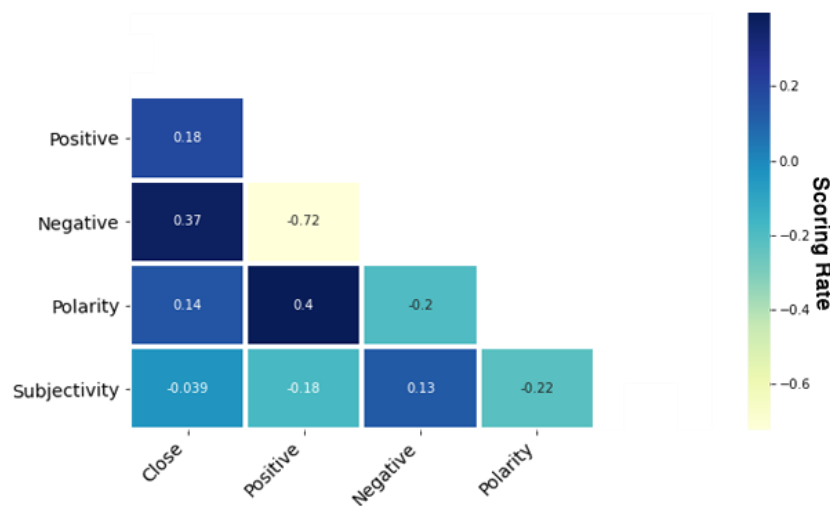


Figure 5.13-B. LTC Harvard IV-4 Sentiment analysis output parameters correlation matrix.

The output of the Harvard IV-4 for ETH the result shows that there is a strong positive relationship between compound and positive analysis, a moderate negative relationship between positive and negative output, and strong relation between positive and polarity and between the negative and closing price. The other parameters have a weak relationship. Harvard IV-4 sentiment analysis that is shown in the correlation matrix figure 5.14-B.

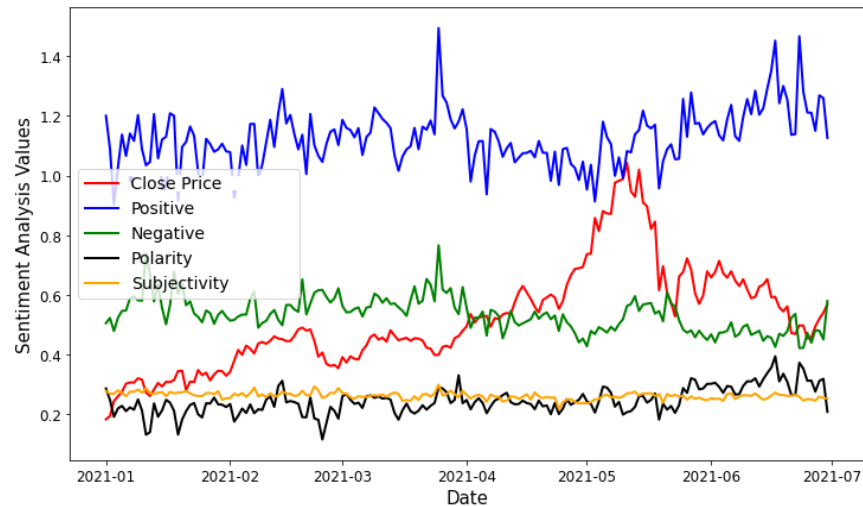


Figure 5.14-A. ETH Harvard IV-4 Sentiment analysis output parameters.

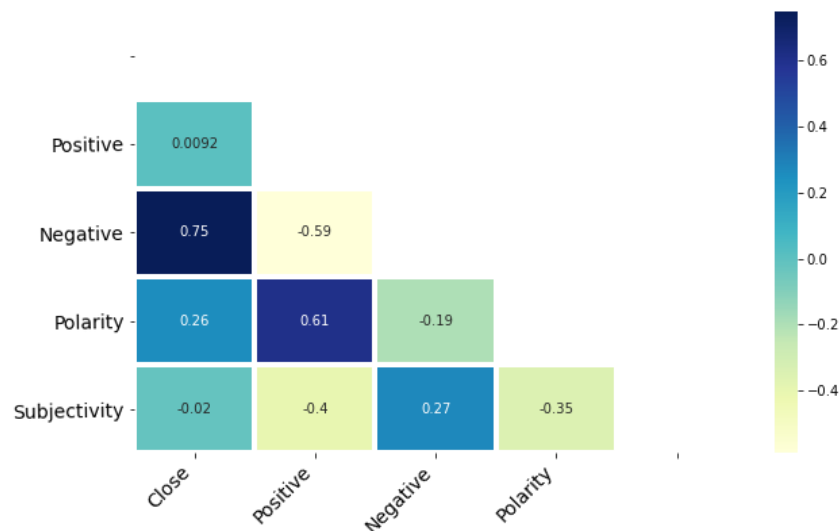


Figure 5.14-B. ETH Harvard IV-4 Sentiment analysis output parameters correlation matrix.

For LTC the output of the Harvard IV-4 result shown in Figure 5.15-B shows that there is a strong positive relationship between polarity and positive analysis, and between the closing price and

polarity, on the other hand, the same strong relation between positive and closing price. The other relations between variables are weak. Harvard IV-4 sentiment analysis is shown in the correlation matrix figure 5.15-B.

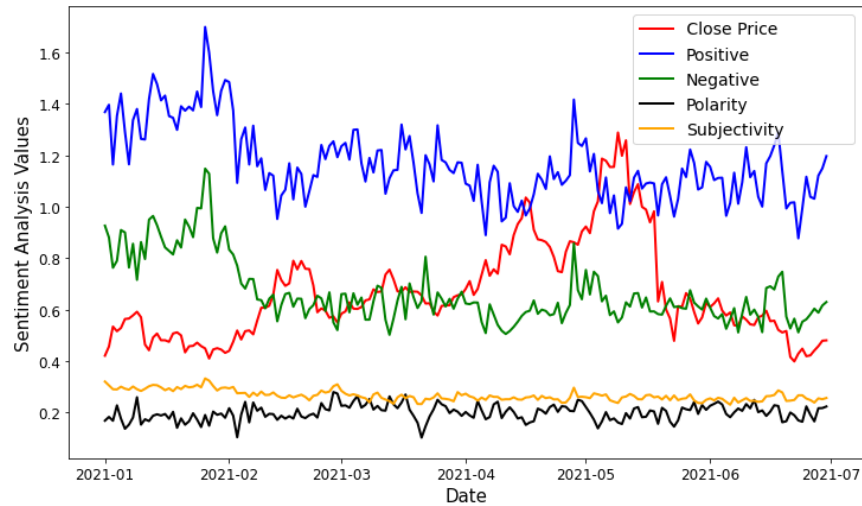


Figure 5.15-A. LTC Harvard IV-4 Sentiment analysis output parameters.

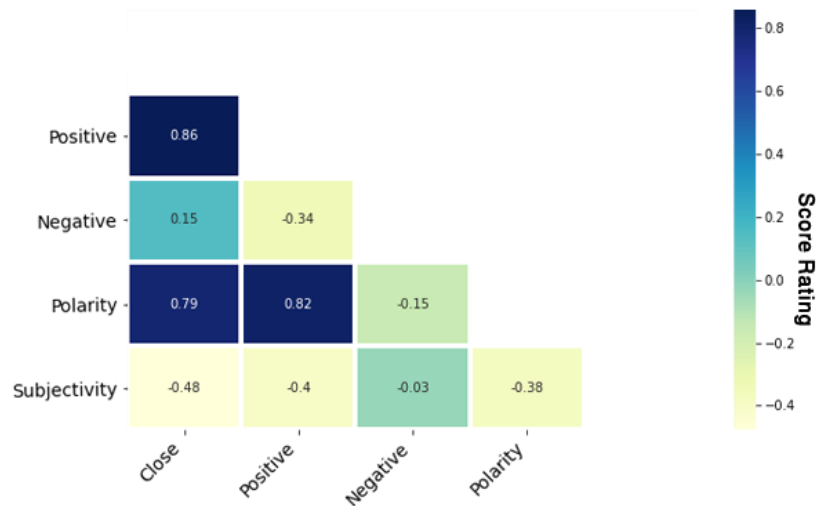


Figure 5.15-B. LTC Harvard IV-4 Sentiment analysis output parameters correlation matrix.

For DOGE the output of the Harvard IV-4 result shown in Figure 5.17-B shows that there is a strong positive relationship between positive and closing price and moderate relation between the

closing price and polarity and between positive and polarity, the other relations are weak. Harvard IV-4 sentiment analysis is shown in the correlation matrix figure 5.15-B.

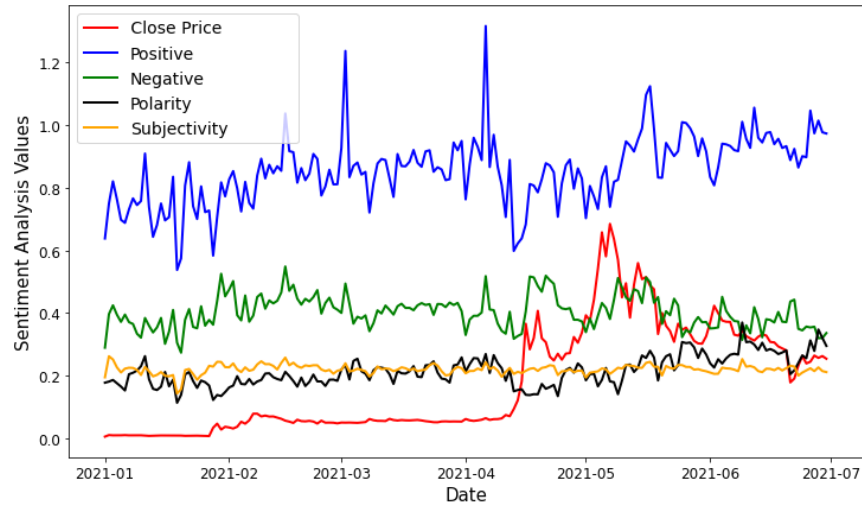


Figure 5.16-A. DOGE Harvard IV-4 Sentiment analysis output parameters.

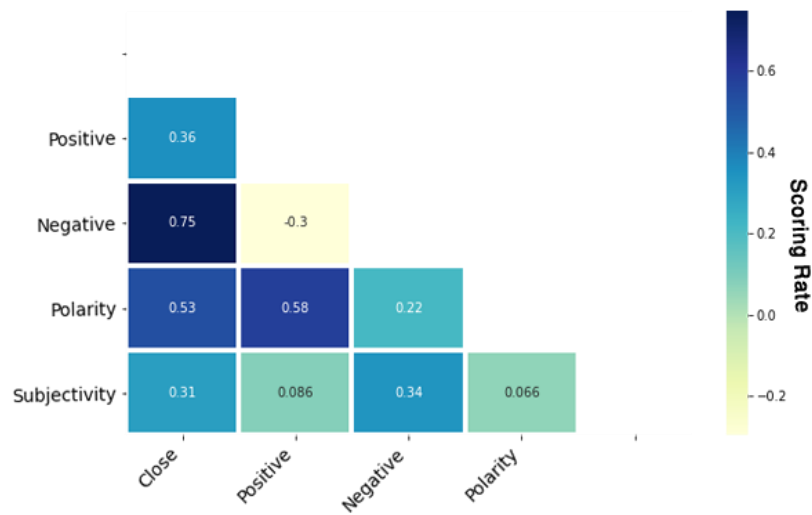


Figure 5.16-B. DOGE Harvard IV-4 Sentiment analysis output parameters correlation matrix.

5.6 Price prediction based on news agency data

Predicting fluctuations in cryptocurrency transactions using user comments and replies examines people's feelings and analyzes them. Many individuals connect on social networking sites such as Twitter and Facebook, as well as news sites such as Reddit [135]. The proposed model for predicting cryptocurrency depends on news agency data. The first step is data collection using the beautifulsoup python library form [110]. The news has been collected from multiple financial websites (mentioned in chapter four). Figure 5.17 illustrates the proposed model architecture.

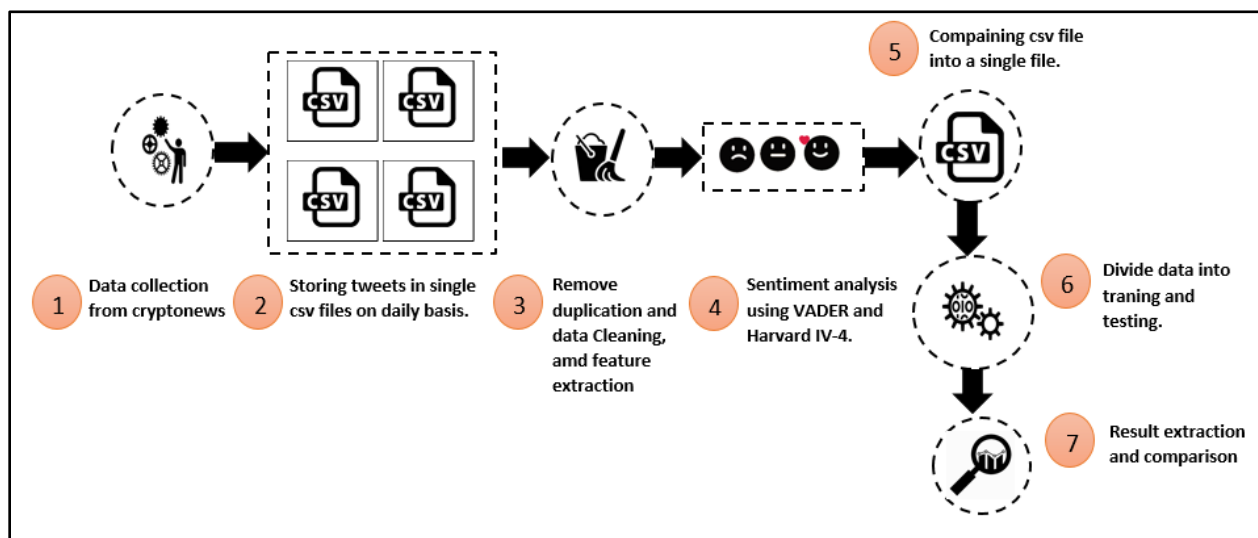


Figure 5.17. The proposed model is to predict cryptocurrency prices using a news channel.

5.6.1 Predicting cryptocurrency using news channel based on VADER sentiment for text analysis

Prediction of cryptocurrency prices using news that comes from reliable sources considers one of the most important channels to achieve the mission, depending on text analysis for news using VADER sentiment analysis, and extract the features like positive, negative, neutral, and compound features, in case of BTC Figure 5.18 show that there is a strong positive correction between BTC closed price and both of positive and neutral VADER output and moderate negative correlation between neutral and negative the other correlation considered as weak correlations.

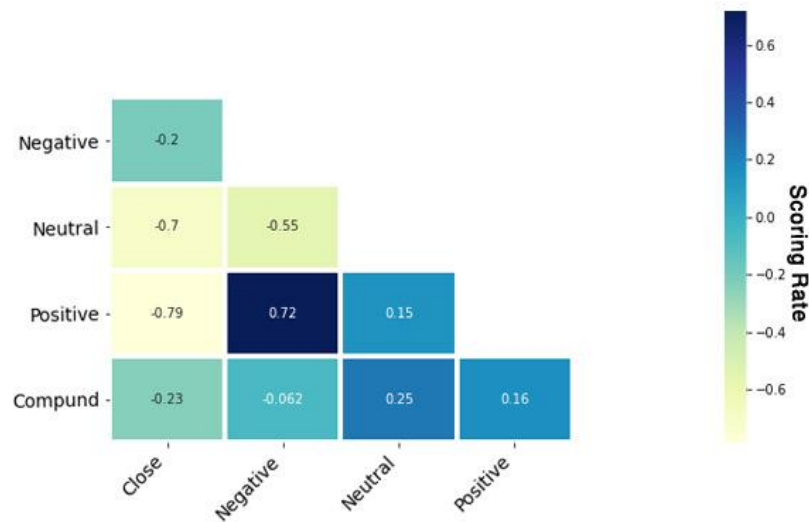


Figure 5.18. BTC VADER Sentiment analysis output parameters correlation matrix for the news channel.

Figure 5.19 illustrates the ETH correlation matrix between the output feature of VADER sentiment analysis it shows that there is a strong positive correlation between positive and negative news and a strong positive between negative and neutral news, for other features the correlation is almost weak.

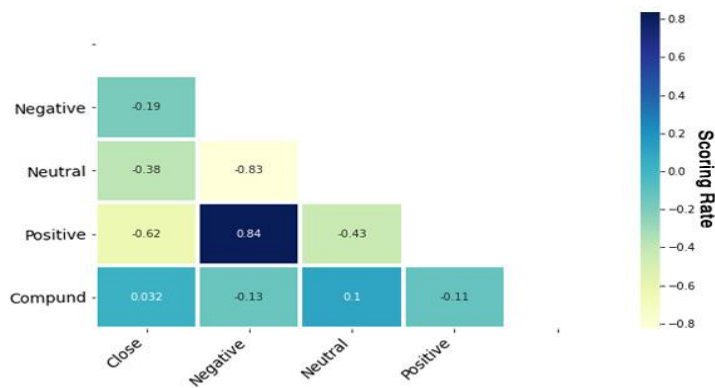


Figure 5.19. ETH VADER Sentiment analysis output parameters correlation matrix for the news channel.

Figure 5.20 illustrates the LTC correlation matrix between the output feature of VADER sentiment analysis it shows that there is a strong positive correlation between positive and negative news, and for other features, the correlation is between moderated and weak correlation.

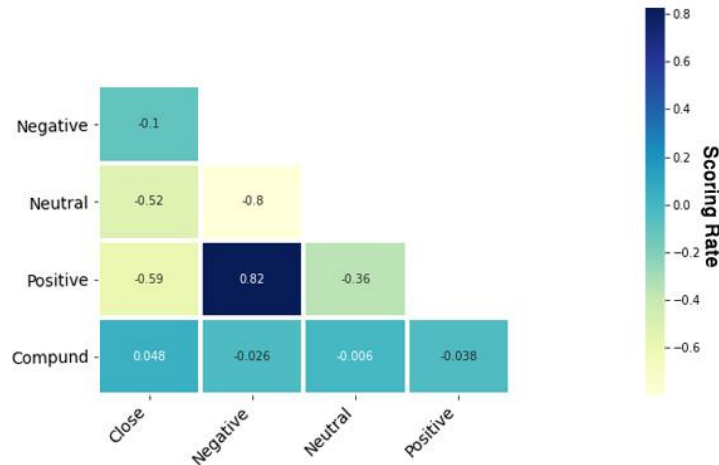


Figure 5.20. LTC VADER Sentiment analysis output parameters correlation matrix for the news channel.

Figure 5.21 illustrates the LTC correlation matrix between the output feature of VADER sentiment analysis it shows that there is a strong positive correlation between positive and negative news and for other features, the correlation is between moderated and weak correlation.

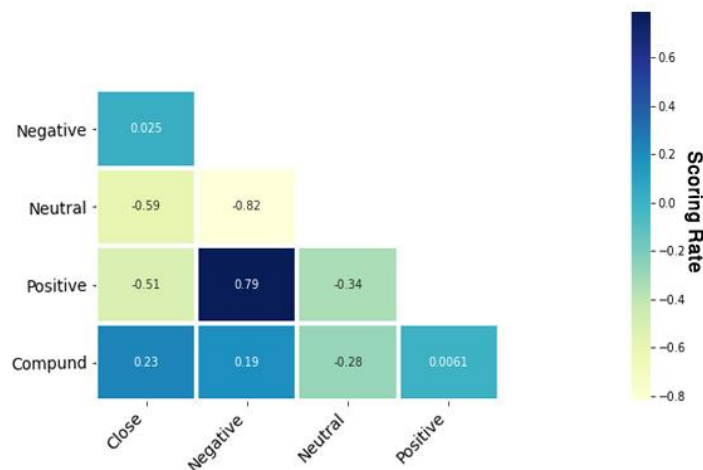


Figure 5.21. DOGE VADER Sentiment analysis output parameters correlation matrix for the news channel.

5.6.2 Predicting cryptocurrency using news channel based von harvard iv-4 dictionary for text analysis

To provide a more context-specific measure of emotion, the Harvard IV-4 vocabulary is made to be used in a range of circumstances. Harvard IV-4 presents positive, negative, subjectivity, and polarity as output to analyze the news text using this type of sentiment. There is a strong positive correlation between positive and polarity and between positive and subjectivity, and a strong negative correlation between polarity and negative for BTC as shown in Figure 5.22.

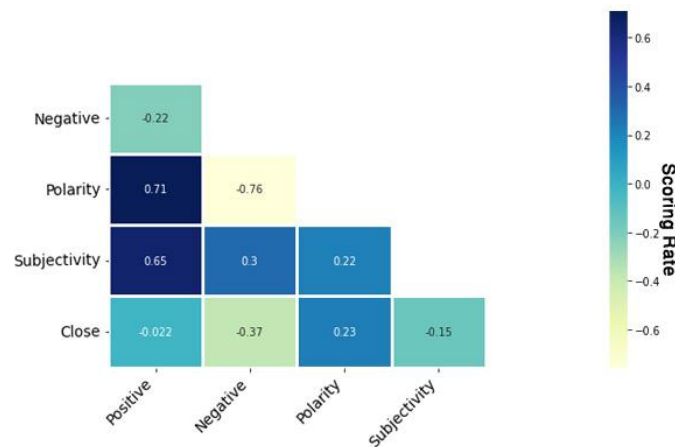


Figure 5.22. BTC Harvard IV-4 Sentiment analysis output parameters correlation matrix for the news channel.

Figure 5.23 illustrate the Harvard IV-4 sentiment analysis for ETH its shows that there is a strong positive correlation between polarity and positive and a strong negative correlation between negative and polarity, and a moderate correlation between subjectivity and positive and between subjectivity and negative, the other correlations are weak.

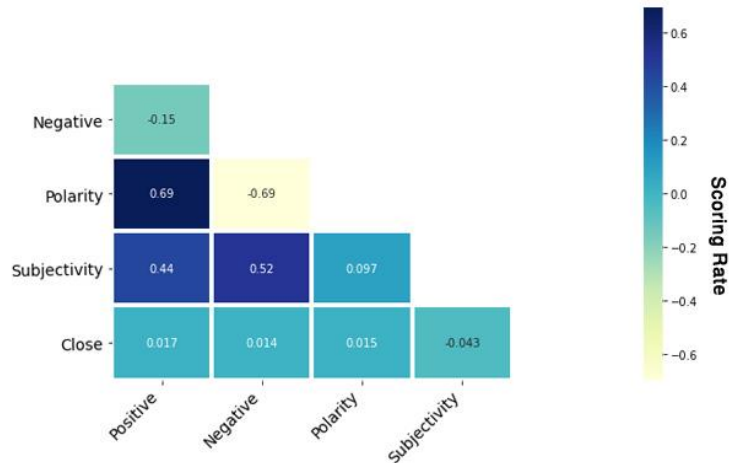


Figure 5.23. ETH Harvard IV-4 Sentiment analysis output parameters correlation matrix for the news channel.

Figure 5.24 illustrate the Harvard IV-4 sentiment analysis for LTC its shows that there is a strong positive correlation between polarity and positive and between positive and subjectivity and there is a strong negative between polarity and negative and there is a moderate correlation between subjectivity and negative, and the other correlation are weak.

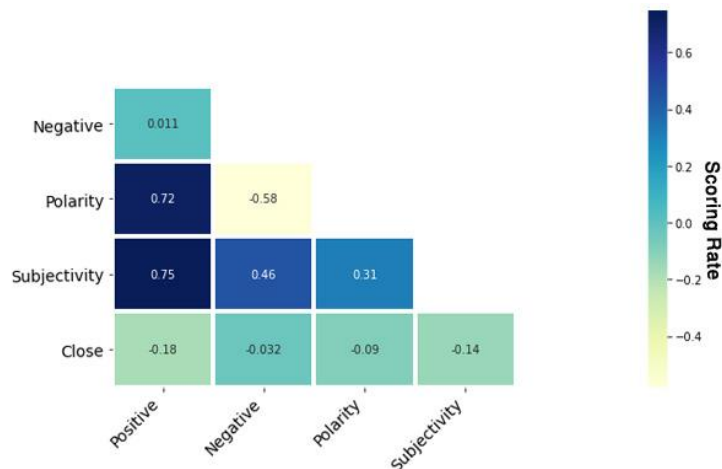


Figure 5.24. LTC Harvard IV-4 Sentiment analysis output parameters correlation matrix for the news channel.

Figure 5.25 illustrate the Harvard IV-4 sentiment analysis for DOGE its shows that there is a strong positive correlation between polarity and positive and between positive and subjectivity and there

is a strong negative between polarity and negative and there is a moderate correlation between subjectivity and negative, and the other correlation are weak.

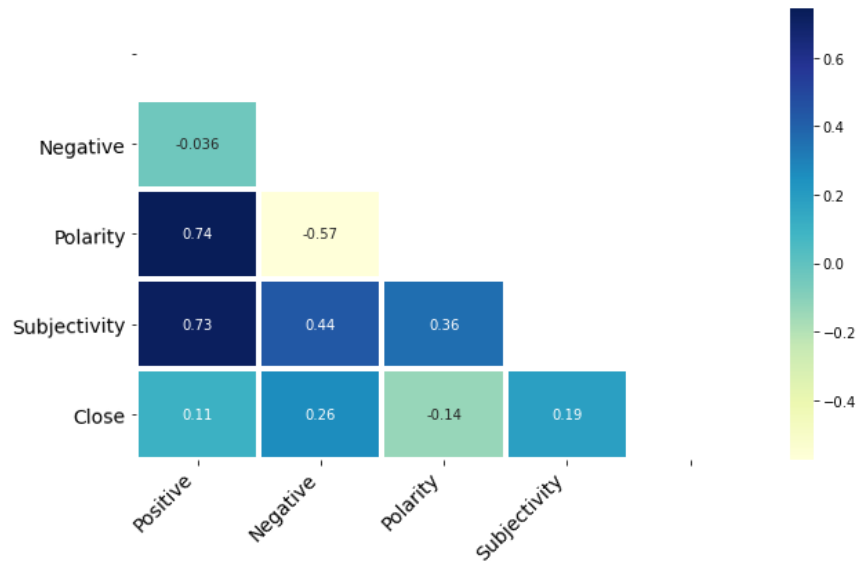


Figure 5.25. DOGE Harvard IV-4 Sentiment analysis output parameters correlation matrix for the news channel.

5.7 Price prediction based on google trends

In this section, I will predict the cryptocurrency price using google trends in general, there is some evidence to suggest that Google Bitcoin Trends may be able to predict Bitcoin prices, though not always over a long period. Also, google trend is a good channel to predict cryptocurrency. Figure 5.26 illustrates the proposed model for the google trends indicator to predict cryptocurrency prices. The prediction process using google trends constant of scene systematic process by starting from google trends data and cryptocurrency price data and storing the result in CSV files, the second step is data visualization, in this process I show the data distribution over the targeted interval and the distribution by countries, the third step is to visualize the collected data and to take a general overview of the data, and the fourth step is to combine the collected data; then the fifth step is to

divide the data into training 75% and testing data 25%. The last two steps are to fit the machine learning model using the LSTM algorithm and finally to extract and plot the results.

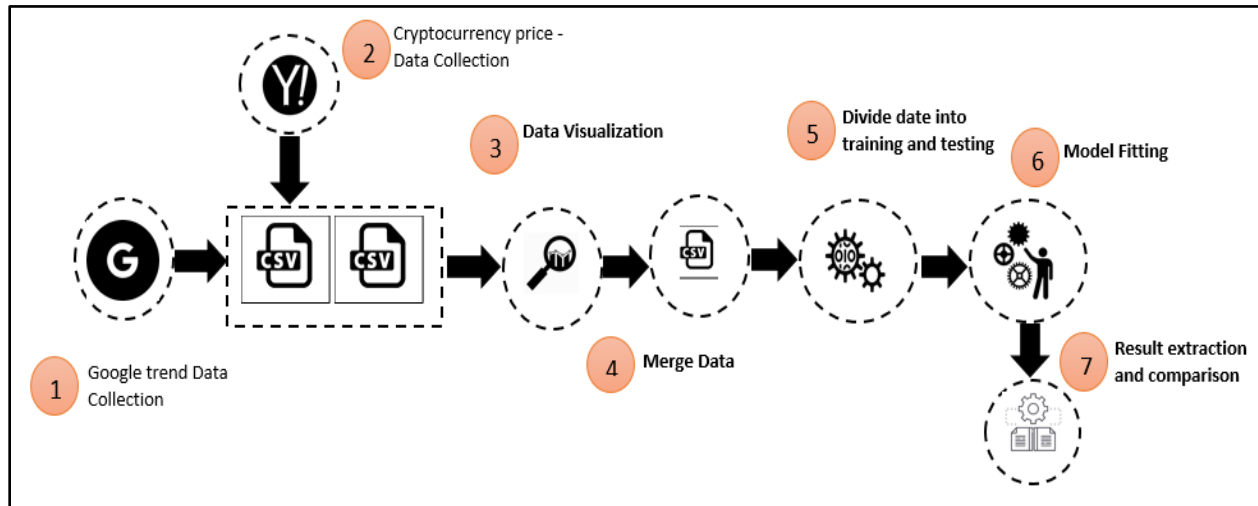


Figure 5.26 Cryptocurrency price prediction using google trend.

In the case of BTC, there is a weak correlation between BTC closing price and Google trend data shown in Figure 5.30-A, on the other hand, the then mean value of google is about 179 with high fluctuation, especially at the end of May 2021 shown in Figure 5.30-B. Also, google trend values are weak negative correlation for BTC as shown in Figure 5.27-A, and a strong positive correlation for ETH, LTC, and DOGE as shown in Figures 5.28-5.30.

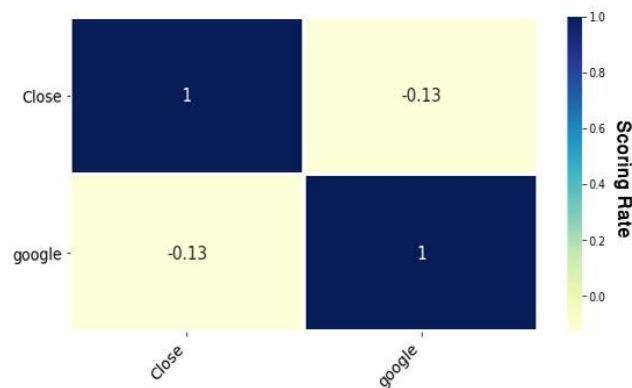


Figure 5.27-A. BTC Google trend correlation matrix with BTC price.

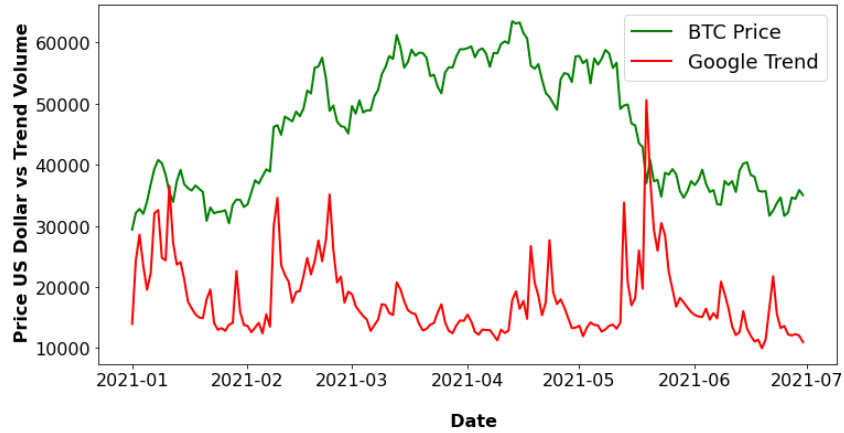


Figure 5.27-B. BTC Google trend vs Price time series.

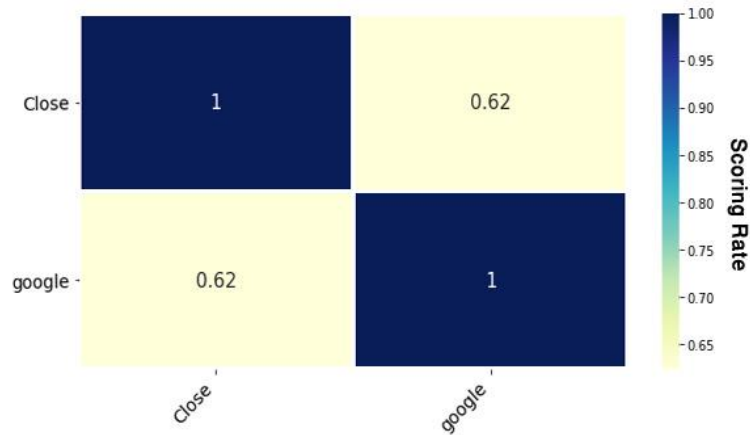


Figure 5.28-A. ETH Google trend correlation matrix with ETH price.

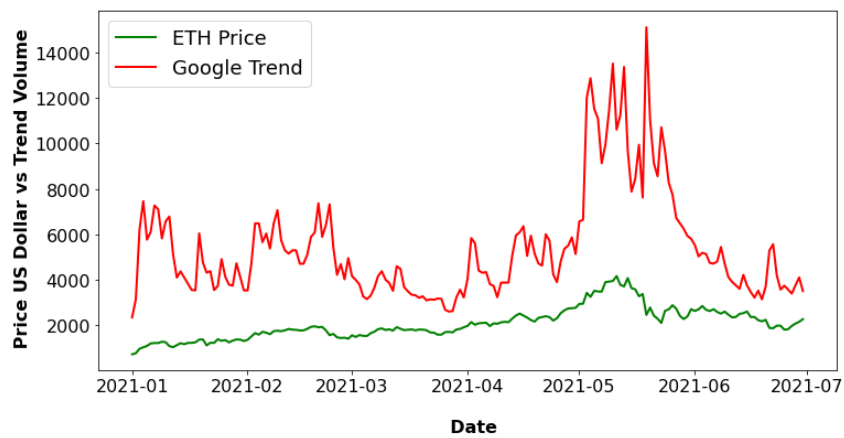


Figure 5.28-B. ETH Google trend vs Price time series.

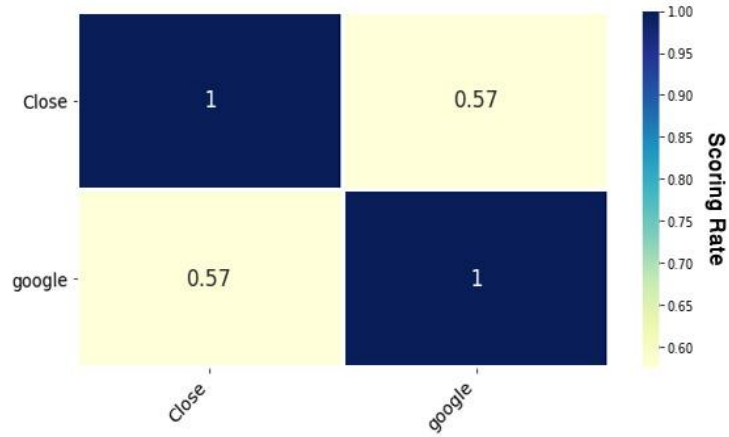


Figure 5.29-A. LTC Google trend correlation matrix with LTC price.

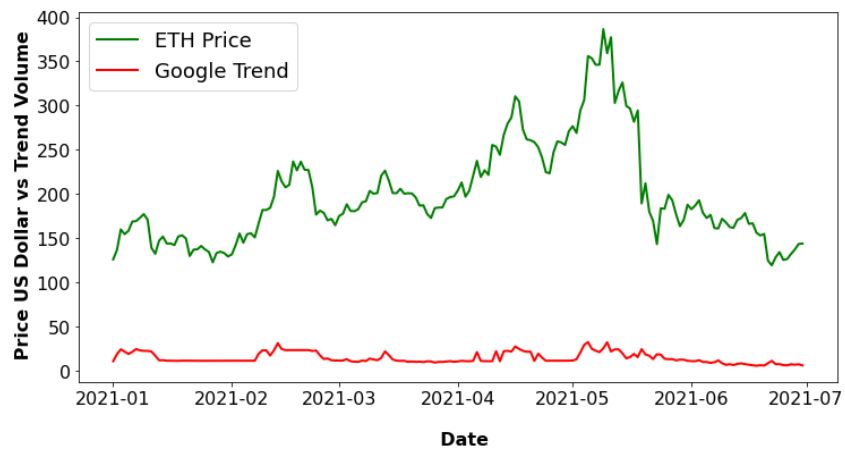


Figure 5.29-B. LTC Google trend vs Price time series.

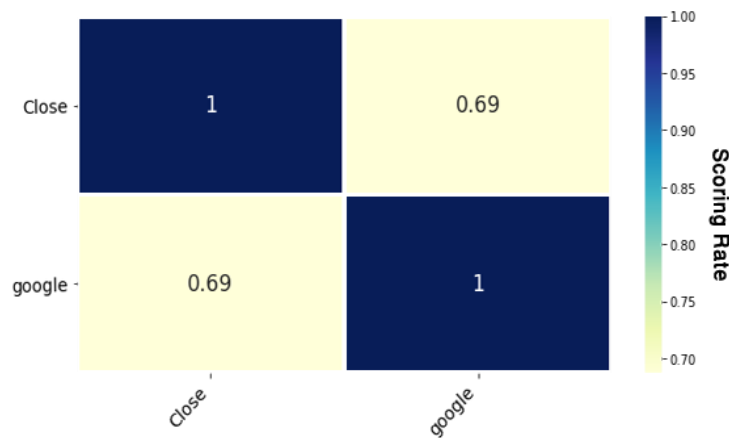


Figure 5.30-A. DOGE Google trend correlation matrix with DOGE price.

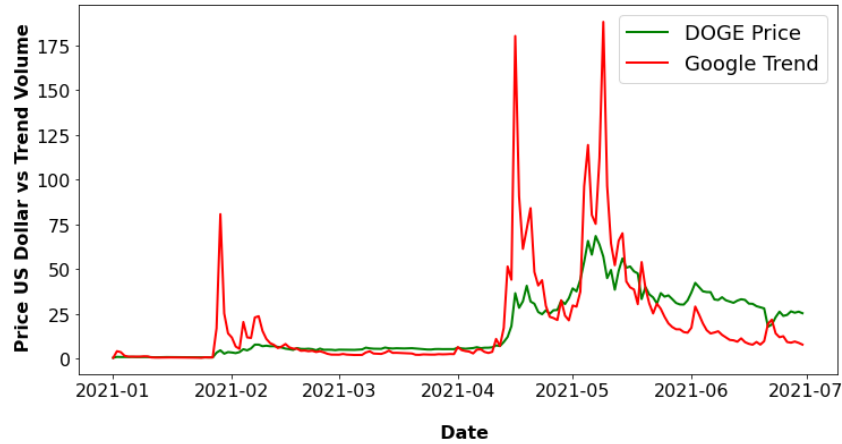


Figure 5.30-B. DOGE Google trend vs Price time series.

5.8 Price prediction based on data fusion

Data fusion is a technique for data collection from multiple data sources to produce consistent and meaningful data that provide a better overview than a single source and also enables us to extract the best features from these sources to formulate a comprehensive dataset to be fed to the machine learning model. In this research, data fusion is used to select the best feature from the multiple data sources to study all internal and external factors that affected cryptocurrency prices to formulate a reliable model for this purpose.

In some scenarios, a single source gives us a good result in direction of accuracy but it's very important to take into account the other factors that affect cryptocurrency prices, because of this if we depend on one source while analyzing the market may shock or influence comes from another source, the best feature from a specific source is like taking the best hereditary traits, as the result of this producing a human with the best features that taking from parents. Figure 5.31 illustrates the proposed model for data fusion. The green arrows represent the features that are shared in the fusion process, and the red arrows represent the excluded features. There are 20 features collected

from multiple data sources that represent the best features, from the Twitter source we applied two techniques for processing the text that is VADER sentiment analysis and Harvard IV-4. The experiment that applied for predicting cryptocurrency depending on Twitter only shows that the VADER output gives a better result than Harvard IV-4, so we take the output from VADER for both Twitter and news sources to be fusion with the other sources.

Mining in cryptocurrency is the procedure used to create new cryptocurrency coins, utilizing a computer to validate the following block on the blockchain is known as bitcoin mining. The ability of cryptocurrencies to function as it does is due to the decentralized network of miners. In addition to any fees paid by the parties exchanging the tokens to the miners, the protocol generates a reward in the form of cryptocurrency tokens as payment.

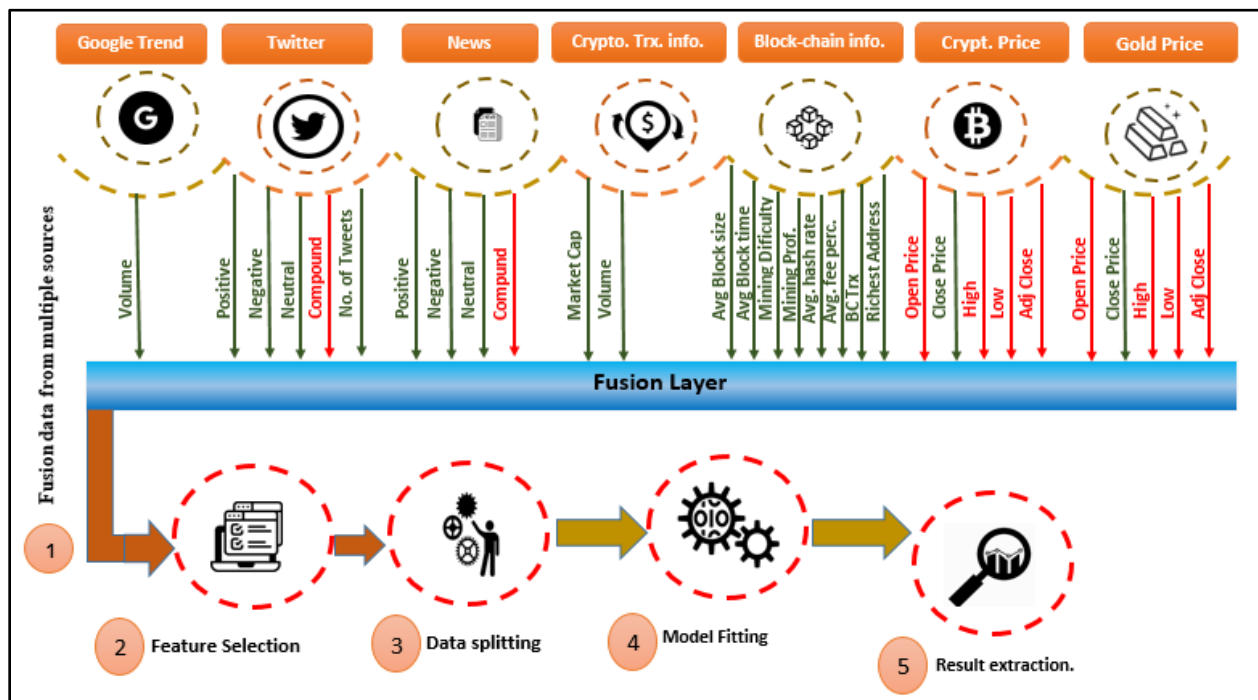


Figure 5.31. Data fusion model to predict cryptocurrency price prediction.

Figures 5.32 – 5.35 illustrate the correlation matrix for fusion data for LTC, BTC, ETH, and DOGE respectively. In general, for some features that are highly correlated in which it is recommended to avoid having correlated features in the dataset for achieving good models. Indeed, A collection of strongly linked features won't provide much new information (if any), but they will complicate the algorithm and raise the possibility of mistakes. Although it relies on the details of the problem, such as the number of variables and the degree of correlation, the correlated features generally don't improve models, they have varied effects on different models and to different degrees.

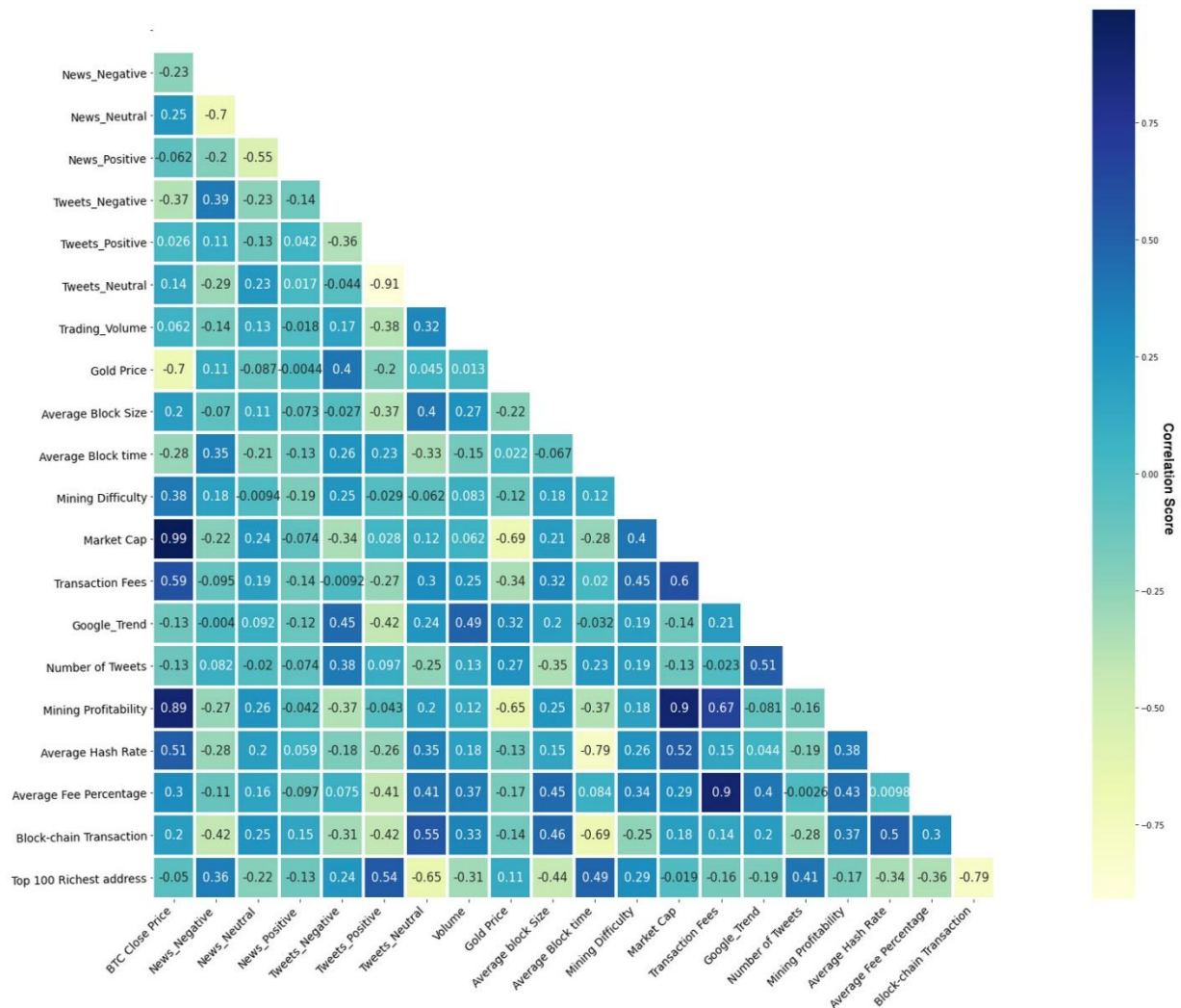


Figure 5.32. Data fusion correlation matrix for BTC.

Figure 5.32 illustrates the fusion result for BTC, there is some feature with a high correlation with each other like average hash rate and mining difficulty which means that the average hash rate could be predicted using mining difficulty and some other highly correlated features. The same for the transaction fee and the average transaction fee.

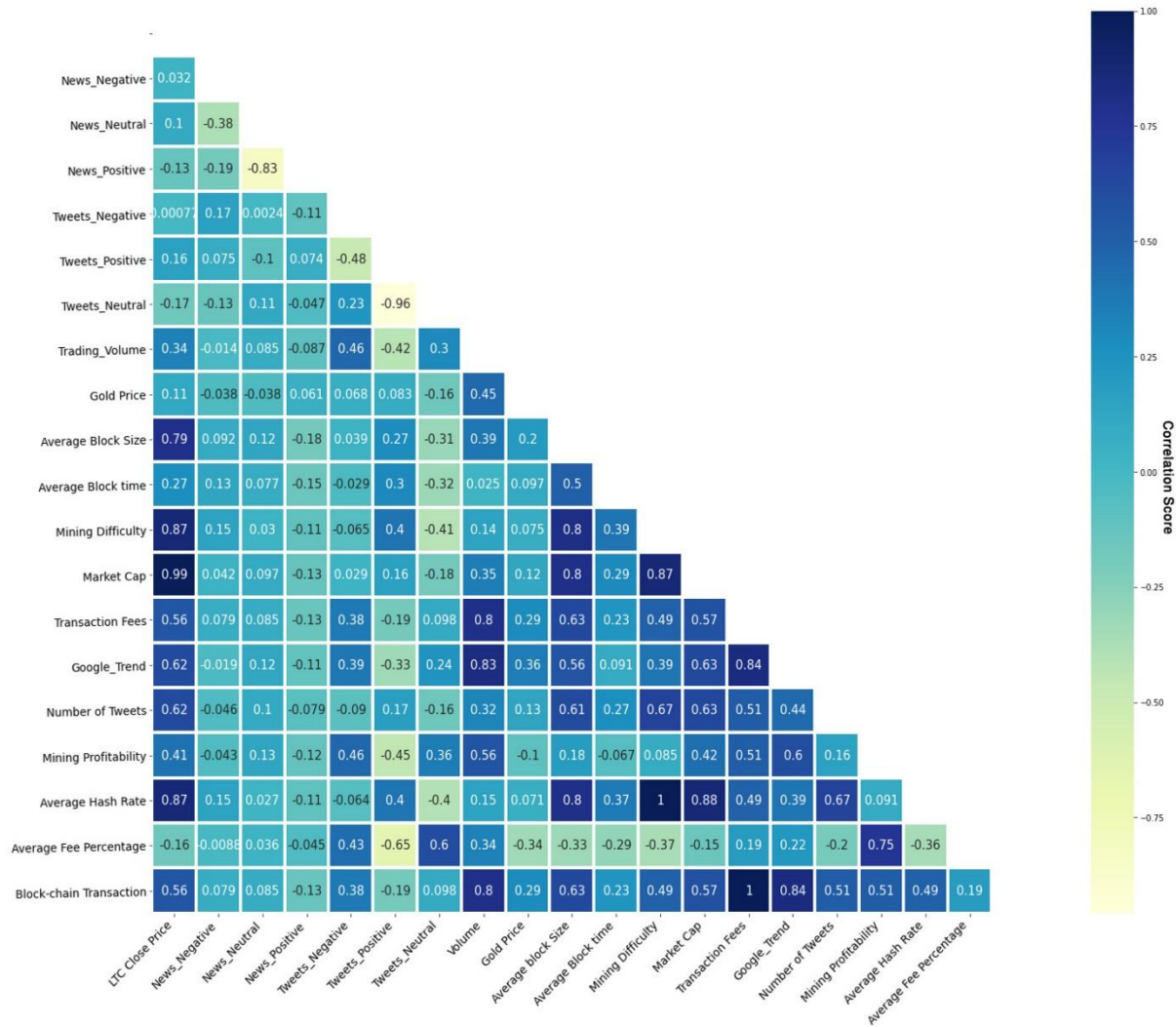


Figure 5.33. Data fusion correlation matrix for ETH.

Figure 5.33 illustrates the fusion result for ETH, there is some feature with a high correlation with each other like market cap and mining profitability which means that the market cap could be

predicted using mining profitability and some other highly correlated features. The same for the transaction fee and the average transaction fee.

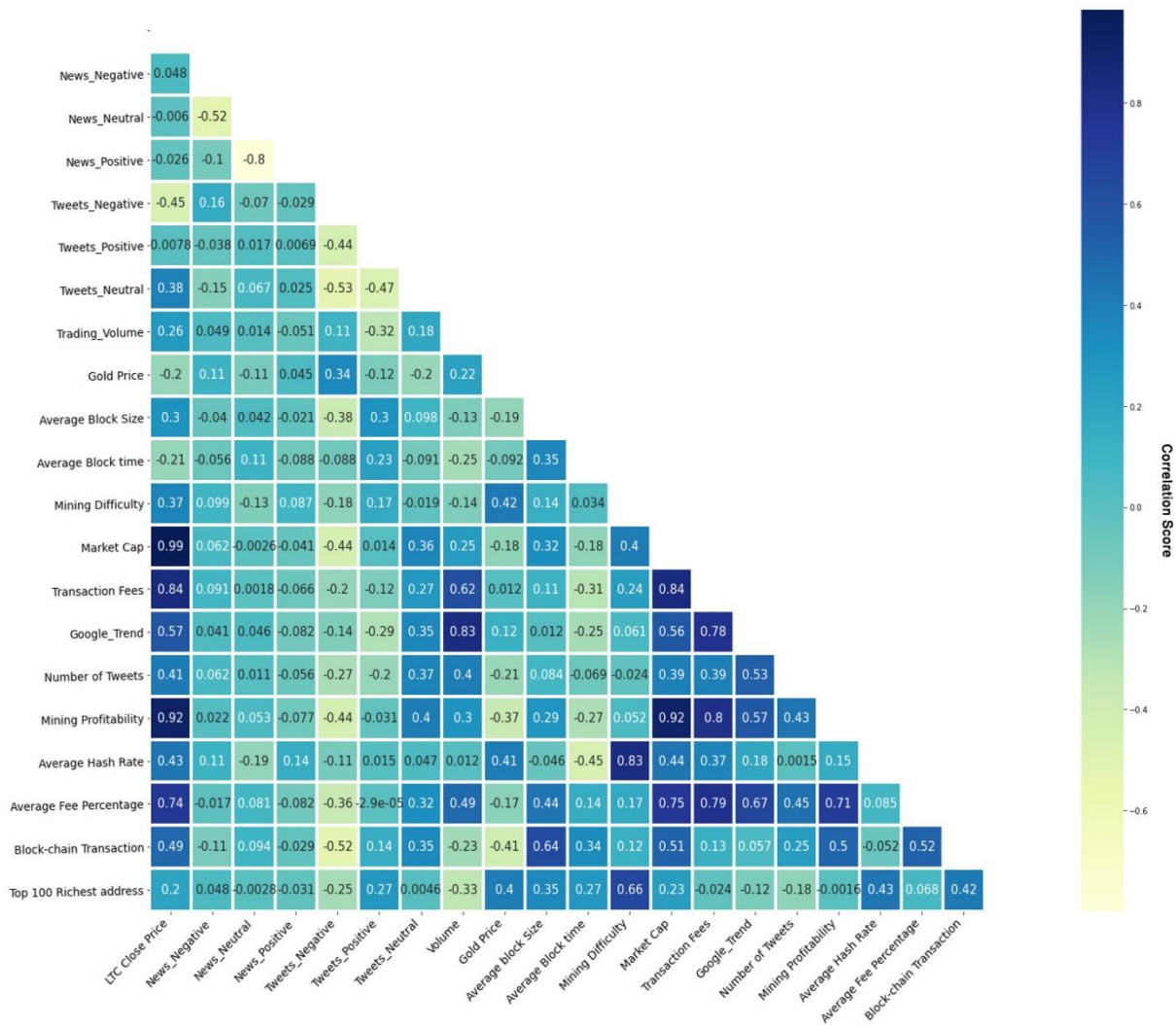


Figure 5.34. Data fusion correlation matrix for LTC.

Figure 5.35 illustrates the fusion result for DOGE, there is some feature with a high correlation with each other like market capitalization and mining profitability which means that the market capitalization could be predicted using mining profitability and some other highly correlated features. The same for the transaction fee and the average transaction fee.

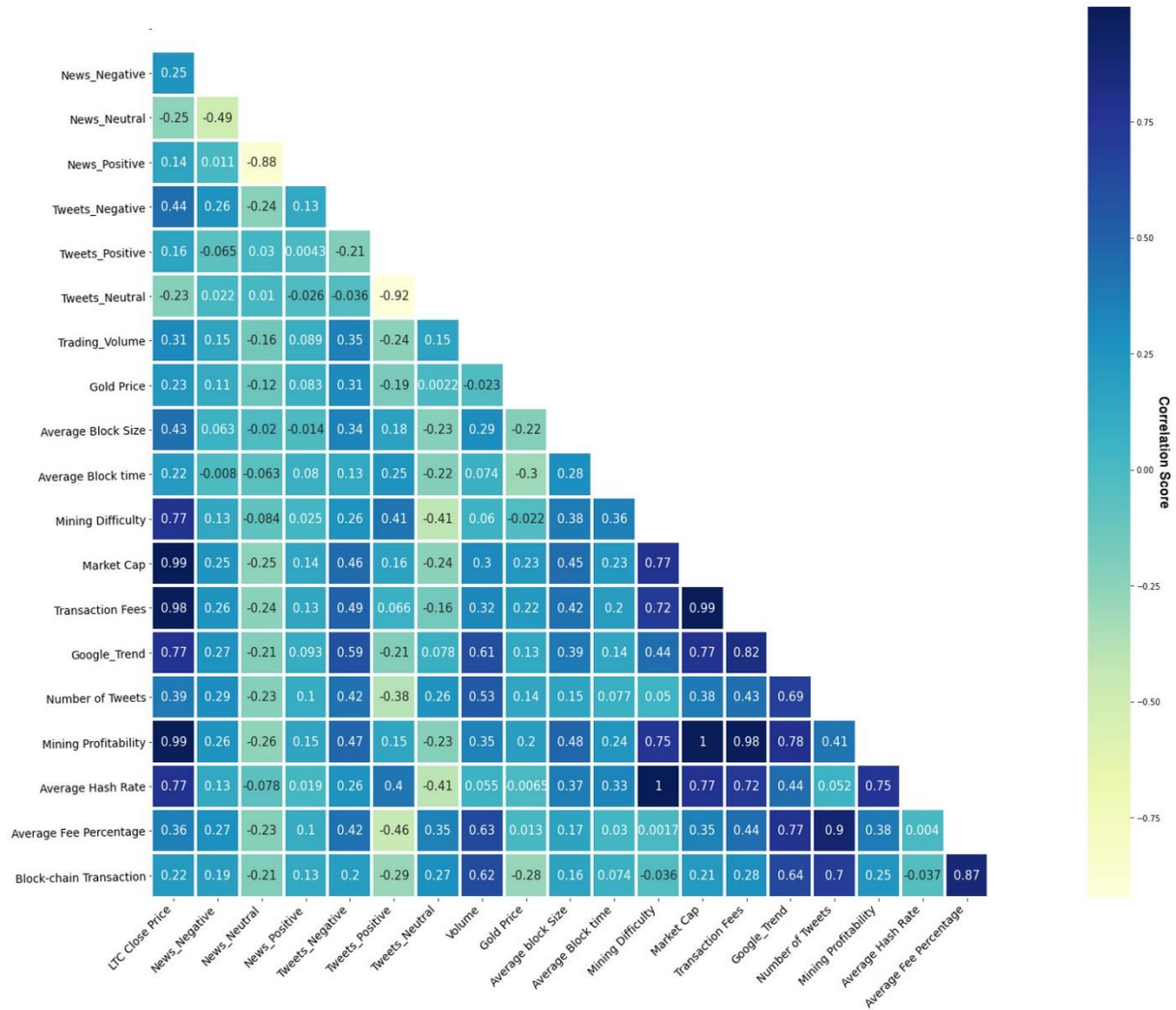


Figure 5.35. Data fusion correlation matrix for DOGE.

5.9 Conclusion.

In this chapter I implement multiple models from multiple channels, may the result for a single channel is good but from the broader concept, we have to study all influences and factors that effected to cryptocurrency price. To achieve this goal, I proposed data fusion techniques to study all factors that effected in a cryptocurrency price. The results have been optimized for all targeted cryptocurrencies, but LTC represents the best results. In the next chapter, I will discuss the result for every single channel and compare them with the fusion model.

Chapter Six

Results and evaluation

6.1 Introduction

In this chapter, I will represent a deep discussion about the extracted result from multiple models that I created to predict cryptocurrency prices for the targeted cryptocurrencies those are BTC, ETH, LTC, and, DOGE within the targeted interval from 1 JAN 2021 and 30 Jun 2021. MAPE and RMSA were used to evaluate the proposed model's performance.

To achieve the thesis goal proposed models have been created in two phases. The first phase is by creating multiple models using a single channel and comparing the result to discover what is the best channel, in phase two the model was formulated using the data from multiple channels to create a machine learning model and then comparing the results between phase one and phase two.

Figure 6.1 illustrates the result comparison and extraction criteria.

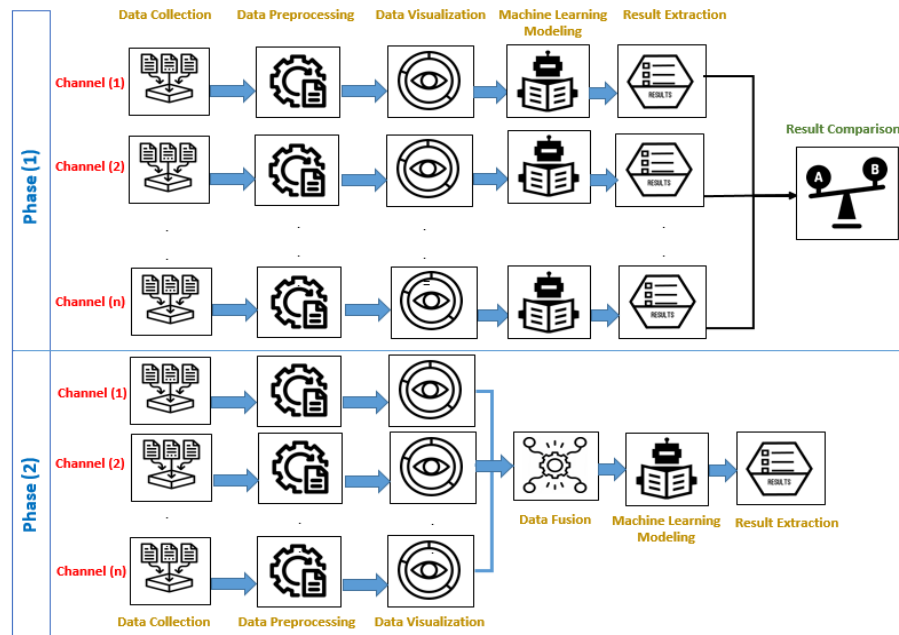


Figure 6.1. The Result Extraction criteria.

LSTM model has been created by training the model with 75% of data and 25% for testing and depends on the following parameters:

single input layer: 8 size.
 Optimizer: Adam.
 Activation: rmse.
 Learning rate: 0.001.
 Dropout: 0.2

6.2 cryptocurrency price prediction using social media (twitter)

By using Twitter as a channel to predict cryptocurrency prices there are many methods and techniques used to find the best model, most robust, and most reliable for the targeted cryptocurrencies with the targeted interval, the findings show that Twitter has a predictive power that is considered as a reliable with taking into account the effected of data preprocessing for the text of tweets. In this phase, I created multiple models using English and non-English tweets the non-English presented better results than using only English tweets.

6.2.1 Prediction depends on English tweets

Because the importance of social especially Twitter is also considered one of the most important channels to predict cryptocurrency prices. Tweets are text which needs to be filtered and analyzed, to perform this, we proposed two types of sentiment analysis techniques that are VADER sentiment analysis and Harvard IV-4, and use the output parameters as input for the machine learning. The output from the VADER sentiment analysis is positive, negative, neutral, and compound, on the other hand, the output of Harvard IV-4 is positive, negative, subjectivity, and polarity. To get the best, we apply feature selection techniques for this purpose. Table 6.1

illustrates briefly the result for four targeted cryptocurrencies using English Tweets by using VADER sentiment and Harvard IV-4 for testing the dataset.

Table 6.1 Predictions of cryptocurrency prices by Tweets channel

	VADER Sentiment Analysis		Harvard IV-4	
Crypto	MAPE	RMSE	MAPE	RMSE
DOGE	0.021%	0.03	0.03%	0.050
LTC	5.29%	173.95	18.36%	713.11
ETH	6.41%	426.78	10.96%	695.36
BTC	11.37%	954.49	18.02%	1469.63

Both techniques present a good result in predicting DOGE cryptocurrency the results are very close, on the other hand, the result became better if we used positive and negative a feature in both techniques. Figures 6.2 – 6.5 illustrate the result of using the Twitter channel.

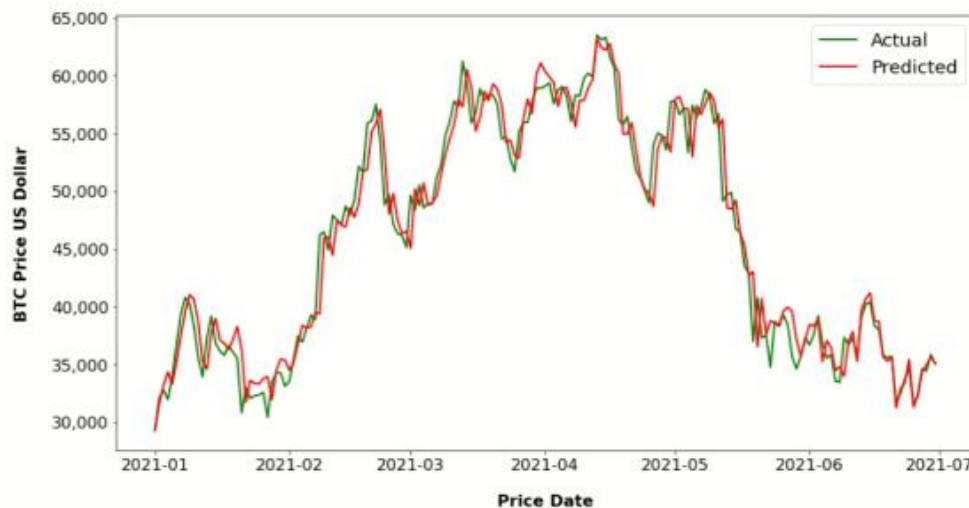


Figure 6.2. BTC price prediction using Twitter and VADER Sentiment analysis.

For BTC using the Twitter channel and VADER sentiment analysis. Figure 6.2 shows that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual is very close, but there

is a clear difference between the predicted and actual at the beginning and end of the targeted interval especially at the middle of May until the middle of Jun. The outcome shows that the value of MAPE is 11.73% and the value of RMSE is 954.49.

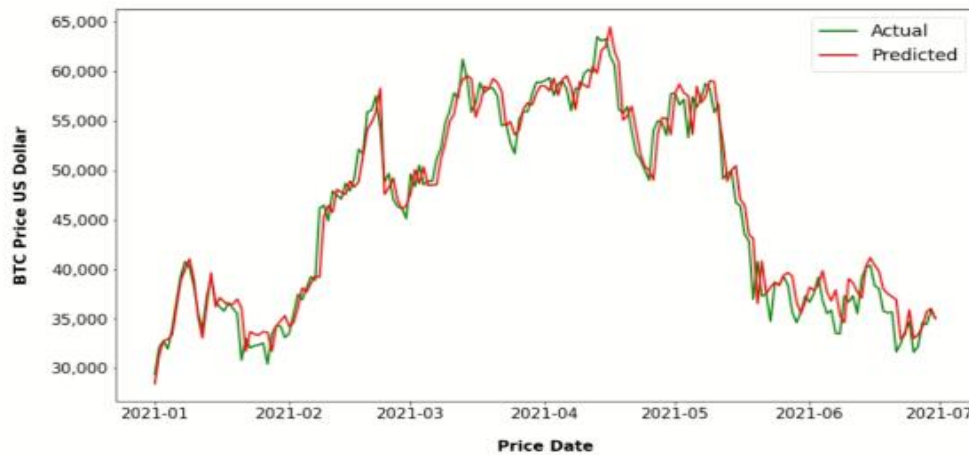


Figure 6.3. BTC price prediction using Twitter and HARVARD IV-4 Sentiment analysis.

For BTC using the Twitter channel and Harvard IV-4 sentiment analysis. Figure 6.3 shows that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual is very close, but there is a clear difference between the predicted and actual at the beginning and end of the targeted interval especially at the middle of May until the end of Jun. Using VADER sentiment analysis shows a better result than using Harvard IV-4. The outcome shows that the value of MAPE is 18.02% and the value of RMSE is 1469.63.

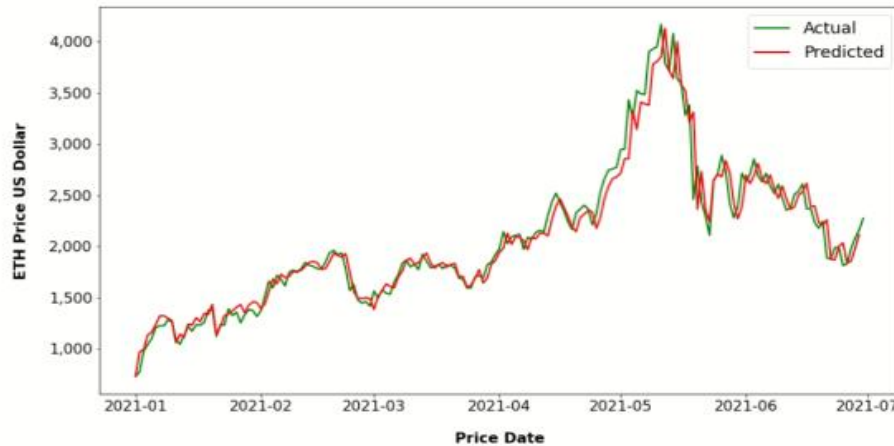


Figure 6.4. ETH price prediction using Twitter and VADER Sentiment analysis.

For ETH, by using the Twitter channel and VADER sentiment analysis. Figure 6.4 shows that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual is very close, but there is a clear difference between the predicted and actual at the beginning of sudden rise at the end of April 2021. The outcome shows that the value of MAPE is 6.41% and the value of RMSE is 426.78.

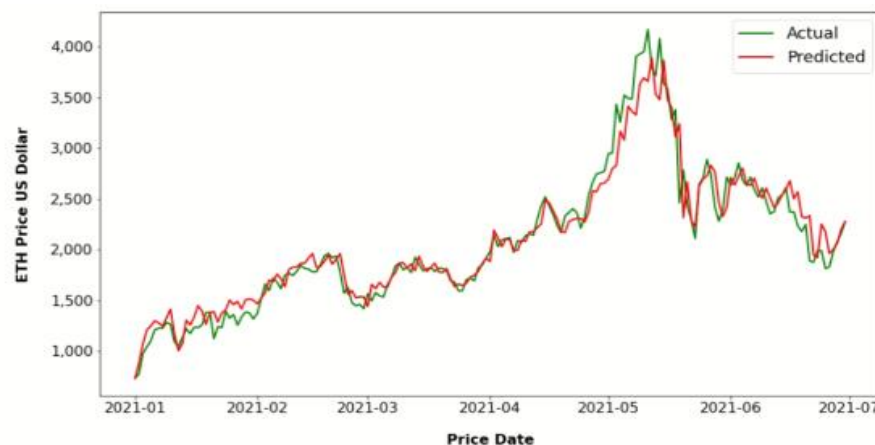


Figure 6.5. ETH price prediction using Twitter and Harvard Sentiment analysis.

For ETH, by using the Twitter channel and Harvard IV-4 sentiment analysis. Figure 6.5 shows that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual is very close, but there is a clear difference between the predicted and actual at the beginning of sudden rise at the end of April 2021, and the end of the period starting at the end of July 2021. The outcome shows that the value of MAPE is 10.96% and the value of RMSE is 695.36.

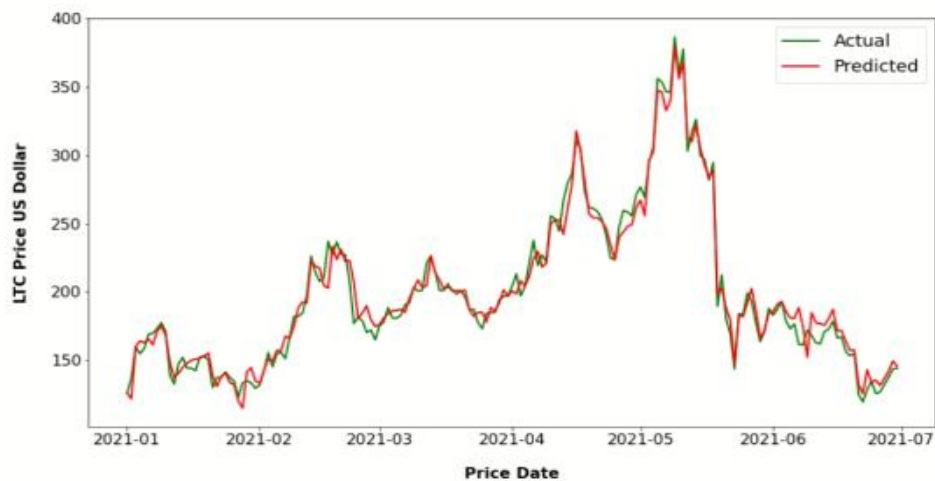


Figure 6.6. LTC price prediction using Twitter and VADER Sentiment analysis.

For LTC, by using the Twitter channel and VADER sentiment analysis. Figure 6.6 shows that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual is very close, but there is a clear difference between the predicted and actual at the beginning Jun 2021 until the end of July 2021. The outcome shows that the value of MAPE is 5.29% and the value of RMSE is 173.95.

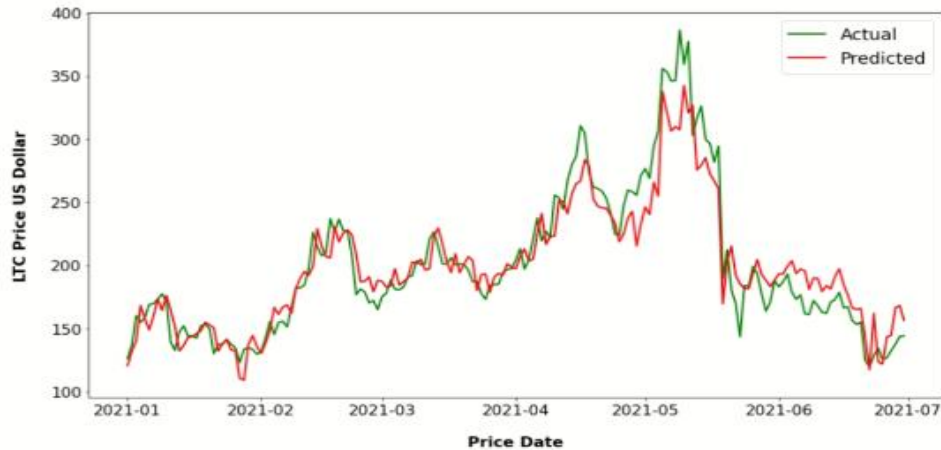


Figure 6.7. LTC price prediction using Twitter and Harvard IV-4 Sentiment analysis.

For LTC, by using the Twitter channel and Harvard IV-4 sentiment analysis. Figure 6.7 shows that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual is unstable over the targeted period, but there is a clear difference between the predicted and actual at the beginning May 2021 until the end of July 2021. The outcome shows that the value of MAPE is 18.36% and the value of RMSE is 713.11. Using VADER sentiment analysis for LTC gives better results than using VADER with a difference of 13.07% in RMSE.

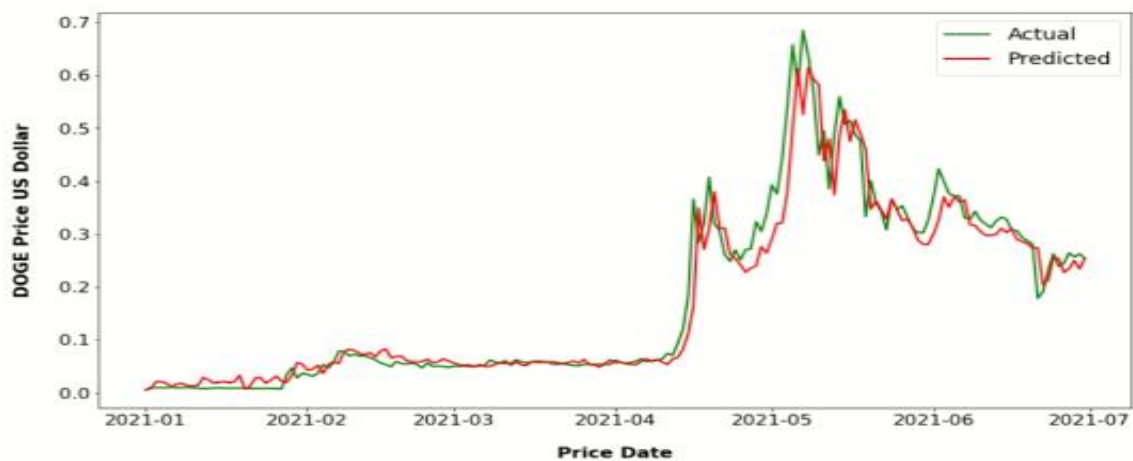


Figure 6.8. DOGE price prediction using Twitter and VADER Sentiment analysis.

For DOGE, by using the Twitter channel and VADER sentiment analysis. Figure 6.8 shows that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was in the sudden increase in the middle of April 2021. The outcome shows that the value of MAPE is 0.021% and the value of RMSE is 0.03.

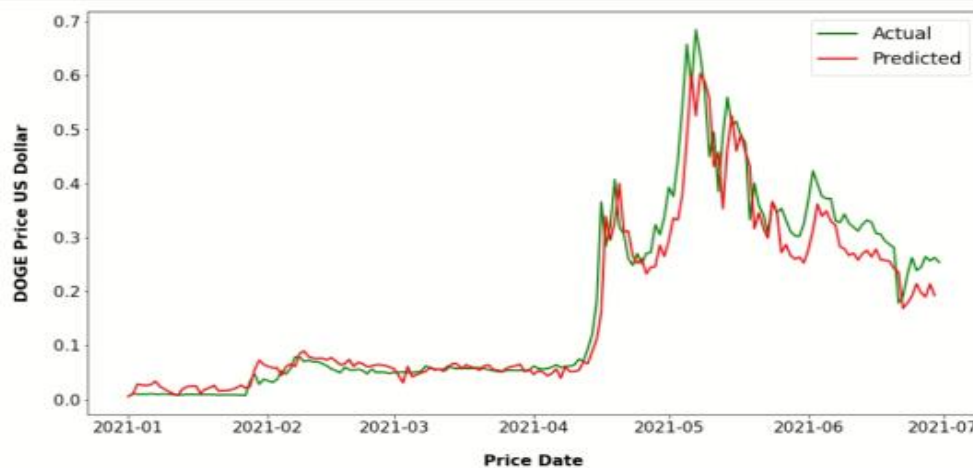


Figure 6.9. DOGE price prediction using Twitter and Harvard IV-4 Sentiment analysis.

For DOGE, by using the Twitter channel and Harvard IV-4 sentiment analysis. Figure 6.9 shows that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was in the sudden increase from the middle of April 2021 until the end of the period. The outcome shows that the value of MAPE is 0.03% and the value of RMSE is 0.05.

Using VADER sentiment analysis gets a better result than using Harvard IV-4, the result is much close. The Figure shows a high difference, but the RMSE result is very close together, this is because the price of DOGE is very low.

6.2.2 Prediction depends on non-English tweets

Previous research depends on English tweets only because there are no packages interacting with multi-language to solve this issue. Therefore, a translation layer is proposed in this research to convert the language to English text. The translation process is very difficult and takes a long time. However, the research contributes to enhancing the result of MAPE. The translation is applied to LTC using VADER sentiment analysis, the outcomes show that RMSE is 196.05 and MAPE is 5.01%, Figure 6.10 illustrates the result of LTC price prediction using English and non-English tweets.

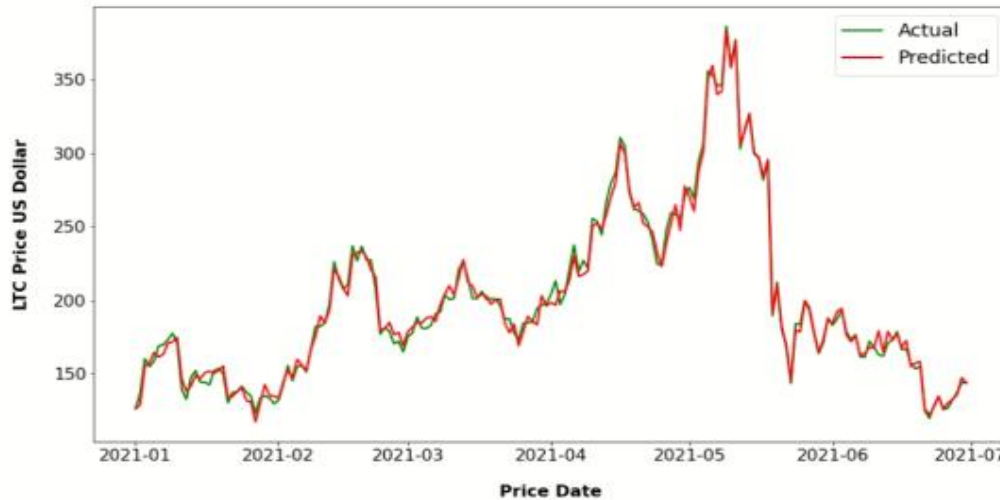


Figure 6.10. LTC price prediction using depends on English and Non-English Tweets.

The translation process that we proposed is to convert the non-English Tweets into English Tweets, this process takes a long time especially when interacting with huge data, For LTC, by using the Twitter channel and VADER sentiment analysis. Figure 6.10 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual is very close over the targeted period with the RMSE being 196.05 and MAPE being 5.01%.

6.3 Cryptocurrency price prediction using news agencies

In this section, the influence of news on cryptocurrency prices is investigated to assist cryptocurrency players in making more informed decisions regarding the sale or purchase of cryptocurrency, we developed a model to predict whether a specific news article will result in an apparent cryptocurrency price increase or decrease using a variety of basic machine learning algorithms.

The experiment applied using two techniques to analyze the news articles text that is VADER sentiment analysis and Harvard IV-4, both two techniques produce good and satisfactory results.

Table 6.2 illustrates the result of using news articles to predict cryptocurrency prices. For DOGE both two proposed techniques are reliable and very close to reality with a value of 0.02% for MAPE. On the other hand, Harvard IV-4 gets better results for text analysis than VADER for the other currencies with a value of 16.44%, 10.58%, and 7.33% for BTC, ETH, and LTC respectively. Figures 6.10 - 6.15 represents news text analysis result using both VADER and Harvard IV-4.

Table 6.2 Predictions of cryptocurrency prices by news articles channel

	VADER Sentiment Analysis		Harvard IV-4	
Crypto	MAPE	RMSE	MAPE	RMSE
BTC	17.95%	1831.95	16.44%	1732.43
ETH	12.25%	146.64	10.58%	142.05
LTC	8.36%	21.66	7.33%	20.22
DOGE	0.02%	0.02	0.03%	0.02

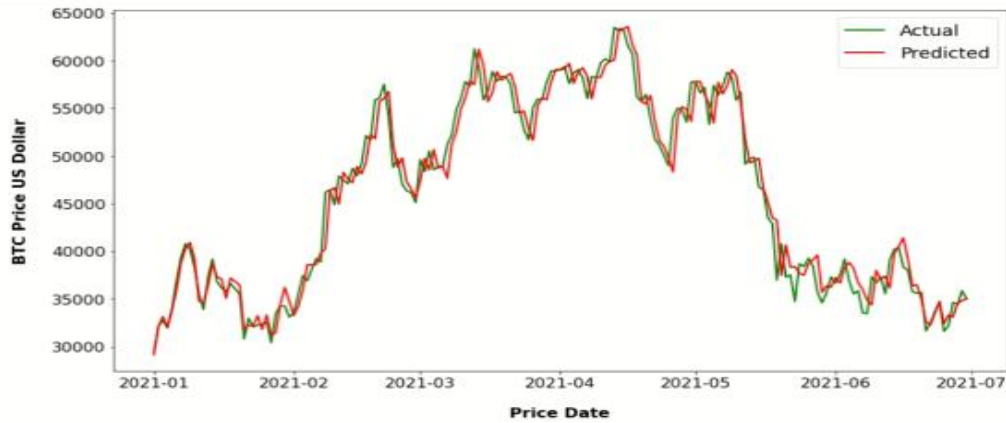


Figure 6.11. BTC price prediction using News channels VADER sentiment analysis.

For BTC, by using the news channel and VADER sentiment analysis. Figure 6.11 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was in the middle of May 2021. The outcome shows that the value of MAPE is 17.95% and the value of RMSE is 1831.95.

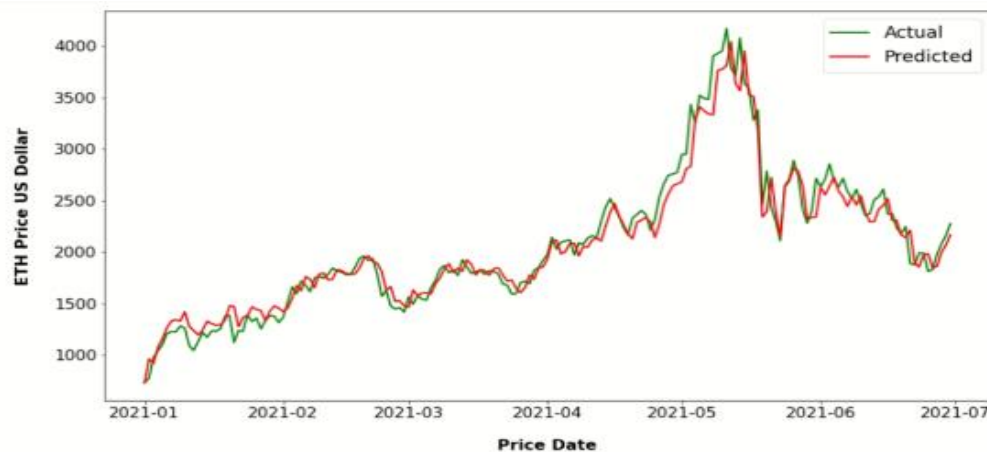


Figure 6.12. ETH price prediction using News channel - VADER sentiment analysis.

For ETH, by using the news channel and VADER sentiment analysis. Figure 6.12 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference

between the predicted and actual was in the sudden increase in the middle of April. The outcome shows that the value of MAPE is 12.25% and the value of RMSE is 146.64.

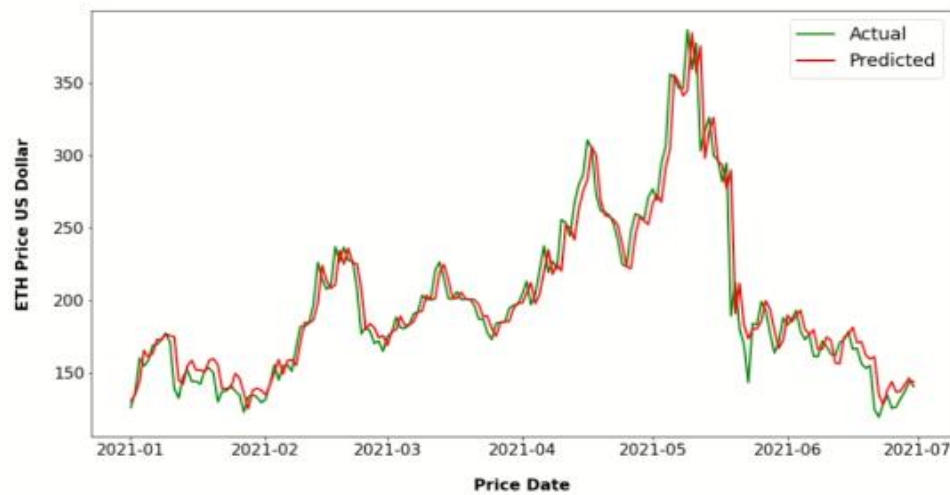


Figure 6.13 LTC price prediction using News channel – VADER sentiment analysis.

For LTC, by using the news channel and VADER sentiment analysis. Figure 6.13 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was at the end of May 2021. The outcome shows that the value of MAPE is 8.36% and the value of RMSE is 21.66.

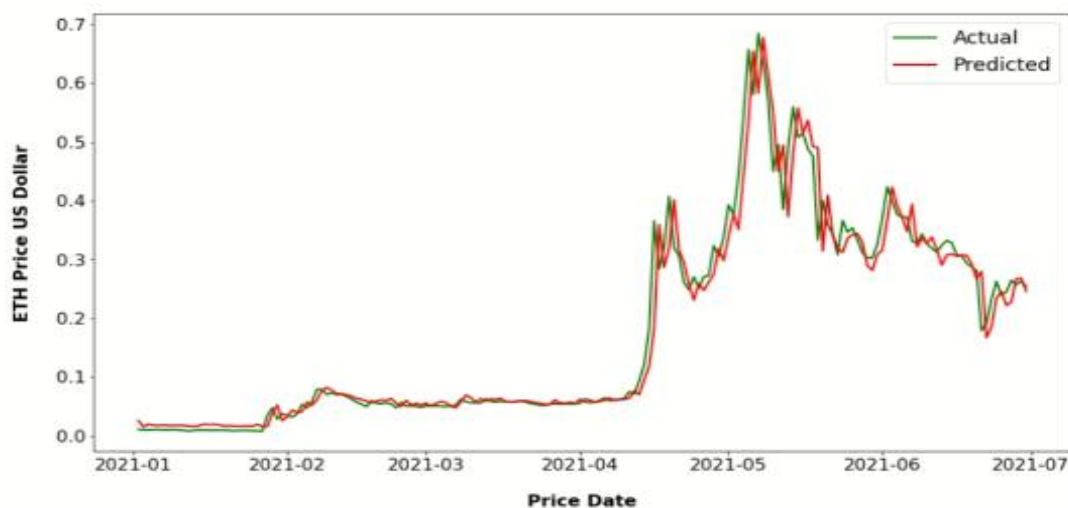


Figure 6.14. DOGE price prediction using News channel – VADER sentiment analysis.

For DOGE, by using the news channel and VADER sentiment analysis. Figure 6.14 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was from the middle of April until the end of the period. The outcome shows that the value of MAPE is 0.02% and the value of RMSE is 0.02.

LTC cryptocurrency represents the best price prediction model using a news agency channel and VADER sentiment analysis with a value of 8.36% for RMSE and 21.66 for MAPE.

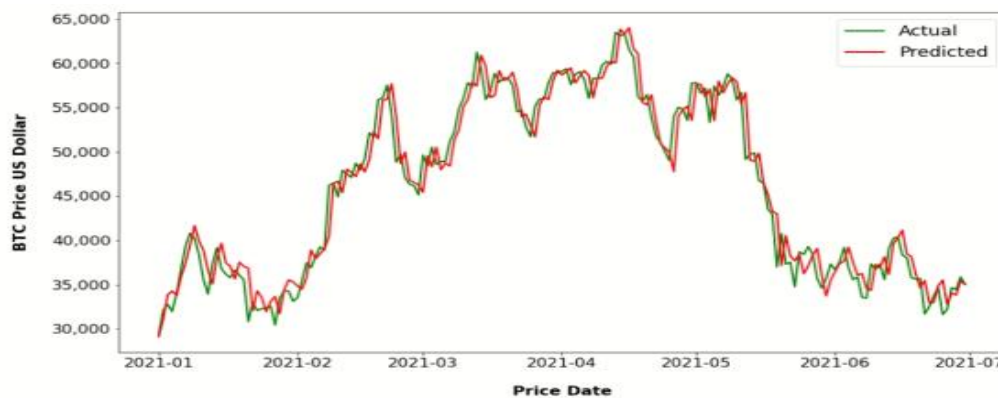


Figure 6.15. BTC price prediction using News channel – Harvard IV-4 sentiment analysis.

For BTC, by using the news channel and Harvard IV-4 sentiment analysis. Figure 6.15 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was in the middle of May 2021. The outcome shows that the value of MAPE is 16.44% and the value of RMSE is 1732.43.

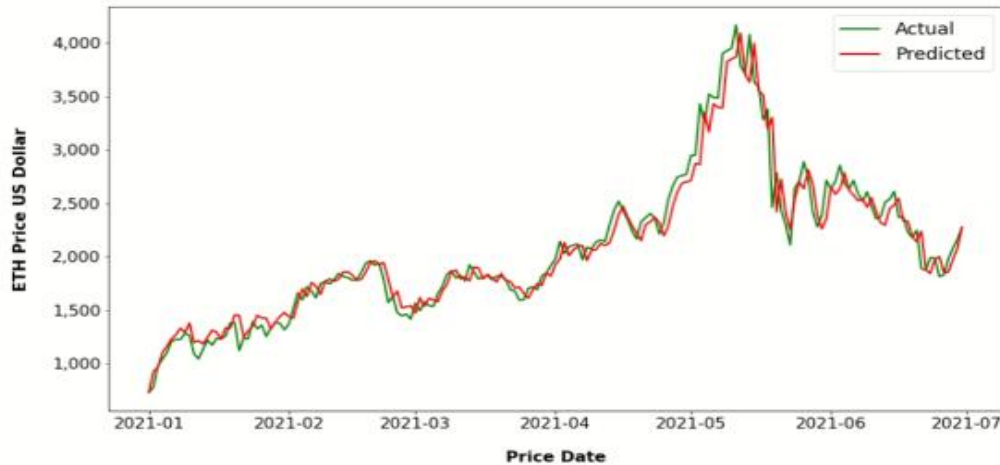


Figure 6.16. ETH price prediction using News channel – Harvard IV-4 sentiment analysis.

For ETH, by using the news channel and VADER sentiment analysis. Figure 6.16 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was in the sudden increase in the middle of April. The outcome shows that the value of MAPE is 10.58% and the value of RMSE is 142.05.

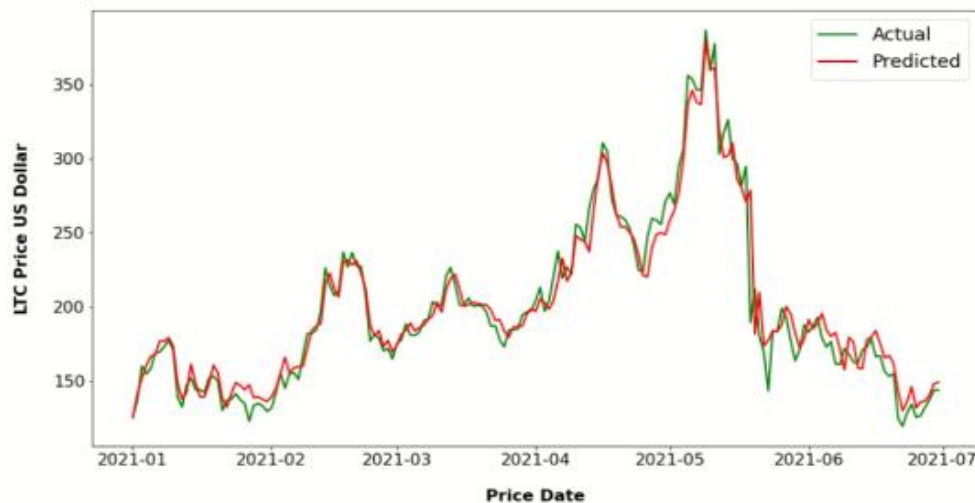


Figure 6.17. LTC price prediction using News channel – Harvard IV-4 sentiment analysis.

For LTC, by using the news channel and VADER sentiment analysis. Figure 6.17 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was at the end of May 2021. The outcome shows that the value of MAPE is 7.33% and the value of RMSE is 20.22.

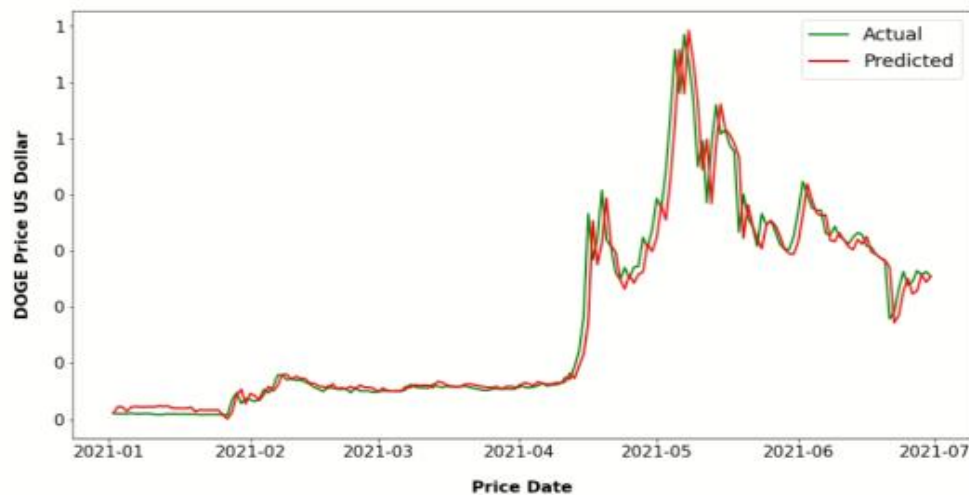


Figure 6.18. DOGE price prediction using News channel – Harvard IV-4 sentiment analysis.

For DOGE, by using the news channel and VADER sentiment analysis. Figure 6.18 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was from the middle of April until the end of the period when the price suddenly increases. The outcome shows that the value of MAPE is 0.03% and the value of RMSE is 0.02.

By using news channel to predict the cryptocurrency, the experiment outcomes show that using Harvard IV-4 gets better results than using VADER sentiment for BTC, ETH, and LTC, and using

VADER sentiment analysis gets better results than using Harvard IV-4 for DOGE with small differences between the two methods.

6.4 Cryptocurrency price prediction using google trend

Google provides access to the number of searches conducted worldwide using their search engine through the Google Trends website. To quantify and analyze the distribution of search trends across countries or within countries, the global trends package downloads search volumes from Google Trends.

With the use of global trends, researchers and analysts can look into patterns within these trends, such as the extent to which businesses and organizations are global in scope or the spread of political, social, or technological trends internationally or inside specific nations. Table 6.3 illustrates that cryptocurrency price prediction depends on the google trends channel.

Results show good performance for the model in predicting cryptocurrencies using google trends. DOGE coin achieved the best MAPE performance with 0.04 %, then LTC with a value of 8.77% for MAPE, on the other hand, BTC and ETH were the latest two cryptocurrencies with a value of 16.79% and 19.23% respectively. Figures 6.19 – 6.22 illustrate the cryptocurrency price prediction using google trends implemented using LSTM.

Table 6.3 Predictions of cryptocurrency prices by using Google trend based on LSTM

Cryptocurrency	MAPE	RMSE
BTC	19.23%	1588.45
ETH	16.79%	112.28
LTC	8.77 %	26.31
DOGE	0.04%	0.08



Figure 6.19. BTC price prediction using Google trend.

For BTC, by using google trends. Figure 6.19 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was from the end of February 2021 and at the beginning of Jun 2021 middle of April until the end of the period when the price suddenly increases. The outcome shows that the value of MAPE is 19.23% and the value of RMSE is 1588.45.

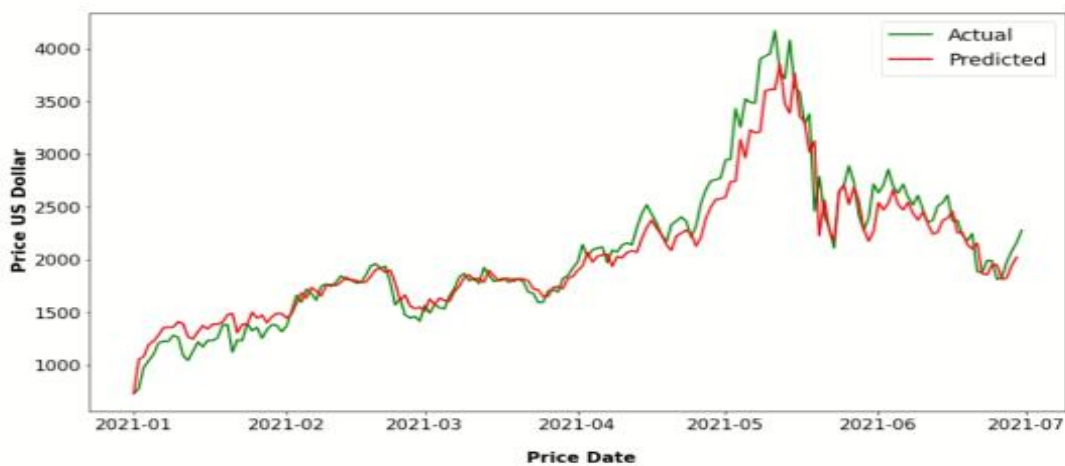


Figure 6.20. ETH price prediction using Google trend.

For ETH, by using google trends. Figure 6.20 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was from the end of February 2021 and at the beginning of Jun 2021 middle of April until the end of the period when the price suddenly increases. The outcome shows that the value of MAPE is 16.79% and the value of RMSE is 112.28.

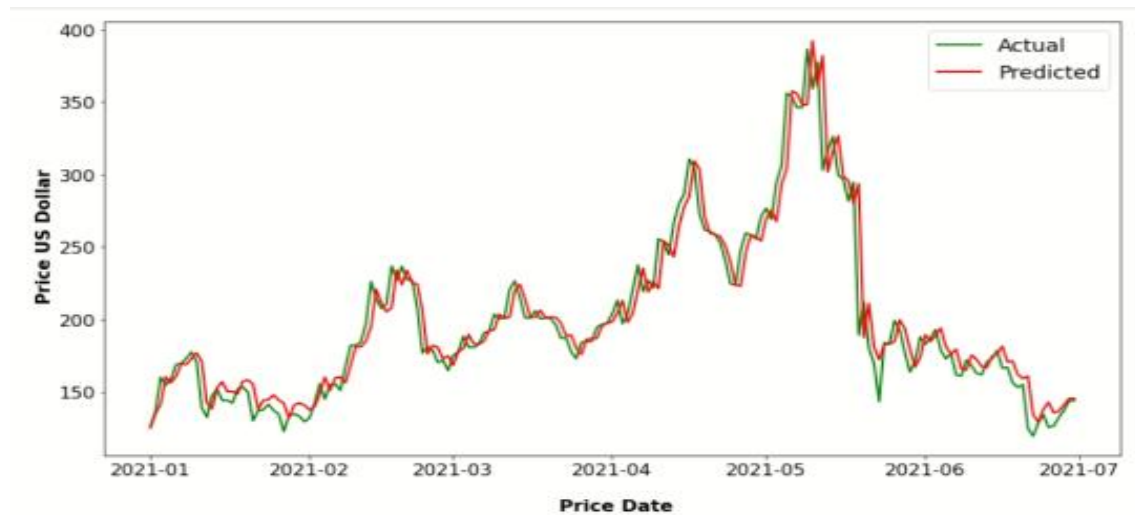


Figure 6.21. LTC price prediction using Google trend.

For LTC, by using google trends. Figure 6.21 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was from the middle of January 2021 until the beginning of February 202m and at the middle of Jun 2021 until the end of July 2021. The outcome shows that the value of MAPE is 8.77 % and the value of RMSE is 26.31.

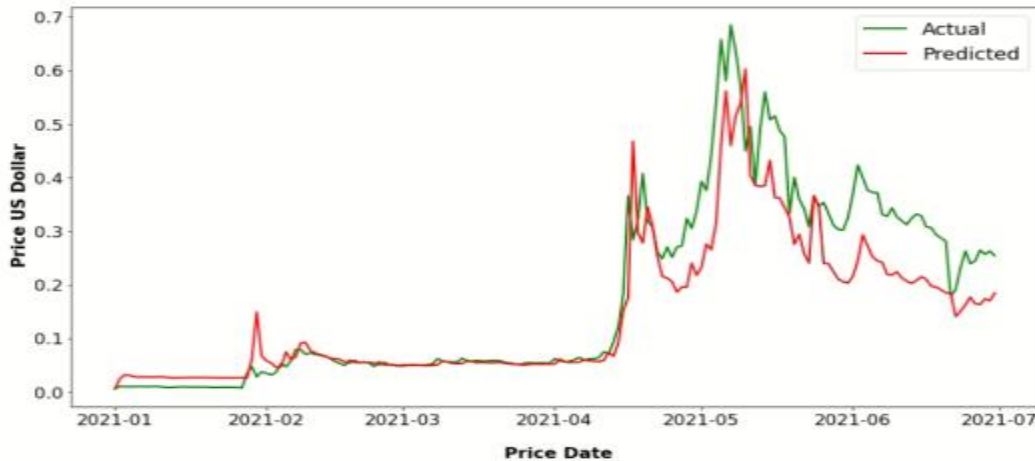


Figure 6.22. DOGE price prediction using Google trend.

For DOGE, by using google trends. Figure 6.22 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The clear difference between the predicted and actual was from the middle of April 2021 until 2021 until the end of July 2021. The outcome shows that the value of MAPE is 0.04% and the value of RMSE is 0.08.

Google trends are considered one of the most popular channels used to predict cryptocurrency prices, the experiment outcomes show that the better model is for DOGE, although the figures represent the opposite because the DOGE price is very low, on other hand the model using Google trends LTC represents the best model with the value of MAPE is 8.77 % and the value of RMSE is 26.31.

By comparing the previous result for the most popular channels those are Twitter, news, and Google trends by using VADER sentiment analysis and Harvard IV-4 for text analysis, the experiment outcomes show that the best model is the LTC model using VADER sentiment analysis with the value of 5.29% for MAPE if ignore DODE because of its low prices.

6.5 Cryptocurrency price prediction using data fusion techniques

Data fusion as a technique is data collection from multiple data sources to produce consistent and meaningful data that provide a better overview than a single source and also enables us to extract the best features from these sources to formulate a comprehensive dataset to be fed to the machine learning model. In this term, we apply the fusion techniques for three cryptocurrencies that are LTC, BTC, and, EHT. Data fusion as a technique contributes to enhancing model results like LTC than using a single channel to predict cryptocurrency. Table 6.4 illustrates the evaluation measurements using MAPE and RMSR for the three cryptocurrencies.

Table 6.4 Predictions of cryptocurrency prices using data fusion techniques

	MAPE	RMSE
LTC	2.88 %	162.33
BTC	5.62 %	437.98
ETH	3.48 %	189.32

Figures 6.23 – 6.25 illustrates the experiment outcomes of using data fusion techniques. It shows that the result is optimized compared with using a single channel. Figure 6.23 shows the data fusion results for the LTC it shows the predicted result is very close to the actual one, except for one point; which is the end of May 2021. The MAPE value is 2.88 % and the RMSE of 162.33% with a difference of 7.66 USD between the mean predicted and the actual. On the other hand, for BTC Figure 6.24 illustrate the model result of using data fusion techniques a MAPE value was 5.62% and 437.98 for RMSE with a difference between the mean of predicted and actual being 352.57 USD, finally figure 6.25 illustrates the model result of using data fusion techniques for ETH a MAPE value was 3.48% and 189.32 for RMSE with a difference between the mean of predicted and actual being 11.21 USD. From a statistical perspective, the maximum value of predicted and

actual BTC is 63772.062 USD and 63503.457 respectively and for LTC the maximum value for actual and predicted is 386.45 and 384.188 USD respectively. At the same time, the maximum value for actual and predicted ETH was 4168.701 and 4156.225 USD respectively.

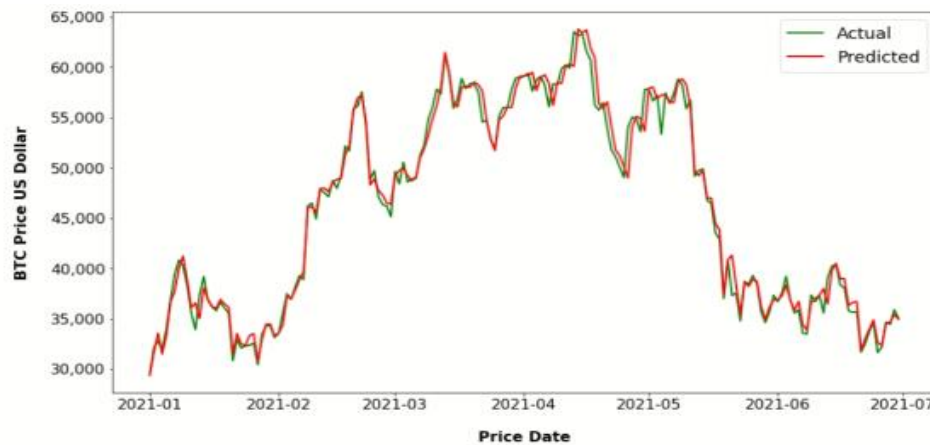


Figure 6.23. Comparison between actual and predicted values for BTC prediction using data fusion techniques.

For BTC, by data fusion techniques. Figure 6.23 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The difference appears on some days but overall the result shows that the actual and predicted are very close together. The outcome shows that the value of MAPE is 5.62 % and the value of RMSE is 437.98.

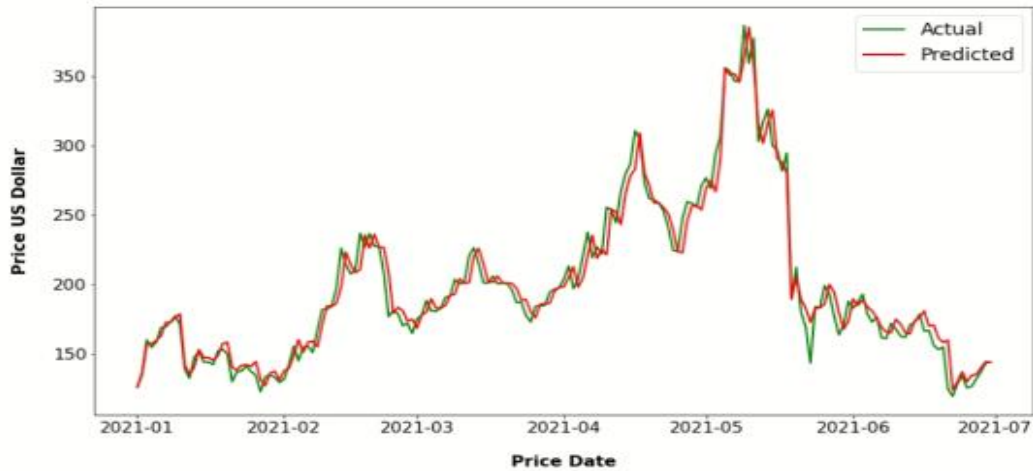


Figure 6.24. Comparison between actual and predicted values for LTC prediction using data fusion techniques.

For ETH, by data fusion techniques. Figure 6.24 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The difference appears on some days but overall the result shows that the actual and predicted are very close together. The outcome shows that the value of MAPE is 3.48 % and the value of RMSE is 189.32.

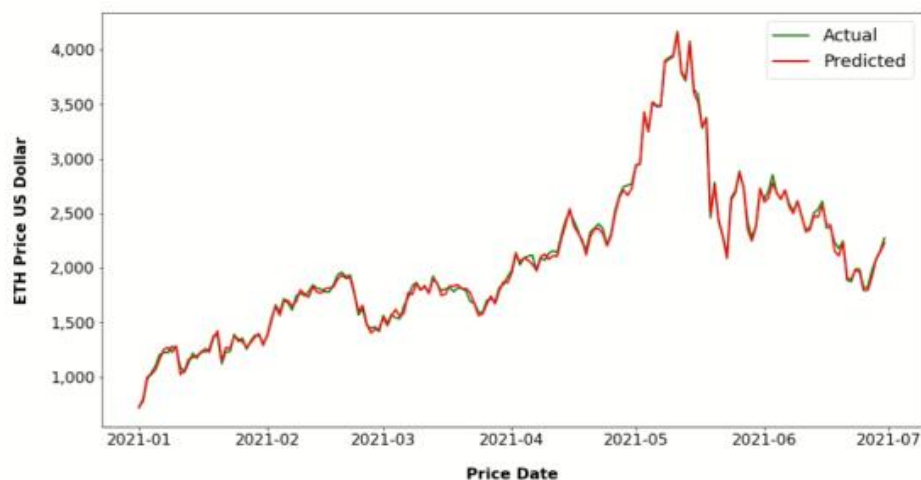


Figure 6.25. Comparison between actual and predicted values for ETH prediction using data fusion techniques.

For LTC, by data fusion techniques. Figure 6.25 show that the actual result was presented using a green line and the predicted presented using a red line, the experiment outcome shows the difference between the predicted and actual. The difference appears on some days but overall the result shows that the actual and predicted are very close together. The outcome shows that the value of MAPE is 2.88 % and the value of RMSE is 162.33.

6.6 Conclusion

Data fusion is a new approach that I proposed in this thesis it depends on collecting the data from multiple data sources, the proposed technique contributes to optimizing the prediction results rather than using a single data source. The approach has been applied using three types of cryptocurrency those are BTC, ETH, and DOGE, the experiment d was done using the collected data on a daily basis from 01 January 2021 to 30 June 2021 and implementing LSTM. The experiment outcomes show that the best model is for LTC, a model with a value of MAPE is 2.88% and a value of RMSE is 162.33.

Chapter Seven

Conclusion and Future Work

7.1 Conclusion

Prediction of cryptocurrency prices is one of the challenges that face investors in this domain because of the high fluctuation, in this research, revolutionary machine learning approaches were used to enable cryptocurrency players to devise a model that will forecast the upcoming cryptocurrency price movements with a smaller percentage of inaccuracy depending on data fusion techniques.

There are multiple techniques, and methods used to collect and preprocess the data from multiple data sources that formulate the model, in the same time interacting with big data is a big challenge so we propose the Divide, Clean, and Combine method (DCC) to interact with this amount of data.

In this research, the impact of news items on the bitcoin price movement is investigated. Other public information sources, such as social media is used to improve prediction accuracy. When it comes to making a more rational decision on whether to sell or purchase bitcoin, many traders rely on news services as their primary source of information. However, the growing volume of current textual information creates a significant information overload, making it more difficult for market participants to select the information that is relevant to them. As a result, we developed a system that forecasts cryptocurrencies price movements based on the subjectivity and polarity of bitcoin-related news headlines and historical price data in this project. We grabbed pertinent news headlines and cryptocurrency historical data from cryptonews, respectively, and then used

sentiment analysis to calculate the polarity and subjectivity scores for each headline. Although the model's accuracy is satisfactory, it might be improved by including more complicated classifiers. We simply looked at the impact of news items on the four different types of cryptocurrency price movement in this study. Other public information sources, such as social media and user comments in forums, can be used to improve prediction accuracy.

Four different cryptocurrencies were tested in this research; which is the most popular cryptocurrency that are BTC, ETH, LTC, and DOGE. The outcome of data represents better results than depending on a single data source. The model was built using Long short-term memory (LSTM) machine learning algorithm and evaluated using mean absolute squared error (MAPE). The results were satisfactory for each cryptocurrency but were the best for LTC, the MAPE results show that 5.62%, 3.48%, and 2.88% for BTC, ETH, and LTC respectively, on the other hand for DOGE there are no effects for data fusion techniques in a result.

7.2 Future Work

Future work will migrate other data sources that can be affected in cryptocurrency prices and implement other machine learning algorithms like NARX Neural Network and compare the result between them. Also applying the approach by using other cryptocurrencies. Also using more data daily prices of cryptocurrencies can give researchers new insights into this area. Use other parameters as input for example the tweets of famous people in the business sector.

References

M. Sami and W. Abdallah, “How does the cryptocurrency market affect the stock market performance in the MENA region?” *Journal of Economic and Administrative Sciences*, vol. 37, no. 4, pp. 741–753, 2021, issn: 1026-4116. doi: 10.1108/jeas-07-2019-0078.

C. Nolasco Braaten and M. S. Vaughn, “Convenience Theory of Cryptocurrency Crime: A Content Analysis of U.S. Federal Court Decisions,” *Deviant Behavior*, vol. 42, no. 8, pp. 958–978, 2021, issn: 15210456. doi: 10.1080/01639625.2019.1706706. [Online]. Available: <https://doi.org/10.1080/01639625.2019.1706706>.

K. H. McIntyre and K. Harjes, “Order Flow and the Bitcoin Spot Rate,” *Applied Economics and Finance*, vol. 3, no. 3, pp. 136–147, 2016, issn: 2332-7294. doi: 10.11114/aef.v3i3.1574.

E. Mnif, A. Jarboui, and K. Mouakhar, “How the cryptocurrency market has performed during COVID 19? A multifractal analysis,” *Finance Research Letters*, vol. 36, p. 101 647, 2020, issn: 15446123. doi: 10.1016/j.frl.2020.101647. [Online]. Available: <https://doi.org/10.1016/j.frl.2020.101647>.

F. Fang, C. Ventre, M. Basios, et al., *Cryptocurrency trading: A comprehensive survey*, 2021. arXiv: 2003.11352 [q-fin.TR].

S. Qureshi, M. Aftab, E. Bouri, and T. Saeed, “Dynamic interdependence of cryptocurrency markets: An analysis across time and frequency,” *Physica A: Statistical Mechanics and its Applications*, vol. 559, no. December 2019, p. 125 077, 2020, issn: 03784371. doi: 10.1016/j.physa.2020.125077. [Online]. Available: <https://doi.org/10.1016/j.physa.2020.125077>.

F. Colon, C. Kim, H. Kim, and W. Kim, “The effect of political and economic uncertainty on the cryptocurrency market,” *Finance Research Letters*, vol. 39, no. February, p. 101 621, 2021, issn: 15446123. doi: 10.1016/j.frl.2020.101621. [Online]. Available: <https://doi.org/10.1016/j.frl.2020.101621>.

A. Brauneis, R. Mestel, R. Riordan, and E. Theissen, “How to measure the liquidity of cryptocurrency markets?” *Journal of Banking and Finance*, vol. 124, p. 106 041, 2021, issn: 03784266. doi: 10.1016/j.jbankfin.2020.106041. [Online]. Available: <https://doi.org/10.1016/j.jbankfin.2020.106041>.

H. Sebastião and P. Godinho, “Forecasting and trading cryptocurrencies with machine learning under changing market conditions,” *Financial Innovation*, vol. 7, no. 1, 2021, issn: 21994730. doi: 10.1186/s40854-020-00217-x. [Online]. Available: <https://doi.org/10.1186/s40854-020-00217-x>.

X. Li and C. A. Wang, “The technology and economic determinants of cryptocurrency exchange rates: The case of bitcoin,” *Decision Support Systems*, vol. 95, pp. 49–60, 2017, issn: 0167-9236. doi: <https://doi.org/10.1016/j.dss.2016.12.001>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167923616302111>.

R. Caferra and D. Vidal-Tomás, “Who raised from the abyss? A comparison between cryptocurrency and stock market dynamics during the COVID-19 pandemic,” *Finance Research Letters*, vol. 43, no. February, p. 101 954, 2021, issn: 15446123. doi: 10.1016/j.frl.2021.101954. [Online]. Available: <https://doi.org/10.1016/j.frl.2021.101954>.

F. Glaser, K. Zimmermann, M. Haferkorn, M. C. Weber, and M. Siering, “Bitcoin - Asset or currency? Revealing users’ hidden intentions,” ECIS 2014 Proceedings - 22nd European Conference on Information Systems, pp. 1–14, 2014.

M. Watorek, S. Drożdż, J. Kwapień, L. Minati, P. Oświcimka, and M. Stanuszek, “Multiscale characteristics of the emerging global cryptocurrency market,” *Physics Reports*, vol. 901, pp. 1–82, 2021, issn: 03701573. doi: 10.1016/j.physrep.2020.10.005. arXiv: 2010.15403. [Online]. Available: <https://doi.org/10.1016/j.physrep.2020.10.005>.

E. Bouri, T. Saeed, X. V. Vo, and D. Roubaud, “Quantile connectedness in the cryptocurrency market,” *Journal of International Financial Markets, Institutions and Money*, vol. 71, p. 101 302, 2021, issn: 10424431. doi: 10.1016/j.intfin.2021.101302. [Online]. Available: <https://doi.org/10.1016/j.intfin.2021.101302>.

G. Aggarwal, V. Patel, G. Varshney, and K. Oostman, “Understanding the Social Factors Affecting the Cryptocurrency Market,” *CoRR*, vol. abs/1901.0, p. 8, 2019. arXiv: 1901.06245. [Online]. Available: <http://arxiv.org/abs/1901.06245>.

I. Amsyar, E. Christopher, A. Dithi, A. N. Khan, and S. Maulana, “The Challenge of Cryptocurrency in the Era of the Digital Revolution: A Review of Systematic Literature,” *Aptisi Transactions on Technopreneurship (ATT)*, vol. 2, no. 2, pp. 153–159, 2020, issn: 2655-8807. doi: 10.34306/att.v2i2.96.

L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, M. Tesconi, and E. Ferrara, “Charting the landscape of online cryptocurrency manipulation,” *IEEE Access*, vol. 8, pp. 113 230–113 245, 2020. doi: 10.1109/ACCESS.2020.3003370.

M. C. S,cheau, S. L. Crăciunescu, I. Brici, and M. V. Achim, “A Cryptocurrency Spectrum Short Analysis,” *Journal of Risk and Financial Management*, vol. 13, no. 8, p. 184, 2020, issn: 1911-8074. doi: 10.3390/jrfm13080184

A. G. Țițan, “The Efficient Market Hypothesis: Review of Specialized Literature and Empirical Research,” *Procedia Economics and Finance*, vol. 32, no. 15, pp. 442–449, 2015, issn: 22125671. doi: 10.1016/s2212-5671(15)01416-1

E. Dorokhova, E. Dorokhova, T. Belykh, and G. Koren’kova, “Economic and Legal Aspects of Cryptocurrency Usage in Russia,” *Proceedings of the 3rd International Conference Spatial Development of Territories (SDT 2020)*, vol. 181, no. Sdt 2020, pp. 327–332, 2021. doi: 10.2991/aebmr.k.210710.055

A. Dutta, G. Pooja, N. Jain, R. R. Panda, and N. K. Nagwani, A Hybrid Deep Learning Approach for Stock Price Prediction. 2021, vol. 141, pp. 1–10, isbn: 9789811571053. doi: 10.1007/978-981-15-7106-0_1.

R. S. Gadey, N. Thakur, N. Charan, and R. Obulakonda Reddy, “Price Prediction of Bitcoin Using Machine Learning,” *International Journal of Engineering Applied Sciences and Technology*, vol. 5, no. 1, pp. 502–506, 2020, issn: 2455-2143. doi: 10.33564/ijeast.2020.v05i01.089

M. Froehlich, M. R. Wagenhaus, A. Schmidt, and F. Alt, “Don’t Stop Me Now! Exploring Challenges of First-Time Cryptocurrency Users,” DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere, no. 1, pp. 138–148, 2021. doi: 10.1145/3461778.3462071.

Badari, Ananda and Chaudhury, Archie, An Overview of Bitcoin and Ethereum White-Papers, Forks, and Prices (April 26, 2021). Available at SSRN: <https://ssrn.com/abstract=3841827> or <http://dx.doi.org/10.2139/ssrn.3841827>

Hileman, G., & Rauchs, M. (2017). Global Benchmarking Study. Cambridge Centre for Alternative Finance, 10. <https://www.crowdfundinsider.com/wp-content/uploads/2017/04/Global-Cryptocurrency-Benchmarking-Study.pdf>

W. Viriyasitavat, D. Hoonsopon, and Z. Bi, “Augmenting cryptocurrency in smart supply chain,” Journal of Industrial Information Integration, vol. 21, p. 100–188, 2021, issn: 2452414X. doi: 10.1016/j.jii.2020.100188. [Online]. Available: <https://doi.org/10.1016/j.jii.2020.100188>.

Agrawal, Chatterjee, J. Kumar, and. Rathore, Blockchain Technology and the Internet of Things: Challenges and Applications in Bitcoin and Security (1st ed.) Apple Academic Press, 2020. doi: <https://doi.org/10.1201/9781003022688>.

C. F. Durach, T. Blesik, M. von Düring, and M. Bick, “Blockchain Applications in Supply Chain Transactions,” Journal of Business Logistics, vol. 42, no. 1, pp. 7–24, 2021, issn: 21581592. doi: 10.1111/jbl.12238.

M. Kouhizadeh, S. Saberi, and J. Sarkis, “Blockchain technology and the sustainable supply chain: Theoretically exploring adoption barriers,” *International Journal of Production Economics*, vol. 231, p. 107 831, 2021, issn: 09255273. doi: 10.1016/j.ijpe.2020.107831. [Online]. Available: <https://doi.org/10.1016/j.ijpe.2020.107831>.

I. Abu-elezz, A. Hassan, A. Nazeemudeen, M. Househ, and A. Abd-alrazaq, “The benefits and threats of blockchain technology in healthcare: A scoping review,” *International Journal of Medical Informatics*, vol. 142, no. February, p. 104 246, 2020, issn: 18728243. doi: 10.1016/j.ijmedinf.2020.104246. [Online]. Available: <https://doi.org/10.1016/j.ijmedinf.2020.104246>.

M. Soni and D. K. Singh, “Blockchain-based security privacy for biomedical and healthcare information exchange systems,” *Materials Today: Proceedings*, no. xxxx, 2021, issn: 22147853. doi: 10.1016/j.matpr.2021.02.094. [Online]. Available: <https://doi.org/10.1016/j.matpr.2021.02.094>.

E. Zaghloul, T. Li, M. W. Mutka, and J. Ren, “Bitcoin and Blockchain: Security and Privacy,” *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10 288–10 313, 2020, issn: 23274662. doi: 10.1109/JIOT.2020.3004273. arXiv: 1904.11435.

K. Martin, I. Alsmadi, M. Rahouti, and M. Ayyash, “Combining blockchain and machine learning to forecast cryptocurrency prices,” in *2020 Second International Conference on Blockchain Computing and Applications (BCCA)*, Nov. 2020, pp. 52–58. doi: 10.1109/BCCA50787.2020.9274454

M. R. Biktimirov, A. V. Domashev, P. A. Cherkashin, and A. Y. Shcherbakov, "Blockchain Technology: Universal Structure and Requirements," *Automatic Documentation and Mathematical Linguistics*, vol. 51, no. 6, pp. 235–238, 2017, ISSN 0005-1055. doi: 10.3103/s0005105517060036.

Tsolakis, Naoum; Niedenzu, Denis; Simonetto, Melissa; Dora, Manoj; Kumar, Mukesh (2020). Supply network design to address United Nations Sustainable Development Goals: A case study of blockchain implementation in Thai fish industry. *Journal of Business Research*, (), S0148296320304914–. doi:10.1016/j.jbusres.2020.08.003

Clohesy, T., Treiblmaier, H., Acton, T., & Rogers, N. (2020). Antecedents of blockchain adoption: An integrative framework. *Strategic Change*, 29(5), 501–515. <https://doi.org/10.1002/jsc.2360>

Biktimirov, M. R.; Domashev, A. V.; Cherkashin, P. A.; Shcherbakov, A. Yu. (2017). *Blockchain Technology: Universal Structure and Requirements. Automatic Documentation and Mathematical Linguistics*, 51(6), 235 –238. doi: 10.3103/S0005105517060036

P. Jay, V. Kalariya, P. Parmar, S. Tanwar, N. Kumar, and M. Alazab, "Stochastic neural networks for cryptocurrency price prediction," *IEEE Access*, vol. 8, pp. 82 804–82 818, 2020. doi: 10.1109/ACCESS.2020.2990659.

Hutto, C.J. and Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for. Eighth International AAAI Conference on Weblogs and Social Media, 18. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109>

O. Kraaijeveld and J. De Smedt, “The predictive power of public twitter sentiment for forecasting cryptocurrency prices,” *Journal of International Financial Markets, Institutions and Money*, vol. 65, p. 101–188, 2020, issn: 1042-4431. doi: <https://doi.org/10.1016/j.intfin.2020.101188>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S104244312030072X>.

J. Abraham, D. Higdon, and J. Nelson, “Cryptocurrency price prediction using tweet volumes and sentiment analysis,” *SMU Data Science Review*, vol. 1, no. 3, p. 22, 2018. [Online]. Available: <https://scholar.smu.edu/datasciencereviewhttp://digitalrepository.smu.edu>. Available at: <https://scholar.smu.edu/datasciencereview/vol1/iss3/1>.

Jahiruddin. (2015). Sentiment Analysis of Twitter Data using Statistical Methods. In *International Journal of Innovative Research in Engineering & Management (IJIREM)* (Issue 2).

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>.

A. Inamdar, A. Bhagtani, S. Bhatt, and P. M. Shetty, “Predicting cryptocurrency value using sentiment analysis,” pp. 932–934, 2019. doi: 10.1109/ICCS45141.2019.9065838.

J. Abraham, D. Higdon, and J. Nelson, “Cryptocurrency price prediction using tweet volumes and sentiment analysis,” *SMU Data Science Review*, vol. 1, no. 3,

p. 22, 2018. [Online]. Available: <https://scholar.smu.edu/datasciencereviewhttp://digitalrepository.smu.edu.Avalableat:https://scholar.smu.edu/datasciencereview/vol1/iss3/1>.

S. Rouhani and E. Abedin, "Crypto-currencies narrated on tweets: a sentiment analysis approach," *International Journal of Ethics and Systems*, vol. 36, no. 1, pp. 58–72, 2020, issn: 25149369. doi: 10.1108/IJOES-12-2018-0185.

H. Jang and J. Lee, "An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information," *IEEE Access*, vol. 6, pp. 5427–5437, 2018. doi: 10.1109/ACCESS.2017.2779181.

T. Pano and R. Kashef, "A complete vader-based sentiment analysis of bitcoin (BTC) tweets during the ERA of COVID-19," *Big Data and Cognitive Computing*, vol. 4, no. 4, pp. 1–17, 2020, issn: 25042289. doi: 10.3390/bdcc4040033.

Hu, M., Li, W., Yan, K., Ji, Z., & Hu, H. (2019). Modern Machine Learning Techniques for Univariate Tunnel Settlement Forecasting: A Comparative Study. *Mathematical Problems in Engineering*, 2019. <https://doi.org/10.1155/2019/7057612>.

A. M. Balfagih and V. Keselj, "Evaluating Sentiment Classifiers for Bitcoin Tweets in Price Prediction Task," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 5499-5506, doi: 10.1109/BigData47090.2019.9006140.

B. Preisler, M. Mieskes, and C. Becker, "Bitcoin value and sentiment expressed in tweets," *CEUR Workshop Proceedings*, vol. 2458, 2019, issn: 16130073.

S. Nair, "Cryptocurrencies Price Movement Prediction Using Machine Learning," International Journal of Innovative Science and Research Technology, vol. 6, no. 2, 2021. [Online]. Available: www.ijisrt.com730.

M. Yasir, M. Attique, K. Latif, et al., "Deep-learning-assisted business intelligence model for cryptocurrency forecasting using social media sentiment," Journal of Enterprise Information Management, 2020, issn: 17410398. doi: 10.1108/JEIM-02-2020-0077.

J. Luo, "Bitcoin price prediction in the time of COVID-19," 2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID), 2020, pp. 243-247, doi: 10.1109/MSIEID52046.2020.00050.

A. M. Balfagih and V. Keselj, "Evaluating Sentiment Classifiers for Bitcoin Tweets in Price Prediction Task," Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019, pp. 5499–5506, 2019. doi: 10.1109/BigData47090.2019.9006140.

O. Sattarov, H. S. Jeon, R. Oh, and J. D. Lee, "Forecasting bitcoin price fluctuation by twitter sentiment analysis," 2020 International Conference on Information Science and Communications Technologies, ICISCT 2020, pp. 0–3, 2020. doi: 10.1109/ICISCT50599.2020.9351527.

A. Salac, "Forecasting of the Cryptocurrency Market Through Social Media Sentiment Analysis," pp. 1–12, 2019. [Online]. Available: http://essay.utwente.nl/78607/1/Salac_BA_BMS.pdf.

A. Ibrahim, “Forecasting the early market movement in bitcoin using twitter’s sentiment analysis: An ensemble-based prediction model,” 2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings, 2021. doi: 10.1109/IEMTRONICS52119.2021.9422647.

Z. H. Kilimci, “Sentiment analysis based direction prediction in bitcoin using deep learning algorithms and word embedding models,” International Journal of Intelligent Systems and Applications in Engineering, vol. 8, no. 2, pp. 60–65, 2020, issn: 21476799. doi: 10.18201/ijisae.2020261585.

D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel, and B. K. Lama, “Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis,” Proceedings on 2018 IEEE 3rd International Conference on Computing, Communication and Security, ICCCS 2018, no. October, pp. 128–132, 2018. doi: 10.1109/CCCS.2018.8586824.

L. Steinert and C. Herff, “Predicting altcoin returns using social media,” PLoS ONE, vol. 13, no. 12, pp. 1–12, 2018, issn: 19326203. doi: 10.1371/journal.pone.0208119.

A. Inamdar, A. Bhagtani, S. Bhatt, and P. M. Shetty, “Predicting cryptocurrency value using sentiment analysis,” pp. 932–934, 2019. doi: 10.1109/ICCS45141.2019.9065838.

M. S. Brown and B. Douglass, “An event study of the effects of cryptocurrency thefts on cryptocurrency prices,” pp. 1–12, 2020. doi: 10.22360/SpringSim.2020.CSE.001.

E. Şaşmaz and F. B. Tek, “Tweet sentiment analysis for cryptocurrencies,” in 2021 6th International Conference on Computer Science and Engineering (UBMK), Sep. 2021, pp. 613–618. doi: 10.1109/UBMK52708.2021.9558914.

D. Shah, H. Isah and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4705-4708, doi: 10.1109/BigData.2018.8621884.

C. Lamon, E. Nielsen, and E. Redondo, “Cryptocurrency Price Prediction Using News and Social Media Sentiment,” Pdfs.Semanticscholar.Org, vol. 25, p. 96, 2016. [Online]. Available: [http : / / cs229 . stanford . edu / proj2017 / final - reports/5237280.pdf%5C%0Ahttps://pdfs.semanticscholar.org/c3b8/0de058596cee95beb20a2d087dbcf8be01ea.pdf](http://cs229.stanford.edu/proj2017/final-reports/5237280.pdf%5C%0Ahttps://pdfs.semanticscholar.org/c3b8/0de058596cee95beb20a2d087dbcf8be01ea.pdf).

A.-D. Vo, “Sentiment Analysis of News for Effective Cryptocurrency Price Prediction,” International Journal of Knowledge Engineering, vol. 5, no. 2, pp. 47–52, 2019, issn: 23826185. doi: 10.18178/ijke.2019.5.2.116.

M. S. Brown and B. Douglass, “An event study of the effects of cryptocurrency thefts on cryptocurrency prices,” pp. 1–12, 2020. doi: 10.22360/SpringSim.2020.CSE.001.

A. F. Vilas, R. P. Díaz Redondo, D. C. Cancela, and A. T. Pazos, “Interplay between cryptocurrency transactions and online financial forums,” Mathematics, vol. 9, no. 4, pp. 1–22, 2021, issn: 22277390. doi: 10.3390/math9040411.

A. Mittal, V. Dhiman, A. Singh, and C. Prakash, “Short-Term Bitcoin Price Fluctuation Prediction Using Social Media and Web Search Data,” 2019 12th International Conference on Contemporary Computing, IC3 2019, pp. 1–6, 2019. doi: 10.1109/IC3.2019.8844899.

M. Al Guindy, “Cryptocurrency price volatility and investor attention,” *International Review of Economics and Finance*, vol. 76, no. June, pp. 556–570, 2021, issn: 10590560. doi: 10.1016/j.iref.2021.06.007. [Online]. Available: <https://doi.org/10.1016/j.iref.2021.06.007>.

Sabalionis, A, Wang, W, Park, H. What affects the price movements in Bitcoin and Ethereum?. The Manchester School. 2020;89: 102– 127. <https://doi.org/10.1111/manc.12352>

V. Karalevicius, “Using sentiment analysis to predict interday Bitcoin price movements,” *Journal of Risk Finance*, vol. 19, no. 1, pp. 56–75, 2018, issn: 09657967. doi: 10.1108/JRF-06-2017-0092.

S. Bibi, S. Hussain, and M. I. Faisal, “Public perception based recommendation system for cryptocurrency,” pp. 661–665, 2019. doi: 10.1109/IBCAST.2019.8667123.

B. Agarwal, P. Harjule, L. Chouhan, U. Saraswat, H. Airan, and P. Agarwal, “Prediction of dogecoin price using deep learning and social media trends,” *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, no. September, p. 171 188, 2021. doi: 10.4108/eai.29-9-2021.171188.

Sun, Xiaolei; Liu, Mingxi; Sima, Zeqian (2018). *A novel cryptocurrency price trend forecasting model based on LightGBM*. *Finance Research Letters*, (), S1544612318307918–. doi:10.1016/j.frl.2018.12.032

M. J. Hamayel and A. Y. Owda, “A novel cryptocurrency price prediction model using gru, lstm and bi-lstm machine learning algorithms,” *AI*, vol. 2, no. 4, pp. 477–496, 2021, issn: 2673-2688. doi: 10.3390/ai2040030. [Online]. Available: <https://www.mdpi.com/2673-2688/2/4/30>.

E. Akyildirim, A. Goncu, and A. Sensoy, “Prediction of cryptocurrency returns using machine learning,” *Annals of Operations Research*, vol. 297, no. 1-2, pp. 3–36, 2021, issn: 15729338. doi: 10.1007/s10479-020-03575-y. [Online]. Available: <https://doi.org/10.1007/s10479-020-03575-y>.

M. M. Patel, S. Tanwar, R. Gupta, and N. Kumar, “A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions,” *Journal of Information Security and Applications*, vol. 55, no. June, p. 102 583, 2020, issn: 22142126. doi: 10.1016/j.jisa.2020.102583. [Online]. Available: <https://doi.org/10.1016/j.jisa.2020.102583>.

V. Derbentsev, V. Babenko, K. Khrustalev, H. Obruch, and S. Khrustalova, “Comparative performance of machine learning ensemble algorithms for forecasting cryptocurrency prices,” *International Journal of Engineering, Transactions A: Basics*, vol. 34, no. 1, pp. 140–148, 2021, issn: 17281431. doi: 10.5829/IJE.2021.34.01A.16.

T. A. Borges and R. F. Neves, “Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods,” *Applied Soft Computing Journal*, vol. 90, p. 106 187, 2020, issn: 15684946. doi: 10.1016/j.asoc.2020.106187. [Online]. Available: <https://doi.org/10.1016/j.asoc.2020.106187>.

H. Pabuccu, S. Ongan, and A. Ongan, “Forecasting the movements of Bitcoin prices: an application of machine learning algorithms,” *Quantitative Finance and Economics*, vol. 4, no. 4, pp. 679–692, 2020, issn: 2573-0134. doi: 10.3934/qfe.2020031.

N. Maleki, A. Nikoubin, M. Rabbani, and Y. Zeinali, *Bitcoin Price Prediction Based on Other Cryptocurrencies Using Machine Learning and Time Series Analysis*, 0. 2020, vol. 0, pp. 0–0, isbn: 9890130750. doi: 10.24200/sci.2020.55034.4040.

S. M. Iqbal, “Forecasting Cryptocurrency Prices using Machine Learning Data Analytics Ashwini Chaudhari National College of Ireland,” *National College of Ireland*, 2020.

I. E. Livieris, N. Kiriakidou, S. Stavroyiannis, and P. Pintelas, “An advanced CNN-LSTM model for cryptocurrency forecasting,” *Electronics (Switzerland)*, vol. 10, no. 3, pp. 1–16, 2021, issn: 20799292. doi: 10.3390/electronics10030287.

R. C. Lakshmi and T. Jaya, “Cryptocurrency Price Prediction: A Machine learning Approache,” *Sensors Transducers Published by IFSA Publishing, S. L.*, 2020, vol. 244, no. 5, pp. 44–47, 2020.

M. K. Salman and A. A. Ibrahim, "Price prediction of different cryptocurrencies using technical trade indicators and machine learning," IOP Conference Series: Materials Science and Engineering, vol. 928, no. 3, 2020, issn: 1757899X. doi: 10.1088/1757-899X/928/3/032007.

Y. Wang and R. Chen, "Cryptocurrency price prediction based on multiple market sentiment," Proceedings of the Annual Hawaii International Conference on System Sciences, vol. 2020-January, pp. 1092–1100, 2020, issn: 15301605. doi: 10.24251/hicss.2020.136.

Y. Indulkar, "Time series analysis of cryptocurrencies using deep learning amp; fbprophet," in 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Mar. 2021, pp. 306–311. doi: 10.1109/ESCI50559.2021.9397004.

T. E. Koker and D. Koutmos, "Cryptocurrency Trading Using Machine Learning," Journal of Risk and Financial Management, vol. 13, no. 8, p. 178, 2020, issn: 1911-8074. doi: 10.3390/jrfm13080178.

M. Asgari and H. Khasteh, "Application of Three Different Machine Learning Methods on Strategy Creation for Profitable Trades on Cryptocurrency Markets," arXiv preprint arXiv:2105.06827, 2021. [Online]. Available: <https://arxiv.org/abs/2105.06827>.

A. M. Khedr, I. Arif, P. V. Pravija Raj, M. El-Bannany, S. M. Alhashmi, and M. Sreedharan, "Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey," Intelligent Systems in Accounting, Fi-

nance and Management, vol. 28, no. 1, pp. 3–34, 2021, issn: 21600074. doi: 10.1002/isaf.1488.

L. Vaddi, “Predicting Crypto Currency Prices Using Machine Learning and Deep Learning Techniques,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 6603–6608, 2020. doi: 10.30534/ijatcse/2020/351942020.

S. A. Alahmari, “Using Machine Learning ARIMA to Predict the Price of Cryptocurrencies,” *The ISC International Journal of Information Security*, vol. 11, no. 3, pp. 139–144, 2019, issn: 2008-2045. [Online]. Available: [http : / / www . isecure-journal.com/article_90865.html](http://www.isecure-journal.com/article_90865.html).

X. Li, Q. He, B. Jiang, X. Qin, and K. Qin, *Blockchain and Trustworthy Systems*. 2020, vol. 1156, pp. 607–620, isbn: 978-981-15-2776-0. doi: 10.1007/978-981-15-9213-3. [Online]. Available: <http://link.springer.com/10.1007/978-981-15-2777-7>.

S. Saadah and A. Ahmad Whafa, “Monitoring financial stability based on prediction of cryptocurrencies price using intelligent algorithm,” in *2020 International Conference on Data Science and Its Applications (ICoDSA)*, Aug. 2020, pp. 1–10. doi: 10.1109/ICoDSA50139.2020.9212968.

A. Politis, K. Doka, and N. Koziris, “Ether price prediction using advanced deep learning models,” in *2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, May 2021, pp. 1–3. doi: 10.1109/ICBC51069.2021.9461061.

M. Kenny, “Investigating the predictability of a Chaotic time-series data using Reservoir computing, Deep-Learning and Machine- Learning on the short- , medium- and long-term pricing of Bitcoin and Ethereum .,” Technological University Dublin, 2020. doi: 10.21427/42ks-q868.

Wei, M. (2021). Essays in cryptocurrencies’ forecasting and trading with technical analysis and advanced machine learning methods.

K. Cortez, M. D. P. Rodríguez-García, and S. Mongrut, “Exchange market liquidity prediction with the k-nearest neighbor approach: Crypto vs. fiat currencies,” *Mathematics*, vol. 9, no. 1, pp. 1–15, 2021, issn: 22277390. doi: 10.3390/math9010056.

H. M. Kim, G. W. Bock, and G. Lee, “Predicting Ethereum prices with machine learning based on Blockchain information,” *Expert Systems with Applications*, vol. 184, no. May, p. 115 480, 2021, issn: 09574174. doi: 10.1016/j.eswa.2021.115480. [Online]. Available: <https://doi.org/10.1016/j.eswa.2021.115480>.

D. G. ANGHEL, “A reality check on trading rule performance in the cryptocurrency market: Machine learning vs. technical analysis,” *Finance Research Letters*, vol. 39, p. 101 655, 2021, issn: 15446123. doi: 10.1016/j.frl.2020.101655. [Online]. Available: <https://doi.org/10.1016/j.frl.2020.101655>.

Q. Wang, “Cryptocurrencies asset pricing via machine learning,” *International Journal of Data Science and Analytics*, vol. 12, no. 2, pp. 175–183, 2021, ISSN: 23644168. doi: 10 . 1007 / s41060 - 021 - 00252 - 6. [Online]. Available: <https://doi.org/10.1007/s41060-021-00252-6>.

T. Zoumpakas, E. Houstis, and M. Vavalis, “ETH analysis and predictions utilizing deep learning,” *Expert Systems with Applications*, vol. 162, no. March, p. 113 866, 2020, issn: 09574174. doi: 10.1016/j.eswa.2020.113866. [Online]. Available: <https://doi.org/10.1016/j.eswa.2020.113866>.

King, R. C., Villeneuve, E., White, R. J., Sherratt, R. S., Holderbaum, W., & Harwin, W. S. (2017). Application of data fusion techniques and technologies for wearable health monitoring. In *Medical Engineering and Physics* (Vol. 42, pp. 1–12). Elsevier Ltd. <https://doi.org/10.1016/j.medengphy.2016.12.011>

A. Sriram, Y. Li and A. Hadaegh, "Mining Social Media to Understand User Opinions on IoT Security and Privacy," *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2021, pp. 252-257, doi: 10.1109/SMARTCOMP52413.2021.00056.

K. H. Manguri, R. N. Ramadhan, and P. R. Mohammed Amin, “Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks”, Kurdistan Journal of Applied Research, vol. 5, no. 3, pp. 54-65, May 2020.

Victor Santoni;Samuel Rufat; (2021). How fast is fast enough? Twitter usability during emergencies . Geoforum, (), –. doi:10.1016/j.geoforum.2021.05.007

R. Gunawan, A. Rahmatulloh, I. Darmawan, and F. Firdaus, “Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath,” vol. 2, no. IcoIESE 2018, pp. 283–287, 2019. doi: 10.2991/icoiese-18.2019.50.

Aggarwal, G., Patel, V., Varshney, G., & Oostman, K. (2019). Understanding the Social Factors Affecting the Cryptocurrency Market. <http://arxiv.org/abs/1901.06245>

Crypto News is a service for crypto enthusiasts made by crypto enthusiasts. Available online <https://cryptonews.net> (accessed on 1 February 2022).

Yahoo Finance provides stocks, foreign exchange, and cryptocurrency prices. Available online at <https://finance.yahoo.com> accessed on 16 June 2021

Market Watch Which Provides the Latest Stock Market, Financial, and Business News. Available online: [**https://www.marketwatch.com**](https://www.marketwatch.com) (accessed on 16 June 2021).

bitinfocharts provides a simple representation of the cryptocurrency data points. online <https://bitinfocharts.com/> (accessed on 2 Oct 2021)

Yahoo Finance provides stocks, foreign exchange, and cryptocurrency prices. Available online at <https://finance.yahoo.com> accessed on 16 June 2021

N. Singh and M. Vardhan, “Computing Optimal Block Size for Blockchain-based Applications with Contradictory Objectives,” *Procedia Computer Science*, vol. 171, no. 2019, pp. 1389–1398, 2020, issn: 18770509. doi: 10.1016/j.procs.2020.04.149. [Online]. Available: <https://doi.org/10.1016/j.procs.2020.04.149>.

Wu, K., Wheatley, S., & Sornette, D. (2018). Classification of cryptocurrency coins and tokens by the dynamics of their market capitalizations. *Royal Society Open Science*, 5(9). <https://doi.org/10.1098/rsos.180381>

X. Zhang, R. Qin, Y. Yuan and F. -Y. Wang, "An Analysis of Blockchain-based Bitcoin Mining Difficulty: Techniques and Principles," 2018 Chinese Automation Congress (CAC), 2018, pp. 1184-1189, doi: 10.1109/CAC.2018.8623140.

Siddiqui, S., Vanahalli, G., & Gujar, S. (2020). BitcoinF: Achieving fairness for bitcoin in transaction-fee-only model. Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, 2020-May, 2008–2010.

Roughgarden, T. (2021). Transaction fee mechanism design. ACM SIGecom Exchanges, 19(1), 52–55. <https://doi.org/10.1145/3476436.3476445>

Huang, M.Y., Rojas, R.R. & Convery, P.D. Forecasting stock market movements using Google Trend searches. *Empir Econ* 59, 2821–2839 (2020). <https://doi.org/10.1007/s00181-019-01725-1>

Crypto News is a service for crypto enthusiasts made by crypto enthusiasts. Available online <https://cryptonews.net> (accessed on 1 February 2022).

Holton, A. E. (2014). Weller et al., Twitter and Society. *The Journal of Media Innovations*, 1(1), 134–137. <https://doi.org/10.5617/jmi.v1i1.825>

Benjamin Lindemann; Benjamin Maschler; Nada Sahlab; Michael Weyrich; (2021). A survey on anomaly detection for technical systems using LSTM networks . *Computers in Industry*, (), – doi:10.1016/j.compind.2021.103498

Sherratt, F.; Plummer, A.; Iravani, P. Understanding LSTM Network Behaviour of IMU-Based Locomotion Mode Recognition for Applications in Prostheses and Wearables. *Sensors* 2021, 21, 1264.

Jozefowicz, R.; Zaremba, W.; Sutskever, I. An empirical exploration of Recurrent Network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, 6–11 July 2015; Volume 3, pp. 2332–2340.

Li, X.; Peng, L.; Yao, X.; Cui, S.; Hu, Y.; You, C.; Chi, T. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environ. Pollut.* 2017, 231, 997–1004.

Kalra, V., & Aggarwal, R. (2018). Importance of Text Data Preprocessing & Implementation in RapidMiner. *Proceedings of the First International Conference on Information Technology and Knowledge Management*, 14(January), 71–75.
<https://doi.org/10.15439/2017km46>

Ek, A., Bernardy, J.-P., Chatzikyriakidis, S. (2020). How does Punctuation Affect Neural Models in Natural Language Inference. *Proceedings of the Probability and Meaning Conference (PaM 2020)*, PaM, 109–116.
<https://www.aclweb.org/anthology/2020.pam-1.15>

Saif, H., Fernandez, M., He, Y., Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, May, 810–817.

Bose, R., Aithal, P. S., & Roy, S. (2021). International MultiConference of Engineers and Computer Scientists : IMECS 2019 : 13-15 March 2019 : The Royal Garden Hotel, Kowloon, Hong Kong. International Journal of Management, Technology, and Social Sciences (IJMTS), 6(1), 2581–6012.

Bose, R., Dey, R. K., Roy, S., & Sarddar, D. (2018). Sentiment Analysis on Online Product Reviews. In Information and Communication Technology for Sustainable Development. Advances in Intelligent Systems and Computing, 933(1), 559 – 569.

Singh, Dalwinder; Singh, Birmohan (2019). Investigating the impact of data normalization on classification performance. Applied Soft Computing, (), 105524–. doi:10.1016/j.asoc.2019.105524

Naldi, M. (2019). A review of sentiment computation methods with R packages. ArXiv, 1–11.

Kearney, Colm; Liu, Sha (2014). Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis, 33(), 171–185. doi:10.1016/j.irfa.2014.02.006

Inamdar, Abid; Bhagtani, Aarti; Bhatt, Suraj; Shetty, Pooja M. (2019). [IEEE 2019 International Conference on Intelligent Computing and Control Systems (ICCS) - Madurai, India (2019.5.15-2019.5.17)] 2019 International Conference on Intelligent Computing and Control Systems (ICCS) - Predicting Cryptocurrency Value using Sentiment Analysis. , (), 932–934. doi:10.1109/ICCS45141.2019.9065838

الملخص

تعتبر عملية الاتجار بالعملات الرقمية واحد من اهم المنافذ التي يستغلها المستثمرين لجني الربح من خلالها وذلك باستخدام اسلوب المضاربة والذي يعتبر احد الاساليب الاستثمارية سواء في اسواق العملات المشفرة او اسواق العملات وذلك بالاعتماد على التقلبات في اسعار العملات الرقمية. تعتبر العمليات الاستثمارية بشكل عام والاستثمار في العملات الرقمية واحد من اكثر عمليات الاستثمار مخاطرة نتيجة التقلبات السريعة في اسواق العملات الرقمية حيث اصبحت محط انظار واهتمامات الباحثين للتنبؤ بالاسعار المستقبلية للعملات المشفرة والتي تهدف الى الاستثمار وجني الربح او دراسة طبيعة وسلوكيات هذه الاسواق.

في هذه الرسالة تم تطبيق العديد من التقنيات للتنبؤ باسعار العملات المشفرة بالاعتماد على دراسة المؤثرات الخارجية للعملات المشفرة والمتلخصة بالعرض، الطلب، منصات التواصل الاجتماعي، وتكلفة التصنيع بالإضافة الى اسعار الذهب حيث تم دراسة كل عامل من العوامل على حده لمعرفة ذلك بغرض معرفة الآلية المناسبة للحصول على اعلى دقة بالإضافة الى استخراج اهم المزايا والتي بدورها تعتبر المكون الاساسي للنموذج الاساسي والمستهدف.

لتحقيق هدف الدراسة تم تطبيق الدراسة على اربعة انواع مختلفة من العملات المشفرة وتتضمن البتكوين، الاثيريوم، اللات كوين بالإضافة الى الدوج كوين، حيث تم اخيار هذه العملات بسبب انها الاكثر شعبية بالمقارنة مع العملات الاخرى، حيث تم اطلاق اسم القناة (Channel) على كل عامل من العوامل المكونة لبيئة العملات الرقمية بحيث تم تحليل كل قناة على حدة لاستخراج افضل الميزات التي ستشارك في بناء النموذج النهائي، فعند تحليل منصات التواصل الاجتماعي وتحديداً منصة التواصل الاجتماعي تويتر (Twitter) تم تطبيق تقنيتين لتحليل المشاعر على النصوص المستخرجة من تويتر وهي تحليل المشاعر باستخدام VADER وتحليل المشار بالاعتماد على Harvard IV-4 حيث تم تطبيق نفس التقنيات لتحليل المشاعر على الاخبار وتم تقسيم البيانات الناشئة من التحليل وتدريب نموذج تعلم الآلة LSTM وتم التوصل بان النموذج المنشأ من خلال بيانات تحليل المشاعر VADER فيما يخص التنبؤ بالاسعار الرقمية من خلال منصات التواصل الاجتماعي افضل من النموذج المبني على اساس البيانات المستخرجة من Harvard

IV-4 و على العكس تمام فيما يخص تحليل البيانات في القنوات الاخبارية حيث كانت النتائج تفوق Harvard IV-4 على VADER.

بعد تقسيم البيانات الناتجة من كلا النموذجين وتدريب نموذج LSTM. تم الاعتماد على بيانات البحث من جوجل Google Trend كقناة ثالثة مشاركة في بناء النموذج الاساسي، بالاضافة الى مجموعة بيانات تتمثل تكلفة التصنيع مثل ما يطلق عليه صعوبة التعدين بالاضافة الى حجم المعاملات الخاصة بعملية تبادل العملات الرقمية.

تعتبر عملية جمع البيانات المشاركة في بناء النموذج الاساسي من اكثر العمليات الاكثر تعقيدا حيث تم جمع اكثر من 220 مليون سجل من منصة التواصل الاجتماعي تويتر وتم معالجة هذا الكم من البيانات من خلال ابتكار طريقة جديدة في التعامل مع البيانات وهي عملية تقسيم الملفات ومعالجتها وتحليلها على حدة ومن ثم تجميع البيانات المحللة لتقسيمها وتغذية النموذج بها. كما نود الاشارة الى ان عملية تنظيف البيانات وخصوصا المجمعة من قبل منصة التواصل الاجتماعي قد خضعت للعديد من المراحل وذلك لتكون جاهو وقادرة على ان تكون بيانات نظيفة ومهيكله وصالة لتغذية نموذج التعلم الالي بها، خضعت عملية تنظيف البيانات الى سلسلة من المعالجة تضمنت حذف السجلات المكررة بالاعتماد على الرقم المرجعي للحقل بالاضافة الى ازالة علامات الترقيم والروابط وايضا علامات التوقف (Stop Words).

تتضمن عملية دمج البيانات دمج البيانات المكونة للنموذج من عدة مصادر تم تسميت كل مصدر من المصادر باسم القناة حيث يتكون النظام من ست قنوات والتي تغطي كافة المؤثرات الداخلية والخارجية والتي تؤثر على اسعار العملات الرقمية ودمجها لتموين قناة واحد لبناء نموذج التنبؤ بالعملات الرقمية القائم على اساس دمج البيانات، تشير النتائج النهائية على تحسن ملحوظ للتنبؤ بالعملات الرقمية للنموذج المبني على اساس دمج البيانات على النماذج الاخرى المبينة على اساس القناة الواحدة حيث تم تقييم النموذج من خلال مقياس متوسط الخطأ المطلق (MAPE) والذي من خلاله يتم قياس دقة التنبؤات، حيث كانت النتيجة النهائية للنموذج للعملات المشفرة الثلاث وهي البتكوين والايثيريوم واللايت كوين هي 5.62% و 3.48% و 2.88% على التوالي حيث كان النموذج الأفضل هو التنبؤ بالعملة الرقمية اللايت كوين والمبني على اساس عملية دمج البيانات مع الاشارة بان النماذج الاخرى نماذج مرضية ويمكن الاعتماد عليها في التنبؤ باسعار العملات الرقمية. في العمل القادم سيتم العمل على بناء النموذج المقترح بالاعتماد على خوارزميات تعلم الالية مختلفة والمقارنة بينها بالاضافة الى تدريب النماذج على بيانات جديدة سيتم جمعها من خلال الطرق المقترحة.