VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belgaum-590018

A PROJECT REPORT **(15CSP78)** ON

"CRYPTOCURRENCY PREDICTION"

Submitted in Partial fulfillment of the Requirements for the VIII Semester of the Degree of

Bachelor of Engineering in Computer Science & Engineering

By

RISHAV PANDEY (1CR16CS133)

SNEHIL ARYA (1CR16CS162)

MRIDUL SINGH (1CR16CS095)

MRIGANK CHAUHAN (1CR16CS096)

Under the Guidance of,

Mr. Kiran Babu T S

Associate Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI,BANGALORE-560037

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **"Cryptocurrency Prediction"** carried out by Mr. **Mridul Singh**, USN **1CR16CS095**, Mr. **Mrigank Chauhan**, USN **1CR16CS096**, Mr. **Rishav Pandey**, USN **1CR16CS133**, Mr. **Snehil Arya**, USN **1CR16CS162**, bonafide students of CMR Institute of Technology, in partial fulfillment for the award of **Bachelor of Engineering** in Computer Science and Engineering of the Visvesvaraya Technological University, Belgaum during the year 2019-2020. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library.

The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Karan Babu T S,	**Dr. Prem Kumar Ramesh**	**Dr. Sanjay Jain**
Associate Professor	**Professor & Head**	**Principal**
Dept. of CSE, CMRIT	**Dept. of CSE, CMRIT**	**CMRIT**

External Viva

Name of the examiners	Signature with date
1.	________________
2.	________________

DECLARATION

We, the students of 8th semester of Computer Science and Engineering, CMR Institute of Technology, Bangalore declare that the work entitled "Cryptocurrency prediction" has been successfully completed under the guidance of Kiran Babu TS, Computer Science and Engineering Department, CMR Institute of technology, Bangalore. This dissertation work is submitted in partial fulfillment of the requirements for the award of Degree of Bachelor of Engineering in Computer Science and Engineering during the academic year 2019 - 2020. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to any university.

Place: Bengaluru

Date: 10-06-2020

Team members:

MRIDUL SINGH(1CR16CS095)

MRIGANK CHAUHAN (1CR16CS096)

RISHAV PANDEY(1CR16CS133)

SNEHIL ARYA(1CR16CS162)

ABSTRACT

In this project, we attempt to predict the Bitcoin price accurately taking into consideration various parameters that affect the Bitcoin value. For the first phase of our investigation, we aim to understand and identify daily trends in the Bitcoin market while gaining insight into optimal features surrounding Bitcoin price. Our data set consists of various features relating to the Bitcoin price and payment network over the course of five years, recorded daily. For the second phase of our investigation, using the available information, we will predict the sign of the daily price change with highest possible accuracy.

Twitter is increasingly used as a news source increasing purchase decisions by informing users of the currency and its increasing popularity. As a result, quickly understanding the impact of tweets on price direction can provide a purchasing and selling advantage to a cryptocurrency user or a trader. By analyzing tweets, we found that tweet volume, rather than tweet sentiment(which is invariably overall positive regardless of price direction), is a predictor of price direction.

Speaking about the logic that will be used for the retrieval of results, we will be using several machine learning algorithms like RNN with LSTM model.

Keywords: Cryptocurrency, Machine Learning, LSTM, Bitcoin, Twitter sentiment.

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude and respect to **CMR Institute of Technology, Bengaluru** for providing me a platform to pursue my studies and carry out my final year project.

I have a great pleasure in expressing my deep sense of gratitude to **Dr. Sanjay Jain,** Principal, CMRIT, Bangalore, for his constant encouragement.

I would like to thank **Dr. Prem Kumar Ramesh,** HOD, Department of Computer Science and Engineering, CMRIT, Bangalore, who has been a constant support and encouragement throughout the course of this project.

I consider it a privilege and honor to express my sincere gratitude to my guide
Mr. Kiran Babu TS, Associate Professor, Department of Computer Science and Engineering, for the valuable guidance throughout the tenure of this review.

I also extend my thanks to all the faculty of Computer Science and Engineering who directly or indirectly encouraged me.

Finally, I would like to thank my parents and friends for all their moral support they have given me during the completion of this work.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

As the economic and social impact of cryptocurrencies continues to grow rapidly, so does the prevalence of related news articles and social media posts, particularly tweets. As with traditional financial markets, there appears to be a relationship between media sentiment and the prices of cryptocurrency coins. While there are many causes of cryptocurrency price fluctuation, it is worthwhile to explore whether sentiment analysis on available online media can inform predictions on whether a coin's price (i.e., perceived value) will go up or down.

The input of our system is text data from news headlines and tweets, aggregated by day and kept in order of occurrence to preserve the time-series nature. Traditional supervised learning binary classification algorithms were then used to assign each news headline and tweet a label of 0 or 1 for each coin, indicating predictions of a price decrease or price increase one day in the future, respectively. The majority label for each coin, on each day, was then used as the final daily prediction.

By May 2018, the two largest cryptocurrencies, measured in terms of market capitalization, had a combined market value of 160.9 billion dollars1. Bitcoin alone made up nearly $115 billion of this value. Given the significant value of these currencies, some people see value in them through use as actual currencies, while others view them as investment opportunities. The result has been large swings in value of both currencies over short periods of time. During 2017 the value of a single Bitcoin increased 2000% going from $863 on January 9, 2017 to a high of $17,550 on December 11, 2017. By eight weeks later, on February 5, 2018, the price of a single Bitcoin had been more than halved with a value of $7,9643. The promising technology behind cryptocurrencies, the blockchain, makes its likely that cryptocurrencies will continue to be used in some capacity, and that their use will grow.

In addition to that, Bitcoin's dominance in market capitalization over the cryptocurrency market has gradually faded from 85% in 2010 to 50% today, showing that an overall attraction to the cryptocurrencies has taken place in the last couple of years.

Lately, as Bitcoin spirals to new lows every day in the year 2018, while dragging the entire cryptocurrency market down with it, market participants are becoming increasingly interested in the factors that lead to such downturns to understand the price dynamics of these digital cryptocurrencies. However, from the perspective of a cryptocurrency trader, whether the prices going up or down is no problem as long as the direction is predictable. In the chase of an expected boom period, investors can take a long position in cryptocurrencies beforehand to realize their returns once the prices reach up to a certain level. Whereas in the case of a bust period foreseen in the future, investors can short sell these cryptocurrencies through margin trading (allowed by many cryptocurrency exchanges) to gain excess returns. Moreover, taking long or short positions has become much easier after the action taken by the CBOE in December 2017 when they introduced Bitcoin futures. Such a financial asset provides investors to speculate on Bitcoin prices in both directions through leverage without even holding Bitcoins. Similar strategies can be implemented lately for other cryptocurrencies through binary options traded in the offshore exchanges.

All these anecdotes lead us to the question of whether cryptocurrency prices are predictable? In other words, does Efficient Market Hypothesis (EMH) hold for cryptocurrencies? In an efficient market (Fama, 1970), any past information should already be reacted into the current prices so that prices might be effected by only future events. However, since the future is unknown, prices should follow a random walk (or a martingale process, to be precise). In the special case of the weak-form efficiency, future returns can not be predicted on the basis of past price changes, however, since the earlier works by Mandelbrot (1971) and others (Fama and French (1988); Lo and Mackinlay (1988); Poterba and Summers (1988); Brock et al. (1992); Cochran et al. (1993)), weak-form of EMH has been found to be violated in various types of asset returns,1 which in turn leads to important problems such as: i) preferred investment horizon being a risk factor (Mandelbrot, 1997); and ii) the fail of common asset pricing models such as CAPM or APT, or derivative pricing models like the Black-Sholes-Merton model (Kyaw et al. (2006); Jamdee and Los (2007)).

As evident by the discussions above, EMH has been an intriguing subject for both academicians and market professionals for a very long time, and naturally, the efficiency of cryptocurrencies (especially of Bitcoin) has gained immediate interest due to this fact. For example, the pricing efficiency of Bitcoin has been studied extensively in the last couple of years in various academic papers: Urquhart (2016) provides the earliest evidence on the status of market efficiency for Bitcoin and concludes that Bitcoin is not weakly efficient, however it has a tendency of becoming weakly efficient over time. Building upon that, Nadarajah and Chu (2017) run various weak-form efficiency tests on Bitcoin prices via power transformations and state that Bitcoin is mostly weak-form efficient throughout their sample period. Excessive amount of studies followed on the same topic with differentapproaches in methodologies, sample frequency, benchmark currency etc (Bariviera (2017); Vidal-Tomas and Ibanez (2018); Jiang et al. (2018); Tiwari et al. (2018); Khuntia and Pattanayak (2018); Sensoy (2018)). Against all different perspectives, the main conclusion is that Bitcoin is inefficient, but to gain weak-form efficiency over time. Even though the related literature on Bitcoin is scarce, other cryptocurrencies has attracted relatively less attention. Brauneis and Mestel (2018) investigate the weak-form efficiency of cryptocurrencies in the cross-section, and show that liquidity and market capitalization has a significant effect on the pricing efficiency. In another study, Wei (2018) analyzes the return predictability of 456 cryptocurrencies and concludes that there is a strong negative relationship between return predictability with cryptocurrency liquidity.

According to our view, the problem with the abovementioned literature is that there are vast amount of studies on the pricing efficiency of cryptocurrencies (especially Bitcoin), with almost all of them rejecting the null hypothesis of weak-form efficiency. However, all those studies rely on common statistical tests and provide no explicit way of exploiting this inefficiency nor state the potential excess gains that could be obtained through consistent active trading.

In this study, we aim to fulfil this gap in the literature. Using returns obtained at various intraday frequencies for the most liquid twelve cryptocurrencies, we test their return predictability via several methods including support vector machines, logistic regression, artificial neural networks and random forest classification algorithms. Naturally, our contribution to the literature is many fold: First, unlike the previous studies that mostly focus on only Bitcoin, we cover a sample of twelve cryptocurrencies. This helps us to understand the overall price dynamics of the cryptocurrency market rather than a single digital currency.

Second, previous studies usually use daily data, whereas we cover a sample ranging from a few minutes to daily frequency. This is especially important since in the current status of the financial markets, algorithmic (especially high-frequency) trading is implemented actively, with average asset holding periods barely extend over a few minutes (Glantz, 2013). Several cryptocurrency exchanges provide algorithmic trading connections to their customers which makes it essential to analyse the cryptocurrency markets pricing efficiency at the intraday level (Sensoy, 2018). Third, rather than using the common statistical methodologies to test the pricing efficiency, we refer to the state of the art methodologies used in the decision sciences that provide us the potential patterns to be exploited and the resulting gains if the selected strategy is implemented. Finally, the use of many cryptocurrencies and different timescales the set of features utilized for prediction can be easily verified in their ability to generalize in different timescales for different cryptocurrencies. This is particularly important since most studies on using machine learning algorithms for forecasting consider a single asset at a single timescale without showing the potential of generalization ability of algorithms in different markets and timescales.

1.1 Relevance of the Project

The current market of cryptocurrency is too unstable. Investing in it may lead to high risks. Before the investment, the investor has to consider many factors like the rules of the country to which he/she belongs. Except it, prices are driven by market news, views and opinions, prediction from respected figures of space and several other factors.

With such unpredictable events, the price can rise fast and plummet equally fast. Unlike in traditional market where risk-averse investors are at peace, the crypto market involves a lot of risks.

1.2 Problem Statement

1.2.1 Goal

The overall goal of the project is to construct a machine learning model that can predict price trends with results superior to that of random selection. The purpose of this study is to find out with what accuracy the direction of the price of Bitcoin can be predicted using machine learning methods. This is fundamentally a time series prediction problem. While much research exists surrounding the use of different machine learning techniques for time series prediction, research in this area relating specifically to Bitcoin is lacking. In addition, Bitcoin as a currency is in a transient stage and as a result is considerably more volatile than other currencies such as the USD. Interestingly, it is the top performing currency four out of the last five years. Thus, its prediction offers great potential and this provides motivation for research in the area. As evidenced by an analysis of the existing literature, running machine learning algorithms on a GPU as opposed to a CPU can offer significant performance improvements. This is explored by benchmarking the training of the RNN and LSTM network using both the GPU and CPU. This provides an answer to the sub research question. Finally, in analysing the chosen dependent variables, each variables importance is assessed using a random forest algorithm. This body of research builds on existing literature in the area which is assessed in section 2.

In addition, the ability to predict the direction of the price of an asset such as Bitcoin offers the opportunity for profit to be made by trading the asset. To implement a full trading strategy based on the results of the models is worthy of a dissertation in itself and as a result this paper will focus solely on the accuracy at which the price direction can be predicted. In basic terms, the model would initiate a short position if the price was predicted to go up and a long position if the price was predicted to go down. Several Bitcoin exchanges offer margin trading accounts to facilitate this. The profitability of this strategy would be based not only on the accuracy of the model, but also on the size of the positions taken. This is outside the scope of this research but could be addressed in future work.

1.2.2 Comparison with Existing System

Existing Systems make use of Machine learning on a given data set to predict the price, instead this project is using sentiment analysis also for better and closer results.

1.2.3 Solution/Implementation

The proposed Solution is an algorithm which analyse the given data set and consider sentiments for predicting the price of cryptocurrency.

1.2.4 Impact

This project would help the user and investors so that they can invest and use it more efficiently.

1.3 Summary

In this chapter, we have discussed about what is cryptocurrency, its importance in the economy according to research. We have also discussed about the scope and the relevance of our project in the current trends. The main objectives of this paper is discussed in this chapter.

CHAPTER 2

LITERATURE SURVEY

Various approaches have been used in the past to carry out the price prediction task. There are mainly two sets of literature that are highly relevant to this work. One is financial data analysis; the other, time series data analysis.

2.1 Financial Data Analysis

Several approaches are described in the literature including, one called technical analysis also known as “charting” that forecasts future prices (Lo et al., 2000). According to it, stock market prices do not follow random walks, that is – the price movements follow a set of patterns. These price movements can be used to predict the future price (Lo & MacKinlay, 1988, 1999). There exist some other empirically designed patterns such as heads-and-shoulders, double-top-and-bottom that can be used to predict future prices. We refer the interested reader to the work of Lo et al. (2000). In this paper, authors have used kernel based regression techniques to find out the patterns in historical data, that is – price is predicted based on past data. This work (Lo et al., 2000) is theoretically close to the current project work. However, it does not employ the same strategies followed in the current project.

2.2 Time Series Data Analysis

In the context of future price predictions, classical methods are quite popular. Autoregressive integrated moving average (ARIMA) models are a popular choice for forecasting over a short term. It works very well when the data exhibits consistent or stable pattern over time with least possible outliers. The ARIMA methodology works well only when the data exhibits “stationarity”, which means that the series remains almost constant. But this is not always possible in the real time scenario, where the data fluctuates drastically, and it is highly volatile. Ediger and Akar used the seasonal ARIMA model to estimate the future fuel energy demand in Turkey over certain years. However, the similar scenario is not guaranteed to work for unseasonal or non-linear data. To solve the real time prediction problems, artificial neural networks are very much useful to increase the speed of computation due to its ability to handle nonlinearities in the data. Examples of nonlinear data include psychological data (Scheier & Tschacher, 1996).

In one of their research papers, Greaves et al., 2015 predicted the price of Bitcoin using support vector machine(SVM) and artificial neural networks (ANNs), concluded that probability of predicting the price in block chain market is challenging and is scope is limited. Despite of all this, neural networks have become a valuable tool for prediction of time series problems due to their ability to handle non-linear data and nonstationary data.

2.3 An Assembly Stock Predictor and Recommender System

As the popularity of cryptocurrencies, and in particular Bitcoin, increased over the years, more studies shifted their focus from the stock market towards the cryptocurrency market.

Articles by Kaminski (2014) and Matta et al. (2015) get close, in terms of methodologies and research questions, to the research that has been done earlier for the stock market.

The studies use Twitter data to analyse relationships between Bitcoin market indicators and Twitter posts containing emotional signals.

The studies find significant correlations between emotional tweets and the closing price, trading volume and intra-day price spread of Bitcoin.

Additionally, the relationship of Google search queries with Bitcoin trading volumes is investigated to identify the impact of search frequencies on cryptocurrency markets (Matta et al., 2015).

Past research has mostly focused on classifying user comments in particular fields. Comments on online communities involve considerable use of neologisms, slang, and emoticons that transcend grammatical usage. C.J. Hutto and Eric Gilbert introduced an algorithm called VADER [44]to parse such expressions, and proposed a method to analyse social media texts by drawing on a rule-based model.

2.4 Source Used

Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and RepliesYoung Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, Chang Hun Kim.

Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis by Jethin Abraham, Daniel Higdon, John Nelson, Juan Ibarra Southern Methodist University.

2.5 Summary

In this chapter we mainly focused on existing system through literature survey and various IEEE paper analysed and we specified some important points of each paper and related diagrams or graphs are included. In comparison section we have mainly highlighted few important advantages and disadvantages in each paper and comparison between those papers. This chapter also introduces drawbacks of existing system and functionality of proposed system.

CHAPTER 3

OBJECTIVE AND METHODOLOGY

The analysis detailed later in this paper requires an understanding of where and why the data was collected, and how cryptocurrencies may vary from standard fait currencies or stocks in companies from traditional stock markets. In this section we will provide more background on these data sources and why they were chosen so that the final analysis is put in the proper context for the reader.

3.1 Blockchain Technology and Cryptocurrency

In this paper we analyse data about the world's two largest cryptocurrencies in terms of market capitalization. The largest is Bitcoin followed by Ethereum. Bitcoin was the first cryptocurrency ever created. The creation of Bitcoin is mysterious as it was created by a person or group of people using the name "Satoshi Nakamoto" and released in 20093. Along with the launch of Bitcoin "Satoshi Nakamoto" published a paper titled "Bitcoin: A Peer-to-Peer Electronic Cash System" which described a peer-to-peer payment system using electronic cash (cryptocurrencies) that could be sent directly from one party to another without the use of a third party to validate the transaction. This innovation is created by the use of the "blockchain" which is like a shared ledger on the peer-to-peer network where all transactions are veri_ed by the network so they cannot be forged.

The applications of blockchain technology go beyond peer-to-peer payment systems. Blockchain technology provides security, privacy, and a distributed ledger which makes them applicable for internet-of-things applications, distributed storage systems, healthcare, and more. The range of applications of the blockchain has led to many more blockchains and cryptocurrencies being created (1,658 cryptocurrencies are in existence4). Cryptocurrencies are tied to the blockchain because they provide the incentive for machines, and the electricity they consume, to run and validate the blockchain. As use of blockchains increases so too will the use of cryptocurrencies. This gives them an inherent value, but what that value is depends on many factors. Because this is a new type of currency, and store of value, improving the understanding of what can lead to price changes brings with it value.

3.2 Twitter

Twitter was launched in July of 2006 as an application in both the social mediaspace (which includes other applications/websites such as Instagram, Facebook, LinkedIn and others) and microblogging. Microblogging is a medium that allows for smaller and more frequent updates than blogging5. Twitter allows users to post messages publicly (which are referred to as "tweets") with a maximum length of 140 characters. In November of 2017 that limit was doubled to 280 characters. In addition, users can add "hashtags" to a tweet, which is the "#" symbol followed a consecutive string of characters. This is used to identify the topic or theme of a tweet and to make them searchable. This is used later when we collect the tweets in the data section.

Since its launch in 2006 Twitter has grown rapidly in popularity. One of the early examples of its reach and power was on January 15, 2009 when a US Airways flight crashed into the Hudson river. An image posted to Twitter broke the news before traditional media outlets did7. Twitter has 330 million monthly active users, 1.3 billion accounts have been created, 83% of the world's leaders have a Twitter account, approximately 23 million of Twitter's active users are actually bots rather than humans, and 500 millions tweets are sent each day8. The result of all of these impressive statistics is that Twitter can be a very rich source of data on how people feel about nearly any given topic. With the ability to see when a tweet was posted it is also possible to tell how those feelings change over time. This makes Twitter an excellent resource to collect text data on a topic like cryptocurrencies to explore the possible relationships between that and prices.

3.3 Sentiment Analysis

It is estimated that 90% of the data in the world has been created in the last two years. Much of that data is in the form of unstructured text data whether it be in the form of tweets, articles posted to the internet, text messages, emails, or other forms. This vast amount of unstructured data has led to the creation of "natural language processing" (NLP) as an area of study or development. NLP is a collection of methods for computers to analyse and understand text. In this paper we use a set of natural language processing tools commonly referred to as "sentiment analysis". Sentiment analysis is the act of extracting and measuring the subjective emotions or opinions that are expressed in text. There are multiple ways to do this. We chose the "VADER" (Valence Aware Dictionary and Entiment Reasoner) [6] system in this analysis, which will be described in more detail in the methodology section. The end goal of this analysis is to apply sentiment analysis to collected tweets so that it can be determined if the tweets are generally positive or negative in their opinions of cryptocurrencies. In addition, we also want to use sentiment analysis to identify tweets which express an opinion (subjective tweets) versus those that just provide information without a positive or negative angle to them (objective tweets).

3.4 Related Work

This paper builds on the ideas of a wide range of research and topics. Behavioral economists like Daniel Kahneman and Amos Tversky established that decisions, even ones involving financial consequences, are impacted by emotion and not just value alone. R. J. Dolan's work in "Emotion, Cognition, and Behavior" further supports that decision making is highly impacted by emotions. The insights from these researchers opens up the possibilities to and advantages through tools like sentiment analysis as it indicates that demand for a good, and therefore price, may be impacted by more than its economic fundamentals.

Later researchers found specifically that purchase decisions people made were being impacted from information gathered online. Galen Thomas Panger found that Twitter sentiment correlated with people's general emotional state.

Additionally, he found that social media like Twitter tended to have a calming affect on the end-user rather than amplifying their emotional state. Chen et al. performed textual analysis on a social platform aimed at investors called "Seeking Alpha" and found that views expressed in articles posted on "Seeking Alpha" were associated with returns and could even predict earnings surprises. In a similar vein Paul Tetlock found that high levels of media pessimism of the stock market impacted trading volumes. Finally, Gartner found in a study that the majority of consumers relied on social media to influence purchase decisions.

Other researchers have specifically studied the efficacy of sentiment analysis of tweets. Kouloumpis et al. found that standard natural language processing techniques such as sentence level and document level sentiment scoring was ineffective due to the short nature of tweets and uniqueness of language used. Alexander Pak and Patrick Paroubek showed that separating tweets into positive, negative, or neutral categories could result in effective sentiment analyses. O'Connor et al. showed that the sentiment found in tweets was reactive of public opinion on various topics in national polling.

Their research identified sentiment analysis as a cost saving option versus national polling, but the implication that sentiments from tweets do accurately react the larger population's feelings on topics suggests that it could also be used to predict demand, and therefore, price changes of products. Web data beyond Twitter and social media has been a rich area of research as well. To our knowledge one of the first papers to and that web search data could be used to predict macroeconomic indicators was by Ettredge et al. where they found that searches relating to employment was associated with unemployment rates. Bordino et al. found that query volumes were correlated with trading volumes for stocks in the NASDAQ. Specific research into Google

Trends data has been done as well by Hyunyoung Choi and Hal Varian with the conclusion that simple seasonal auto-regressive models which included Google Trends data as inputs outperformed models that did not use Google Trends data by 5% to 20%. Asur et al. found that tweet volume about recently released movies accurately predicted box-office receipts.

Having established that decisions are inuenced by emotions, that social media can impact decisions, that sentiment analysis of social media can accurately react the larger population's opinions towards something, and that web search data can predict changes in macroeconomic statistics, much research built of those findings to see if they applied to the stock market. Alan Dennis and Lingyao Yuan collected valence scores on tweets about the companies in the S&P 500 and found that they correlated with stock prices. Pieter de Jong et al. analysed minute-by-minute stock price and tweet data for 30 stocks in the DOW Jones Industrial Average and found that 87% of stock returns were influenced by the tweets. However, they also looked for the inverse happening, that stock prices were influencing tweets and found little evidence for it. Bollen et al. used a self-organizing fuzzy neural network, with Twitter mood from sentiment as an input, to predict price changes in the DOW Jones Industrial average and achieved 86.7% accuracy.

With the introduction of cryptocurrencies similar work has been done to see if such methods effectively predict cryptocurrency price changes. In the paper "Predicting Bitcoin price fluctuation with Twitter sentiment analysis" by Evita Stenqvist and Jacob, the authors describe their process in which they collected tweets related to Bitcoin, and Bitcoin prices from May 11 to June 11 in 2017. Tweets were cleaned of non-alphanumeric symbols (using \#" and \@" as examples of symbols removed). Then tweets which were not relevant or determined to be too influential were removed from the analysis. The authors then used VADER (Valence Aware Dictionary and sentiment Reasoner) to analyse the sentiment of each tweet and classify it as negative, neutral, or positive. Only tweets that could be considered positive or negative were kept in the sentimental analysis. Connor Lamon et al. used sentiment of news headlines and tweets to regression performed best to classify these tweets and that they were able to correctly predict 43.9% of price increases correctly and 61.9% of price decreases. Colianni et al. collected tweets from November 15, 2015 to December 3, 2015 and used Naive Bayes and Support Vector Machines to classify tweets and achieved a 255-accuracy increase.

Finally, Shah et al. successfully established a trading strategy using historical prices and Bayesian regression analysis. Another branch of research in this area involves various applications of neural networks. Kimoto et al. used a modular neural network to create a buying and selling timing system for stocks on the Tokyo stock exchange and achieved profitability using their system with simulated stock purchases. Guresen etal. compared various neural network performance in forecasting stock exchange rates and found that a multilayer perceptron (MLP) neural network performed best. Xhu et al. used stock trading volumes as a neural network input and found that they modestly improved prediction performance over the medium and long terms.

The research presented in this paper builds on of everything above, but is unique in that we solve the problem of predicting cryptocurrency prices changes by combining web search data (in the form of Google Trends) and tweet volume as inputs into a linear model. In addition, we show why sentiment analysis is less useful in its predictive capabilities of cryptocurrencies despite its potential in other areas.

3.5 Summary

In this chapter we mainly focused on providing more background on these data sources and why they were chosen so that the final analysis is put in the proper context for the reader.

CHAPTER 4

REQUIREMENTS SPECIFICATION

4.1 Functional Requirements

The system should be able to provide these functionalities efficiently.

• Resource Visualization: The visualizations should be self-explanatory which can be easily understood by the user. There will be line plots and graphs which can be used as an effective measure while devising any new program.

• ML algorithm should be able to predict the output efficiently and accurately.

• On exceeding the critical conditions, alert should be sent to the aqua farmers.

• Predict the bitcoin price in hourly manner. i.e., predict for next hour based on past data.

4.2 Non-Functional Requirements

Non-functional requirements are requirements that specifies criteria that can be used to judge the operation of a system rather than the behaviour.

• **Usability:** System has been made user friendly by including a readme file in the program so that any user facing difficulty can refer it and easily solve there problem.

• **Scalability**: If more parameters required, it can be added easily. Number of visualizations can be increased. Currently the system predicts for hourly manner this interval can be changed accordingly.

• **Reliability:** System should give reliable predicted results.

Performance: Our LSTM model will have improved performance because of the use of datasets with lowest time intervals and has high precession. For checking the accuracy, we have shown the performance metrics using RMSE.

Documentation: Coding standards are maintained throughout the project.

Maintainability: This project has easy maintainability of the web application, can be modifiable and integrated with advanced computational and operational technologies.

4.3 Hardware Requirements

- System: Core i5 Processor
- Hard Disk :1 TB.
- GPU: for matrix calculation
- RAM: 8GB

4.4 Software Requirement

- IDE: Anaconda(with Tensor flow)
- Programming language: Python
- Library: Numpy, pandas, keras, Tensorflow

4.5 Summary

This chapter gives an insight into the functional and non-functional requirements that the system provides. It also describes the hardware and software requirements that are required for building the system.

CHAPTER 4

SYSTEM ANALYSIS AND DESIGN

4.1 System Architecture

The proposed system architecture (Figure 4.1) shows the complete working of the system starting from training the model using the collected dataset to showing the predicted result and appropriate message on the web application.

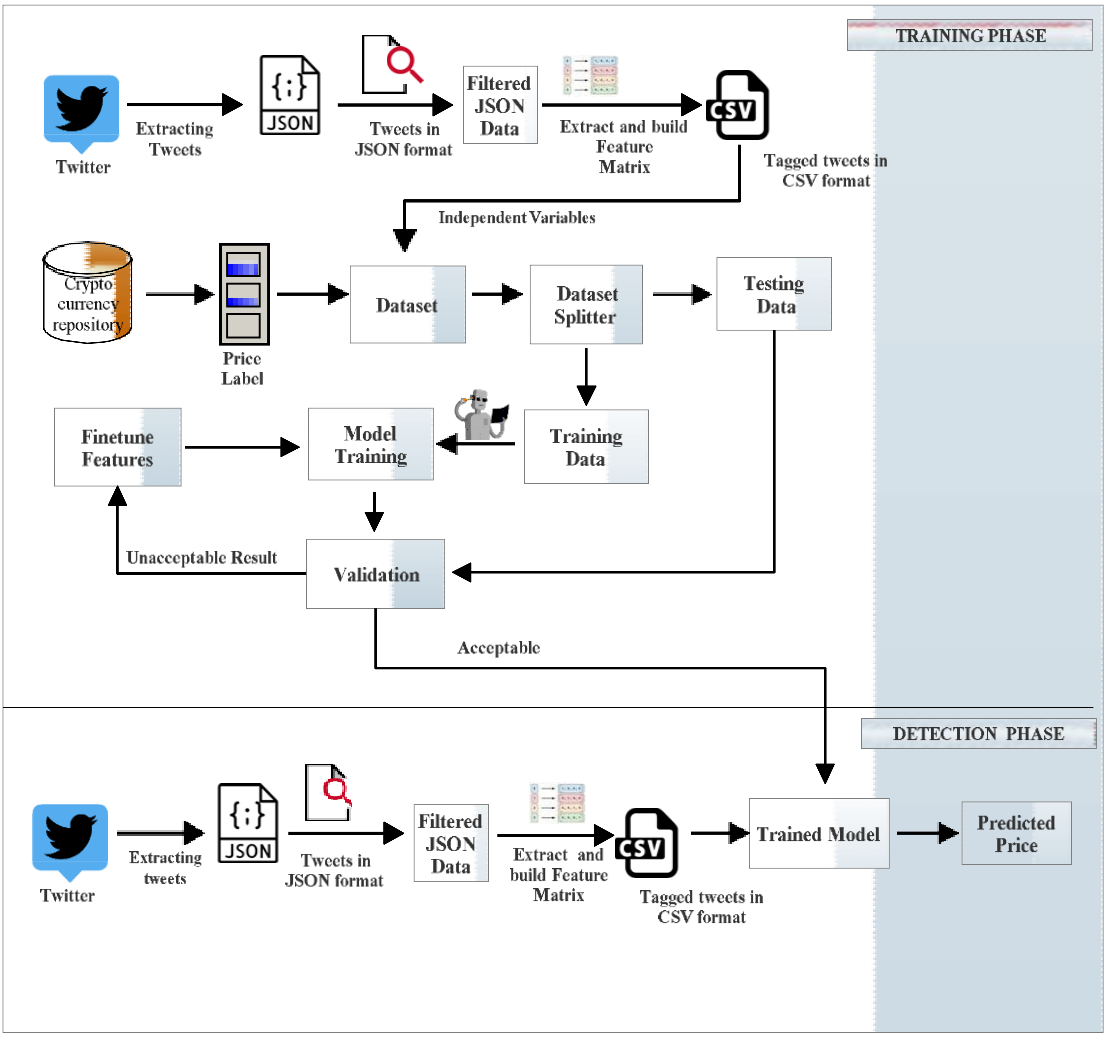


Fig 4.1 Proposed System Architecture

The proposed Architecture (Fig – 4.1) works in two phases Training phase and Detection phase. The training phase is a one-time activity. For carrying out training phase, we have collected Twitter data and the concurrent Bitcoin and Litecoin prices. The collected Twitter data and prices data are not in same format, the former being in JSON format and the latter being in the CSV format. So, in order to make synchronization in between these two, the Twitter data is converted into CSV format. The process of conversion of JSON file to CSV file is highlighted in Figure 3. The tweets in the data are analysed for their sentiment polarity. The tweets having polarity above 0 are tagged as positive tweets. The tweets having polarity equals to 0 are tagged as neutral tweets. The tweets having polarity less than 0 are tagged as negative tweets. All the tagged tweets are stored and the stored data is broken into chunks containing tagged tweets which are posted in the time duration of two hours. The number of positive tweets, neutral tweets and negative tweets present in one chunk, are counted. These counted numbers are then mapped with the average of the prices that occur in corresponding two hours' time duration. The count of positive tweets, neutral tweets, and negative tweets are the features of the dataset, and the mapped average price is the label of the dataset. Model is validated with the original labels of the given dataset. If the result of validation is acceptable, then the model is ready to be used for predicting future price by analysing real-time tweets. If not, then a new model is to be formed. The training and testing process are repeated until an acceptable model is formed. Once the acceptable model is formed, the detection phase starts. In the detection phase, real-time tweets are inputted to the model, and the model predicts the average price for the duration of two hours.

4.2 Flow Chart

The below flow chart shows the step by step execution implemented at the backend and the frontend of the system:

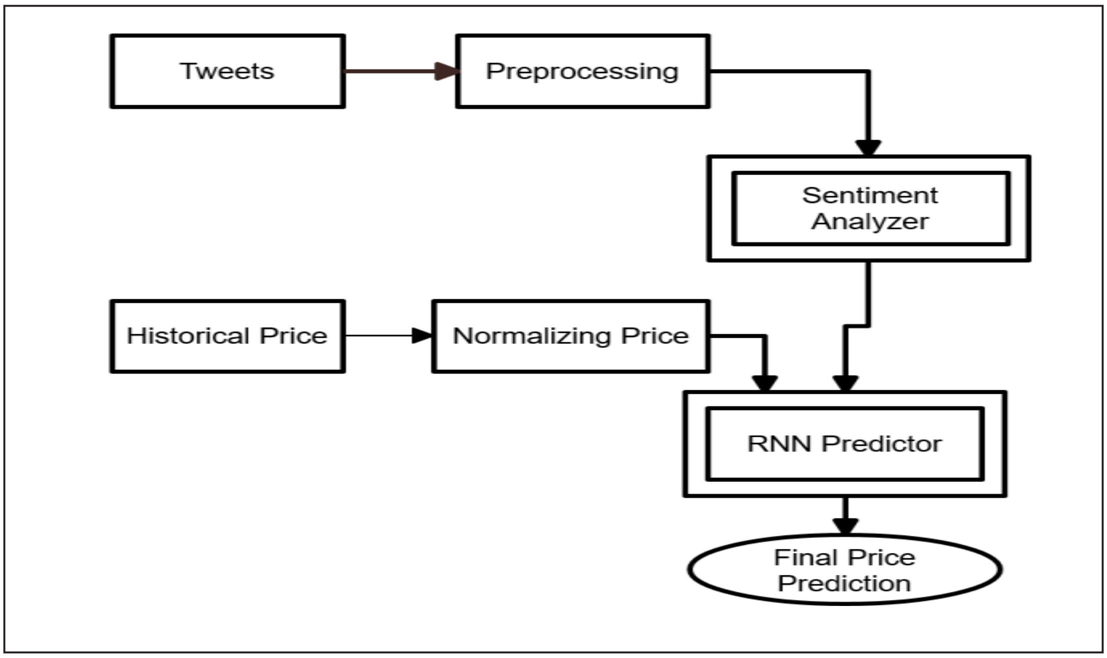


Fig:4.2-Flow Diagram

4.3 State Diagram

Figure 4.3 represents the transition between various states of the prediction system. It gives an idea about the various states and the events involved from data collection to generating an alert when any parameter is not in range.

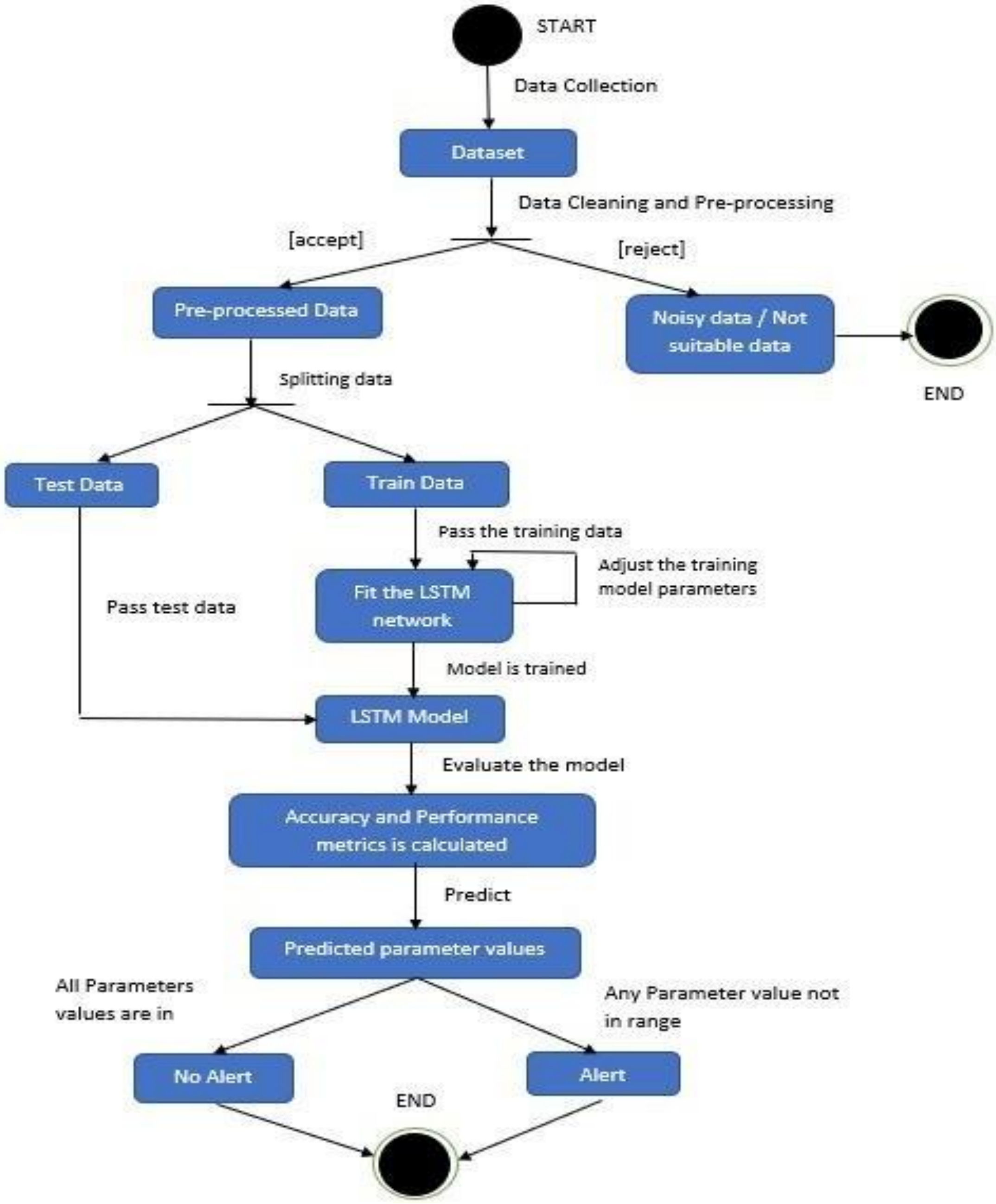


Figure 4.3 State Diagram of Cryptocurrency Prediction System

4.4 Use Case Diagram

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. Here in Figure 4.3 Supervisor is involved in the use cases View prediction result, View Visualization, View Alert and Managing the application

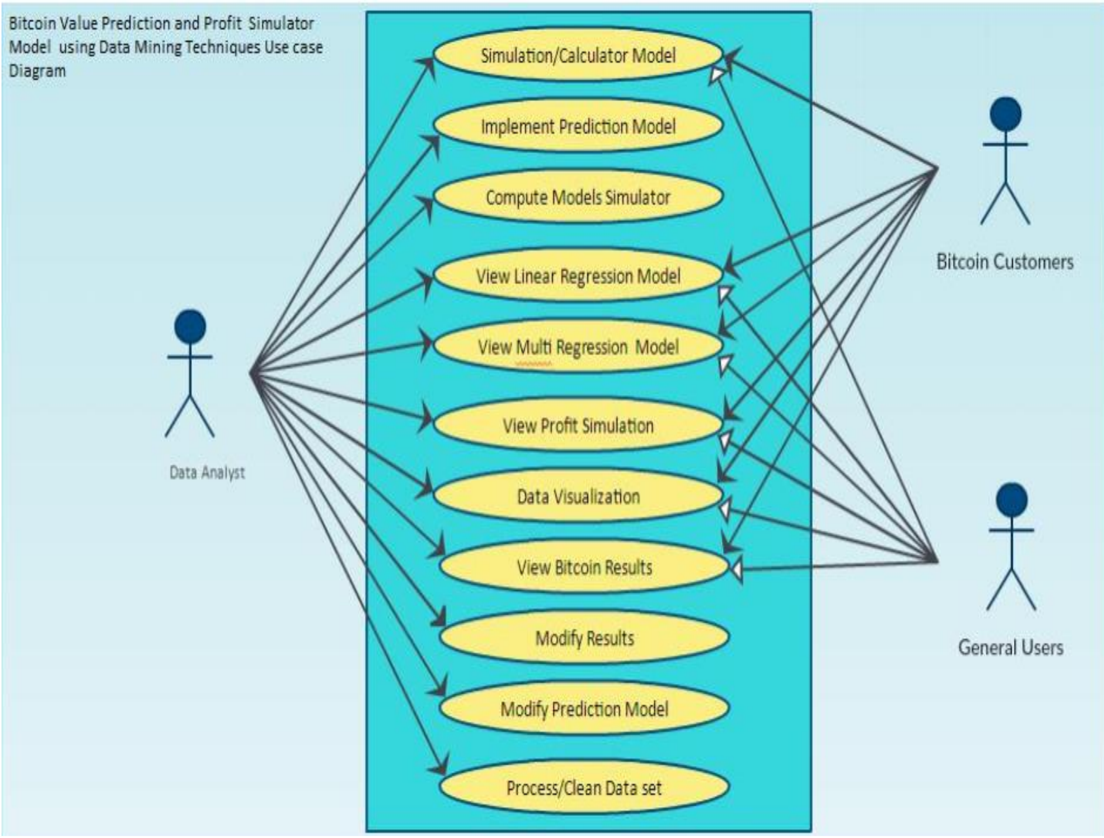


Figure 4.4 Use-Case Diagram of Cryptocurrency Prediction System

4.3 Summary

This chapter discusses the working of the system through proposed system architecture. The flow diagram shows the working of ML algorithm. The use-case diagram shows interaction between actors and the system. The sequence diagram is shown.

CHAPTER 5

ALGORITHAM

In this section we will discuss about the different algorithms we have used while developing our search engine.

5.1 RNN

Recurrent Neural Network is a generalization of feedforward neural network that has an internal memory. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it has learned from the previous input. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. In other neural networks, all the inputs are independent of each other. But in RNN, all the inputs are related to each other.

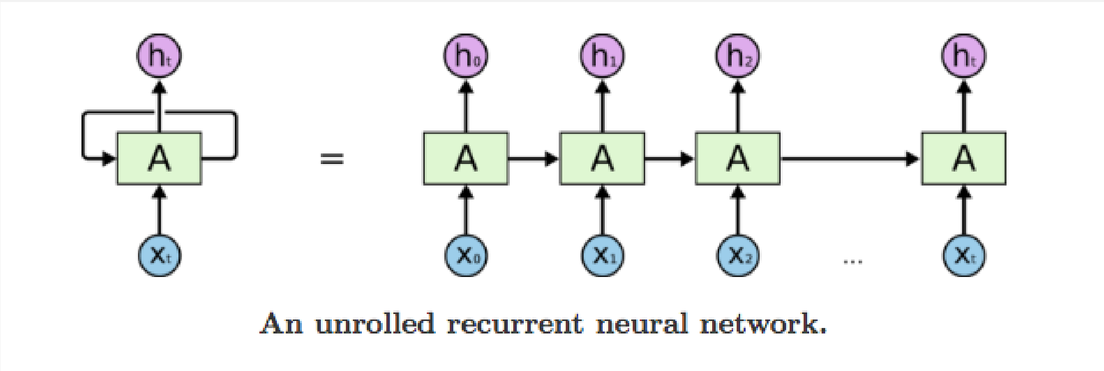


Fig 5.1 Unrolled recurrent neural network

First, it takes the X(0) from the sequence of input and then it outputs h(0) which together with X(1) is the input for the next step. So, the h(0) and X(1) is the input for the next step. Similarly, h(1) from the next is the input with X(2) for the next step and so on. This way, it keeps remembering the context while training.

The formula for the current state is

$$h_t = f(h_{t-1}, x_t) \tag{5.1}$$

Applying Activation Function:

$$h_t = \tanh (W_{hh}h_{t-1} + W_{xh}x_t) \tag{5.2}$$

W is weight, h is the single hidden vector, Whh is the weight at previous hidden state, Whx is the weight at current input state, tanh is the Activation function, that implements a Non-linearity that squashes the activations to the range[-1.1]Output:

$$y_t = W_{hy}h_t \tag{5.3}$$

Yt is the output state. **Why** is the weight at the output state.

5.2 LSTM

Long Short-Term Memory (LSTM) networks are a modified version of recurrent neural networks, which makes it easier to remember past data in memory. The vanishing gradient problem of RNN is resolved here. LSTM is well-suited to classify, process and predict time series given time lags of unknown duration. It trains the model by using back-propagation. In an LSTM network, three gates are present:

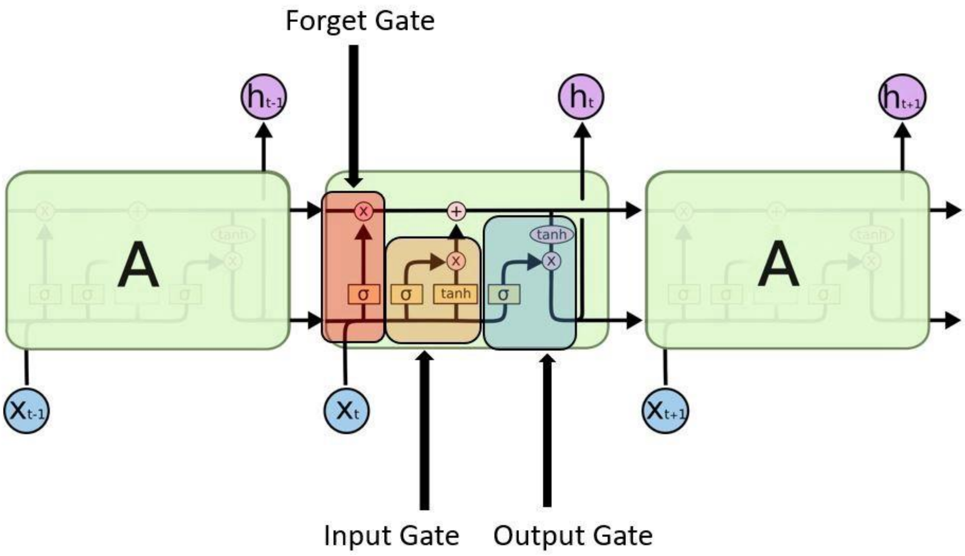


Fig 5.2 - LSTM gates

1. Input gate — discover which value from input should be used to modify the memory. **Sigmoid** function decides which values to let through **0,1.** and **tanh** function gives weightage to the values which are passed deciding their level of importance ranging from**-1** to **1.**

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C) \quad (5.4)$$

2. Forget gate — discover what details to be discarded from the block. It is decided by the **sigmoid function.** it looks at the previous state(**ht-1**) and the content input(**Xt**) and outputs a number between **0(***omit this*)and **1(***keep this***)**for each number in the cell state **Ct−1**.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right) \qquad (5.5)$$

Forget gate

3. Output gate — the input and the memory of the block is used to decide the output. **Sigmoid** function decides which values to let through **0,1.** and **tanh** function gives weightage to the values which are passed deciding their level of importance ranging from**-1** to **1** and multiplied with output of **Sigmoid.**

$$\begin{aligned} o_t &= \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right) \\ h_t &= o_t * \tanh\left(C_t\right) \end{aligned} \qquad (5.6)$$

5.3 Summary

In this section we will discuss about the different algorithms we have used while developing our search engine.

CHAPTER 6

IMPLEMENTATION

6.1 Data Representation

The used data set is a history of Bitcoin prices per minute from March 1st, 2020 to April 1st, 2020. That is to say, there are 129316 data samples. Each of them has the associated timestamp and Bitcoin price information. For this work, several fields are ignored, starting with the timestamp, since the interval is constant and it is enough knowing the order of the data, being redundant. The lowest and highest value of the minute are not considered for simplicity; there are close in value and somewhat redundant to the weighted price, which is the prediction target. The opening and closing values are also ignored for the same reason as the previous fields. That is, the only value to consider is the weighted price, which can be conceptually seen as the average price of the Bitcoin (in United States dollars, USD), for each minute.

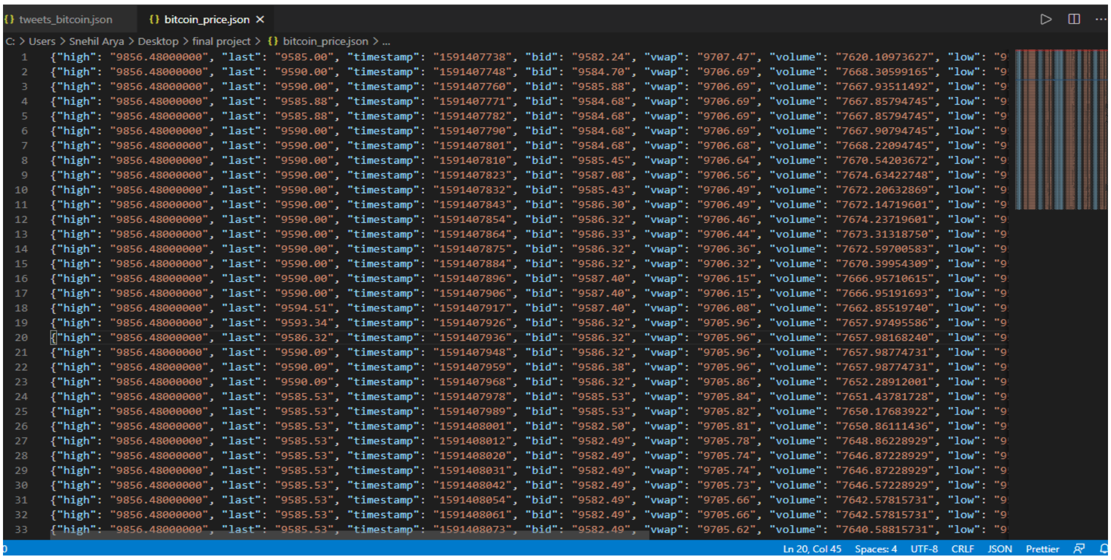


Fig 6.1 Live Bitcoin Price

6.2 Data Pre-processing

The data has been normalized using a minimum-maximum scaler, that is, translating the whole set of price values to the range 0-1, by assigning 0 to the lowest original value, 1 to the highest, and a linear equivalent to the rest of values, lying in between the extremes. Once the scaler has been fitted to the data, it can be used after producing the model to invert the transformation on the predicted values, to recover the original ranges of values for the predictions. The reason to normalize the data is to help the RNN and specifically back-propagation and gradient descent learn faster by reducing the magnitude of the value search space.

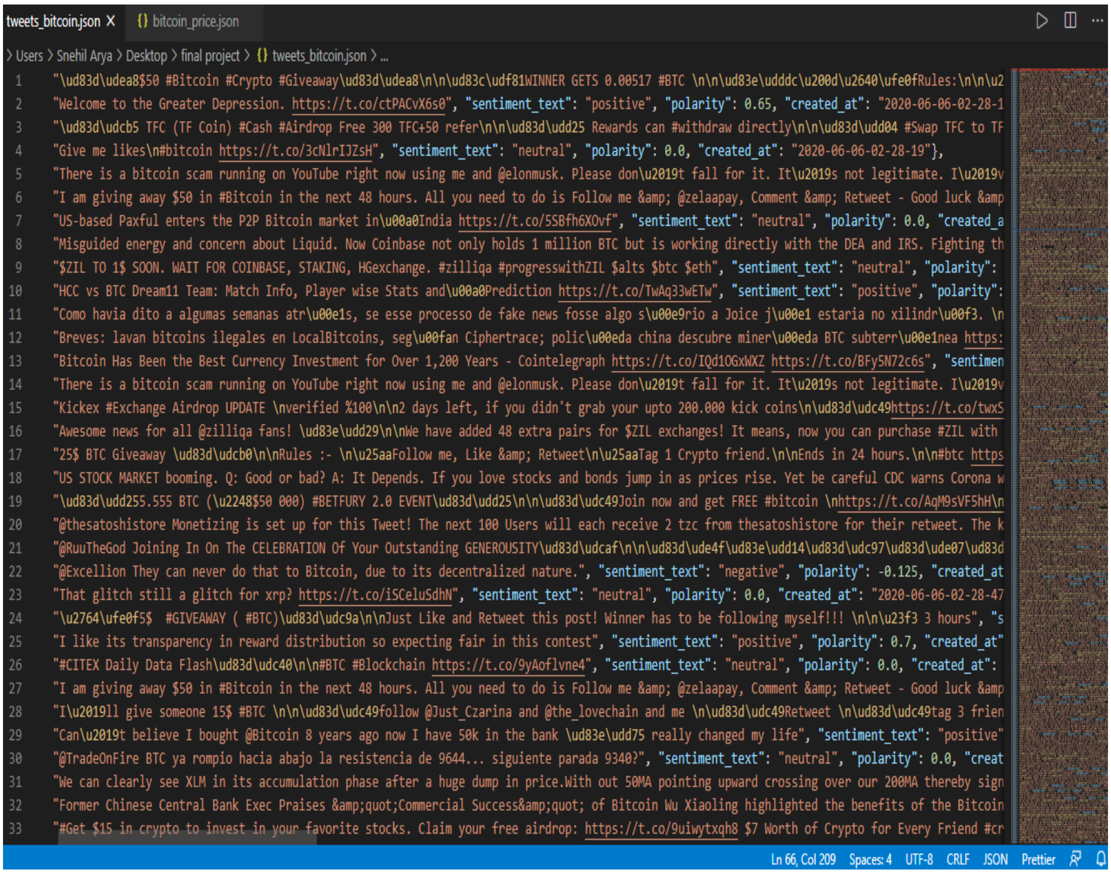
6.3 Data Split

As previously mentioned, the data is composed of 1293167 instances. The decision of how to split the data was taken trying to both have a large percentage of the data to learn (the more data, the more a model can opt to learn) and to keep a reasonably long and heterogeneous sample as the test one. Therefore, the model was trained with the first 120000 instances (92.2% of the data), while the remaining 9316 (7.8%) were used as a test. They correspond to a period of time, during the first half of 2020, with large, fast changes in the value of the Bitcoin, which make the prediction task really challenging.

6.4 Collection of tweets

For this research work, data is extracted as tweets with the name of cryptocurrencies-Bitcoin, concurrent price data which is extracted from Coindesk. For Bitcoin, data is collected from (30 days) March 2020 using the REST API5 of Twitter. For Litecoin, data is collected from March–April 2020. At the same time, per minute price of the Bitcoin is also collected. The collected tweets are obtained in JSON format, and the prices are obtained in .csv format. We tagged tweets as positive, neutral, and negative. For this, Textblob sentiment polarity is used for knowing tweet's sentiments. The value returned by "Textblob.sentiment.polarity" is in between -1 and

1. The tweets whose polarity value is 0 are tagged as neutral. The tweets whose polarity value is in between -1 and 0 are tagged as negative. The tweets whose polarity value is in between 0 and 1 are tagged as positive. After collection of tweets and prices, we have counted the number of tagged tweets in two hours duration and the final dataset comprises of the total count of positive, negative and neutral tagged tweets at the end of every two hours. We have considered the count of the tagged tweets for two hours because the efficiency of the model increases up to the second hour and after that, it starts decreasing, making two-hour duration an ideal for the consideration.



```
"\ud83d\udea8$50 #Bitcoin #Crypto #Giveaway\ud83d\udea8\n\n\ud83c\udf81WINNER GETS 0.00517 #BTC \n\n\ud83e\udddc\u200d\u2640\ufe0fRules:\n\n\u2
"Welcome to the Greater Depression. https://t.co/ctPACvX6s0", "sentiment_text": "positive", "polarity": 0.65, "created_at": "2020-06-06-02-28-1
"\ud83d\udcb5 TFC (TF Coin) #Cash #Airdrop Free 300 TFC+50 refer\n\n\ud83d\udd25 Rewards can #withdraw directly\n\n\ud83d\udd04 #Swap TFC to TF
"Give me likes\n#bitcoin https://t.co/3cNlrIJZsH", "sentiment_text": "neutral", "polarity": 0.0, "created_at": "2020-06-06-02-28-19"},
"There is a bitcoin scam running on YouTube right now using me and @elonmusk. Please don\u2019t fall for it. It\u2019s not legitimate. I\u2019v
"I am giving away $50 in #Bitcoin in the next 48 hours. All you need to do is Follow me &amp; @zelaapay, Comment &amp; Retweet - Good luck &amp
"US-based Paxful enters the P2P Bitcoin market in\u00a0India https://t.co/5SBfh6XOvf", "sentiment_text": "neutral", "polarity": 0.0, "created_a
"Misguided energy and concern about Liquid. Now Coinbase not only holds 1 million BTC but is working directly with the DEA and IRS. Fighting th
"$ZIL TO 1$ SOON. WAIT FOR COINBASE, STAKING, HGexchange. #zilliqa #progresswithZIL $alts $btc $eth", "sentiment_text": "neutral", "polarity":
"HCC vs BTC Dream11 Team: Match Info, Player wise Stats and\u00a0Prediction https://t.co/TwAq33wETw", "sentiment_text": "positive", "polarity":
"Como havia dito a algumas semanas atr\u00e1s, se esse processo de fake news fosse algo s\u00e9rio a Joice j\u00e1 estaria no xilindr\u00f3. \n
"Breves: lavan bitcoins ilegales en LocalBitcoins, seg\u00fan Ciphertrace; polic\u00eda china descubre miner\u00eda BTC subterr\u00e1nea https:
"Bitcoin Has Been the Best Currency Investment for Over 1,200 Years - Cointelegraph https://t.co/IQd1OGxWXZ https://t.co/BFy5N72c6s", "sentimen
"There is a bitcoin scam running on YouTube right now using me and @elonmusk. Please don\u2019t fall for it. It\u2019s not legitimate. I\u2019v
"Kickex #Exchange Airdrop UPDATE \nverified %100\n\n2 days left, if you didn't grab your upto 200.000 kick coins\n\ud83d\udc49https://t.co/twxS
"Awesome news for all @zilliqa fans! \ud83e\udd29\n\nWe have added 48 extra pairs for $ZIL exchanges! It means, now you can purchase #ZIL with
"25$ BTC Giveaway \ud83d\udcb0\n\nRules :- \n\u25aaFollow me, Like &amp; Retweet\n\u25aaTag 1 Crypto friend.\n\nEnds in 24 hours.\n\n#btc https
"US STOCK MARKET booming. Q: Good or bad? A: It Depends. If you love stocks and bonds jump in as prices rise. Yet be careful CDC warns Corona w
"\ud83d\udd255.555 BTC (\u2248$50 000) #BETFURY 2.0 EVENT\ud83d\udd25\n\n\ud83d\udc49Join now and get FREE #bitcoin \nhttps://t.co/AqM9sVF5hH\n
"@thesatoshistore Monetizing is set up for this Tweet! The next 100 Users will each receive 2 tzc from thesatoshistore for their retweet. The k
"@RuuTheGod Joining In On The CELEBRATION Of Your Outstanding GENEROUSITY\ud83d\udcaf\n\n\ud83d\ude4f\ud83e\udd14\ud83d\udc97\ud83d\ude07\ud83d
"@Excellion They can never do that to Bitcoin, due to its decentralized nature.", "sentiment_text": "negative", "polarity": -0.125, "created_at
"That glitch still a glitch for xrp? https://t.co/iSCeluSdhN", "sentiment_text": "neutral", "polarity": 0.0, "created_at": "2020-06-06-02-28-47
"\u2764\ufe0f5$  #GIVEAWAY ( #BTC)\ud83d\udc9a\n\nJust Like and Retweet this post! Winner has to be following myself!!! \n\n\u23f3 3 hours", "s
"I like its transparency in reward distribution so expecting fair in this contest", "sentiment_text": "positive", "polarity": 0.7, "created_at"
"#CITEX Daily Data Flash\ud83d\udc40\n\n#BTC #Blockchain https://t.co/9yAoflvne4", "sentiment_text": "neutral", "polarity": 0.0, "created_at":
"I am giving away $50 in #Bitcoin in the next 48 hours. All you need to do is Follow me &amp; @zelaapay, Comment &amp; Retweet - Good luck &amp
"I\u2019ll give someone 15$ #BTC \n\n\ud83d\udc49follow @Just_Czarina and @the_lovechain and me \n\ud83d\udc49Retweet \n\ud83d\udc49tag 3 frien
"Can\u2019t believe I bought @Bitcoin 8 years ago now I have 50k in the bank \ud83e\udd75 really changed my life", "sentiment_text": "positive"
"@TradeOnFire BTC ya rompio hacia abajo la resistencia de 9644... siguiente parada 9340?", "sentiment_text": "neutral", "polarity": 0.0, "creat
"We can clearly see XLM in its accumulation phase after a huge dump in price.With out 50MA pointing upward crossing over our 200MA thereby sign
"Former Chinese Central Bank Exec Praises &amp;quot;Commercial Success&amp;quot; of Bitcoin Wu Xiaoling highlighted the benefits of the Bitcoin
"#Get $15 in crypto to invest in your favorite stocks. Claim your free airdrop: https://t.co/9uiwytxqh8 $7 Worth of Crypto for Every Friend #cr
```


Fig 6.2-Bitcoin tweets

These are csv files for twitter sentiment and bitcoin coin prices which are obtained by cleaning the above JSON files.

	A	B	C
2771	-0.0082	17-12-22-00-48	
2772	-0.0082	17-12-22-00-49	
2773	-0.0082	17-12-22-00-50	
2774	-0.0082	17-12-22-00-51	
2775	-0.0082	17-12-22-00-52	
2776	-0.0082	17-12-22-00-53	
2777	-0.0082	17-12-22-00-54	
2778	-0.0082	17-12-22-00-55	
2779	-0.0082	17-12-22-00-56	
2780	-0.0082	17-12-22-00-57	
2781	-0.0082	17-12-22-00-58	
2782	-0.0082	17-12-22-00-59	
2783	-0.0082	17-12-22-01-00	
2784	-0.0082	17-12-22-01-01	
2785	-0.0082	17-12-22-01-02	
2786	1.63E-06	17-12-22-01-03	
2787	1.63E-06	17-12-22-01-04	
2788	0.001986	17-12-22-01-05	
2789	0.000398	17-12-22-01-06	
2790	0.000398	17-12-22-01-07	
2791	0.000398	17-12-22-01-08	
2792	0.000398	17-12-22-01-09	
2793	0.000398	17-12-22-01-10	

Twitter Sentiment

```
20200112,8021.49
20200115,8842.42
20200118,8900.34
20200121,8626.47
20200124,8388.11
20200127,8588.42
20200130,9279.81
20200202,9378.09
20200205,9162.14
20200208,9807.54
20200211,9854.79
20200214,10242.43
20200217,9937.67
20200220,9604.72
20200223,9669.63
20200226,9309.15
20200229,8712.35
20200303,8912.82
20200306,9067.39
20200309,8039.38
20200312,7936.65
20200315,5166.26
20200318,5357.61
20200321,6226.44
20200324,6502.16
```

Bitcoin Price

6.5 Model Code

Training a simple random model:

In [5]:
```
from sklearn.preprocessing import MinMaxScaler
values = datag['Price'].values.reshape(-1,1)
sentiment = datag['Sentiment'].values.reshape(-1,1)
values = values.astype('float32')
sentiment = sentiment.astype('float32')
scaler = MinMaxScaler(feature_range=(0, 1))
scaled = scaler.fit_transform(values)
```

In [6]:
```
train_size = int(len(scaled) * 0.7)
test_size = len(scaled) - train_size
train, test = scaled[0:train_size,:], scaled[train_size:len(scaled),:]
print(len(train), len(test))
split = train_size
```

```
2523 1082
```

In [7]:
```
def create_dataset(dataset, look_back, sentiment, sent=False):
    dataX, dataY = [], []
    for i in range(len(dataset) - look_back):
        if i >= look_back:
            a = dataset[i-look_back:i+1, 0]
            a = a.tolist()
            if(sent==True):
                a.append(sentiment[i].tolist()[0])
            dataX.append(a)
            dataY.append(dataset[i + look_back, 0])
    #print(len(dataY))
    return np.array(dataX), np.array(dataY)
```

Fig 6.3-Training sample model

with lookback = 1 (kind of unigram)

In [30]:
```
look_back = 1
trainX, trainY = create_dataset(train, look_back, sentiment[0:train_size])
testX, testY = create_dataset(test, look_back, sentiment[train_size:len(scaled)])
```

In [31]:
```
trainX = np.reshape(trainX, (trainX.shape[0], 1, trainX.shape[1]))
testX = np.reshape(testX, (testX.shape[0], 1, testX.shape[1]))
```

In [32]:
```
model = Sequential()
model.add(LSTM(100, input_shape=(trainX.shape[1], trainX.shape[2])))
model.add(Dense(1))
model.compile(loss='mae', optimizer='adam')
history = model.fit(trainX, trainY, epochs=300, batch_size=100, validation_data=(testX, testY), verbose=0,
shuffle=False)
```

In [33]:
```
yhat = model.predict(testX)
pyplot.plot(yhat, label='predict')
pyplot.plot(testY, label='true')
pyplot.legend()
pyplot.show()
```

In [34]:
```
yhat_inverse_1 = scaler.inverse_transform(yhat.reshape(-1, 1))
testY_inverse_1 = scaler.inverse_transform(testY.reshape(-1, 1))
```

In [35]:
```
rmse_1 = sqrt(mean_squared_error(testY_inverse_1, yhat_inverse_1))
print('Test RMSE: %.3f' % rmse_1)
```

```
Test RMSE: 60.814
```

In [164]:
```
model_1 = model
```

Fig 6.4 Training sample model with lookback=1

with lookback = 2(kind of biram)

```
In [8]: look_back = 2
        trainX, trainY = create_dataset(train, look_back, sentiment[0:train_size])
        testX, testY = create_dataset(test, look_back, sentiment[train_size:len(scaled)])

In [9]: trainX = np.reshape(trainX, (trainX.shape[0], 1, trainX.shape[1]))
        testX = np.reshape(testX, (testX.shape[0], 1, testX.shape[1]))

In [10]: model = Sequential()
         model.add(LSTM(100, input_shape=(trainX.shape[1], trainX.shape[2])))
         model.add(Dense(1))
         model.compile(loss='mae', optimizer='adam')
         history = model.fit(trainX, trainY, epochs=300, batch_size=100, validation_data=(testX, testY), verbose=0,
         shuffle=False)

In [11]: yhat = model.predict(testX)
         pyplot.plot(yhat, label='predict')
         pyplot.plot(testY, label='true')
         pyplot.legend()
         pyplot.show()

         <IPython.core.display.Javascript object>
```

Fig 6.5 Training sample model with lookback=2

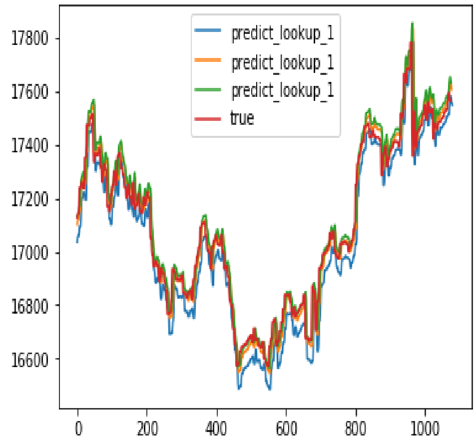
With lookback = 3

```
In [33]: look_back = 3
         trainX, trainY = create_dataset(train, look_back, sentiment[0:train_size])
         testX, testY = create_dataset(test, look_back, sentiment[train_size:len(scaled)])

         2517
         1076

In [34]: trainX = np.reshape(trainX, (trainX.shape[0], 1, trainX.shape[1]))
         testX = np.reshape(testX, (testX.shape[0], 1, testX.shape[1]))

In [36]: model = Sequential()
         model.add(LSTM(100, input_shape=(trainX.shape[1], trainX.shape[2])))
         model.add(Dense(1))
         model.compile(loss='mae', optimizer='adam')
         history = model.fit(trainX, trainY, epochs=300, batch_size=100, validation_data=(testX, testY), verbose=0,
         shuffle=False)

In [37]: yhat = model.predict(testX)
         pyplot.plot(yhat, label='predict')
         pyplot.plot(testY, label='true')
         pyplot.legend()
         pyplot.show()
```

Fig 6.6 Training sample model with lookback=3

Plotting different lookups

In [43]:
```python
pyplot.plot(yhat_inverse_1, label='predict_lookup_1')
pyplot.plot(yhat_inverse_2, label='predict_lookup_1')
pyplot.plot(yhat_inverse_3, label='predict_lookup_1')
pyplot.plot(testY_inverse_3, label='true')
pyplot.legend()
pyplot.show()
```



In [56]:
```python
len(datag.index.values)
```

Out[56]: 3605

In [63]:
```python
btc_1_trace = go.Scatter(x=datag.index.values[3605-1080-1:], y=yhat_inverse_1.reshape(1080), name= 'predic
t_lookup_1')
btc_2_trace = go.Scatter(x=datag.index.values[3605-1078-1:], y=yhat_inverse_2.reshape(1078), name= 'predic
t_lookup_2')
btc_3_trace = go.Scatter(x=datag.index.values[3605-1076-1:], y=yhat_inverse_3.reshape(1076), name= 'predic
t_lookup_3')
```

Fig 6.7-Plotting different lookups

Lookups with sentiments

```
In [14]: look_back = 2
         trainX, trainY = create_dataset(train, look_back, sentiment[0:train_size],sent=True)
         testX, testY = create_dataset(test, look_back, sentiment[train_size:len(scaled)], sent=True)

In [15]: trainX = np.reshape(trainX, (trainX.shape[0], 1, trainX.shape[1]))
         testX = np.reshape(testX, (testX.shape[0], 1, testX.shape[1]))

In [26]: model = Sequential()
         model.add(LSTM(100, input_shape=(trainX.shape[1], trainX.shape[2]), return_sequences=True))
         model.add(LSTM(100))
         model.add(Dense(1))
         model.compile(loss='mae', optimizer='adam')
         history = model.fit(trainX, trainY, epochs=300, batch_size=100, validation_data=(testX, testY), verbose=0,
         shuffle=False)

In [27]: yhat = model.predict(testX)
         pyplot.plot(yhat, label='predict')
         pyplot.plot(testY, label='true')
         pyplot.legend()
         pyplot.show()

In [28]: yhat_inverse_sent = scaler.inverse_transform(yhat.reshape(-1, 1))
         testY_inverse_sent = scaler.inverse_transform(testY.reshape(-1, 1))

In [29]: rmse_sent = sqrt(mean_squared_error(testY_inverse_sent, yhat_inverse_sent))
         print('Test RMSE: %.3f' % rmse_sent)

         Test RMSE: 44.563

In [ ]: len(yhat)

In [36]: btc_1_trace = go.Scatter(x=datag.index.values[3605-1078-1:][0:500], y=yhat_inverse_sent.reshape(1078)[0:50
         0], name= 'With_Sentiment')
```

Fig 6.8-Plotting with sentiments

Now go live with the model:

In [67]:
```
import MySQLdb
#Enter the values for you database connection
dsn_database = "bitcoin"         # e.g. "MySQLdbtest"
dsn_hostname = "173.194.231.244"      # e.g.: "mydbinstance.xyz.us-east-1.rds.amazonaws.com"
dsn_port = 3306                  # e.g. 3306
dsn_uid = "demo"             # e.g. "user1"
dsn_pwd = "qwerty@123"              # e.g. "Password123"
```

In [68]:
```
conn = MySQLdb.connect(host=dsn_hostname, port=dsn_port, user=dsn_uid, passwd=dsn_pwd, db=dsn_database)
```

In [69]:
```
cursor=conn.cursor()
cursor.execute("""SELECT * FROM live_data""")
cursor.fetchone()
```

Out[69]: (1234.0, 123.0, 456.0, datetime.date(2009, 2, 2))

In [70]:
```
print ("\nShow me the records:\n")
rows = cursor.fetchall()
import pprint
pprint.pprint(rows)
```

```
Show me the records:

()
```

In [71]:
```
cursor.execute("""INSERT INTO live_data values(15625,16000,0.8,'1000-01-01 00:00:00');""")
```

Out[71]: 1

In [73]:
```
conn.commit()
```

In [66]:
```
cursor.close()
```

Fig 6.9-Live with Model

6.6 Summary

This chapter explains about lstm working and its architecture and its importance. Process of developing machine learning model and briefing out each steps, and tools used to develop the model.

CHAPTER 7

TESTING

System testing is actually a series of different tests whose primary purpose is to fully exercise the computer-based system. Although each test has a different purpose, all work to verify that all the system elements have been properly integrated and perform allocated functions. The testing process is actually carried out to make sure that the product exactly does the same thing what is supposed to do. In the testing stage following goals are tried to achieve: -

- To affirm the quality of the project.
- To find and eliminate and residual errors from previous stages.
- To validate the software as solution to the original problem.
- To provide operational reliability of the system.

Configuration change	MSE	R^2	Forecast bias	MAE	ME%
2 RNN layers	3.00632E-5	0.99930	60.07$	69.10$	0.46
1 RNN neuron	0.00180	0.95799	549.53$	549.73$	3.59
32 RNN neurons	3.755927E-6	0.99991	17.42$	23.22$	0.17
128 RNN neurons	2.70684E-6	0.99994	16.9$	21.39$	0.17
LSTM architecture	6.73593E-6	0.99984	18.5$	32.83$	0.25
500-neuron dense layer	1.66677E-5	0.99961	55.89$	57.66$	0.42
Dropout 0.25	5.76688E-6	0.99986	-35.52$	37.16$	0.30
Dropout 0.5	2.58831E-5	0.99940	-62.97$	68.50$	0.47
20 epochs	4.27518E-6	0.99990	22.01$	26.45$	0.20
30 epochs	2.68944E-6	0.99994	10.72$	18.91$	0.14
50 epochs	2.68130E-6	0.99994	8.79$	19.68$	0.15
Batch size 250	3.82469E-6	0.99991	20.41$	24.59$	0.18
Batch size 1000	4.35678E-6	0.99988	25.35$	29.71$	0.22
Nadam optimizer	2.11560E-5	0.99950	54.85$	58.52$	0.39
RMSprop optimizer	0.00037	0.99146	229.04$	243.55$	1.59
0.0001 learning rate	2.64888E-6	0.99994	13.12$	19.08$	0.14
0.01 learning rate	2.88746E-6	0.99993	16.22$	21.21$	0.16

Fig 7.1 - Testing Values

The model works relatively well for identifying general trends in coin prices, but struggles to accurately predict daily price fluctuations that are not in line with the general trend. Specifically, there is a general increase in bitcoin prices during the test set time period, and the model correctly picks up on this via the text input and most often predicts additional price increases. As a result, the final model was not able to predict the very large increase in price during the test set time period.

7.2 Summary

This chapter discuss about the importance of testing and varies methods that are used to test the model built. This helps us to understand the performance of the system and make the necessary changes accordingly.

CHAPTER 8

RESULTS AND DISCUSSION

8.1 Discussion

To give context to our results, it is important to understand the general price behaviour of Bitcoin during the test set time period.

This graph (Figure 6.10) shows how the relation between the price of bitcoin and total volume of the tweets that is made on that day addressing bitcoin.

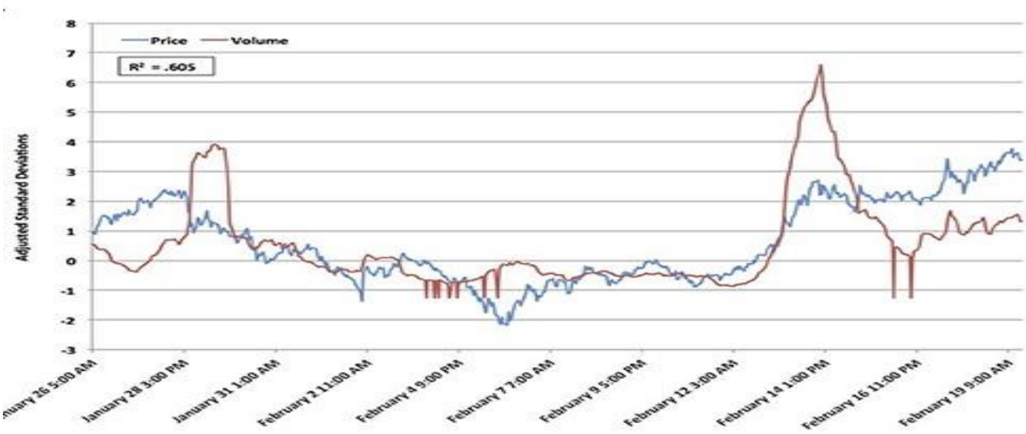


Fig 6.10 price vs volume

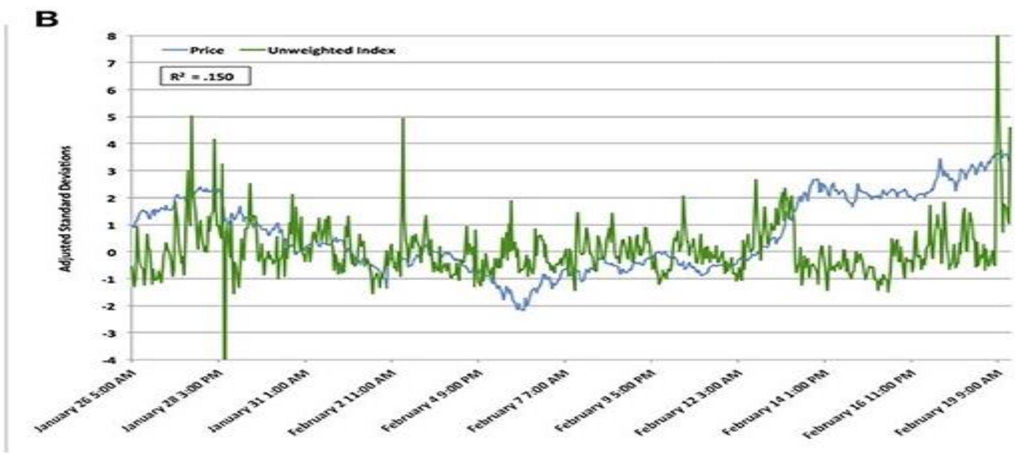


Fig 6.11 Price vs Unweighted Index

This graph (Figure 6.11) shows the relation between the price of bitcoin and the tweets that does not have polarity assigned to them i.e. all tweets have same weight.

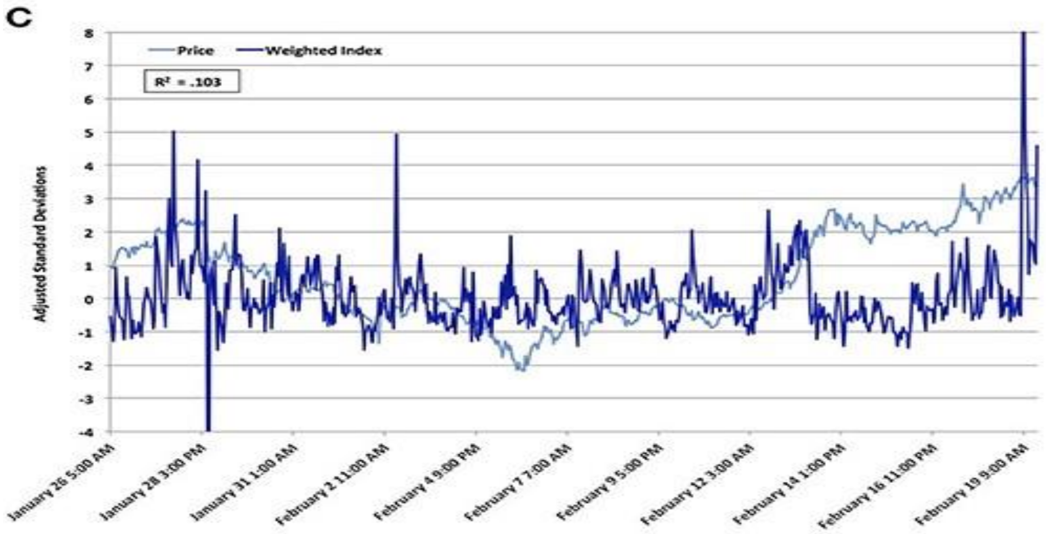


Fig 6.12 Price vs Weighted Index

This graph (Figure 6.12) shows the relation between the price of bitcoin and the tweets that have polarity assigned to them i.e. different tweets have different weights and they affect the price of bitcoin differently.

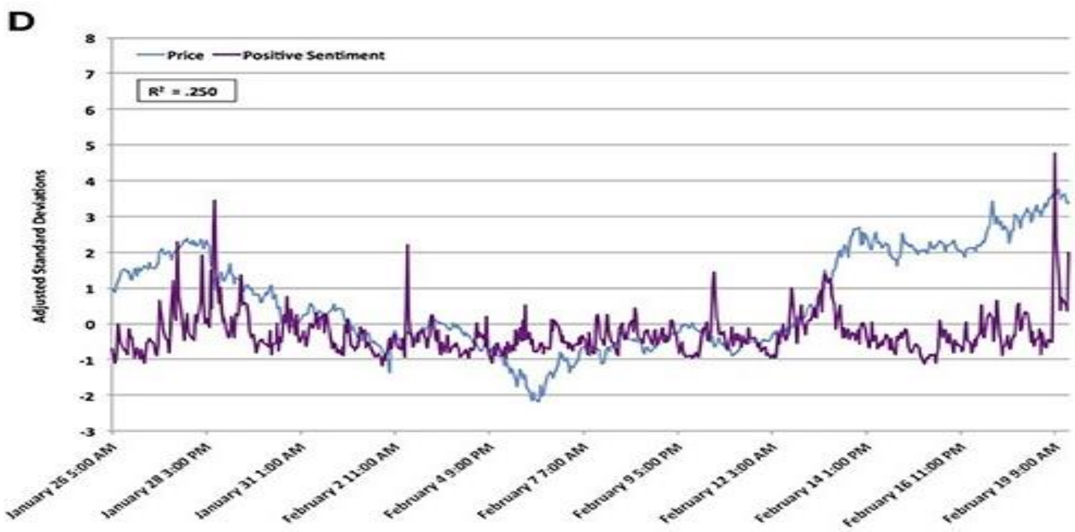


Fig 6.13 Price vs Positive Sentiment

This graph (Figure 6.13) shows the relation between the price of bitcoin and the tweets that have positive polarity i.e. we can see that if positive tweets have been made in context of bitcoin than the price of bitcoin increases.

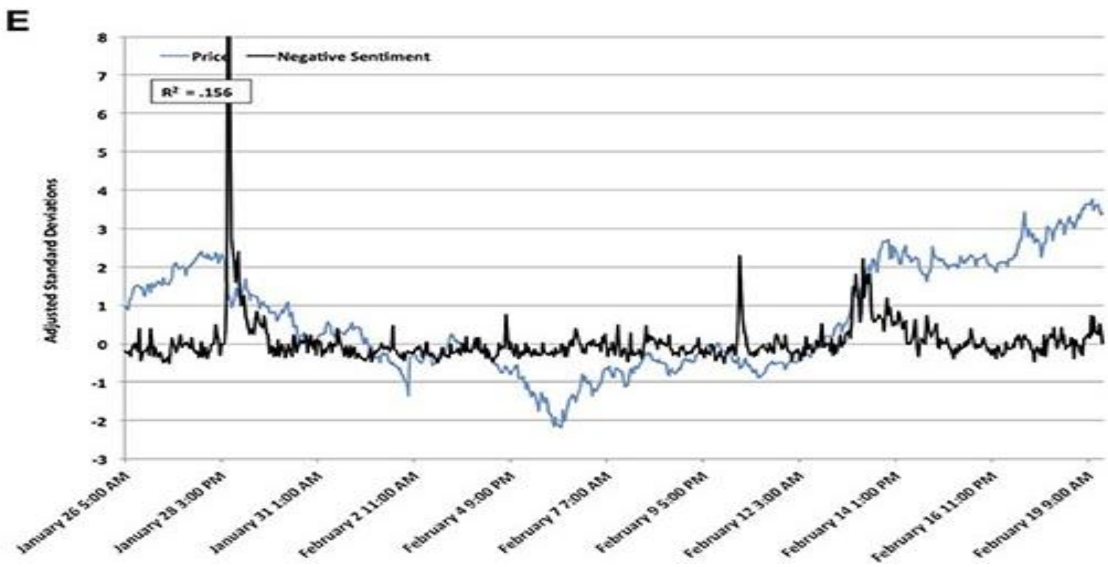


Fig 6.14 Price vs Negative Sentiment

This graph (Figure 6.14) shows the relation between the price of bitcoin and the tweets that have negative polarity i.e. we can see from the graph that if the volume of negative tweets increases the price of bitcoin decreases.

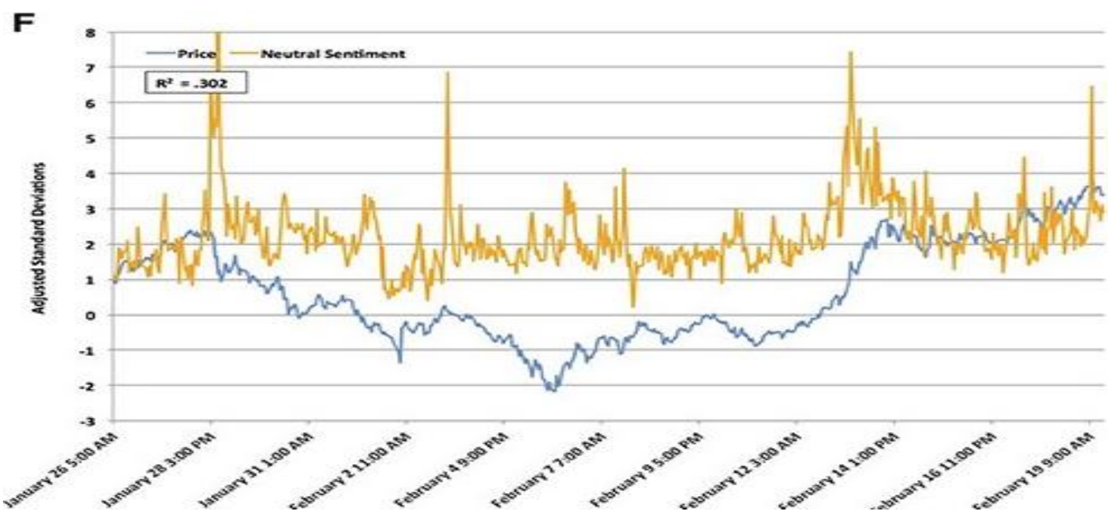


Fig 6.15 Price vs Neutral sentiment

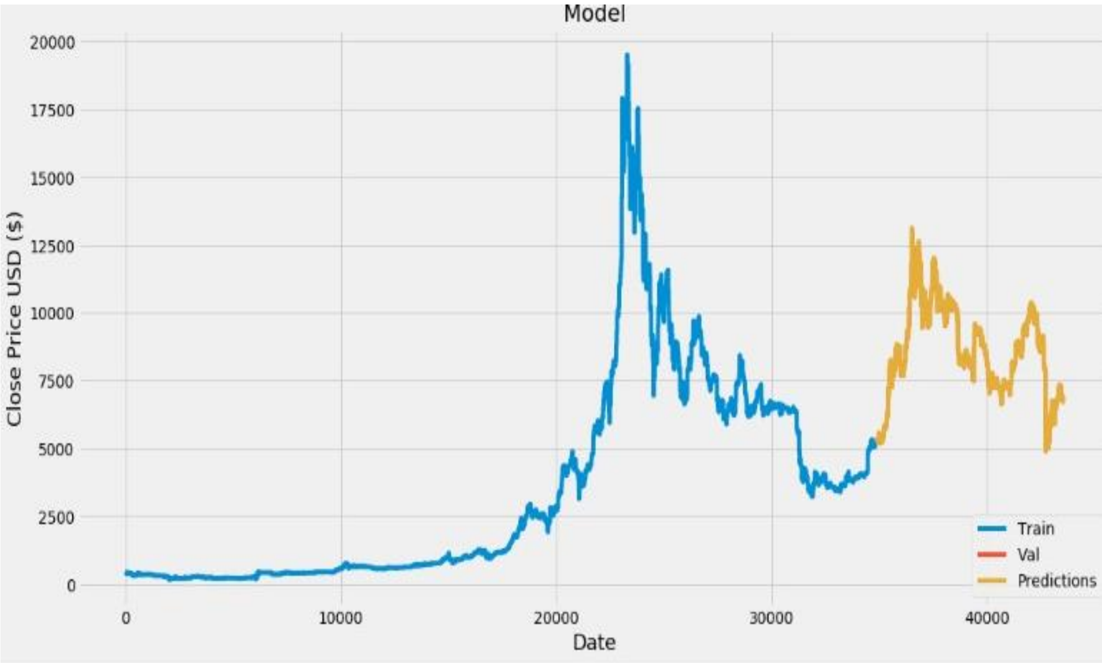


Fig 6.16 Train vs Test

This graph (Figure 6.16) shows the Training and Testing phase of the model the curve with the blue colour is the period in which the model is being trained with the historical data and after that portion of the curve we can see the predictions that the model made and we compare those prediction with rest of the historical data to get the accuracy of the model.

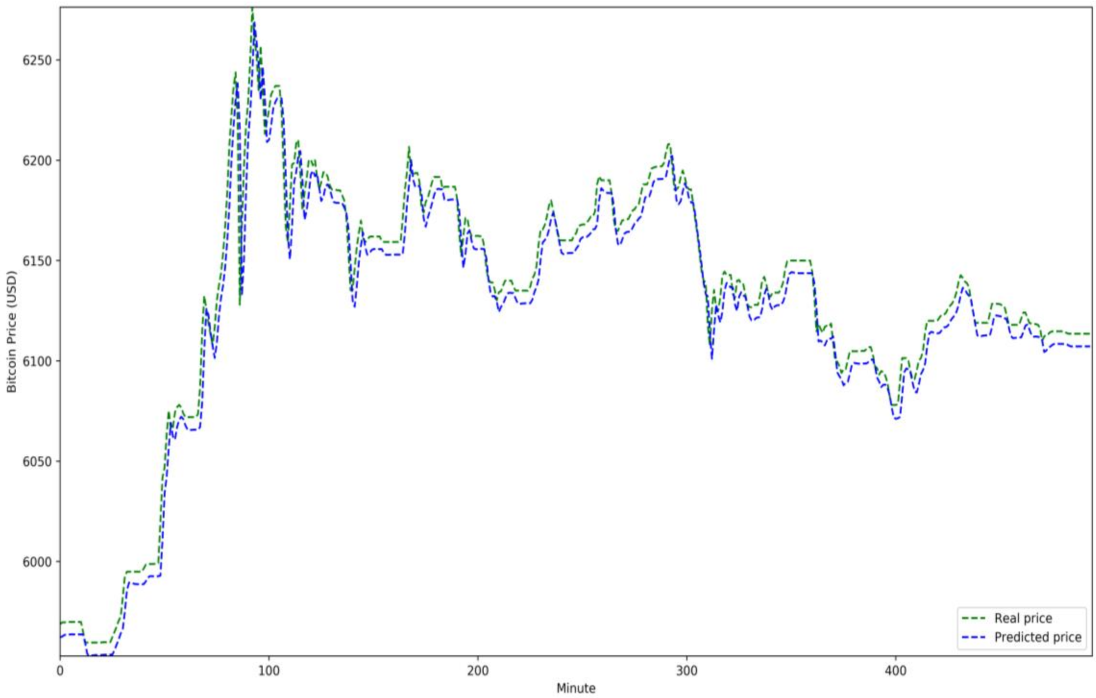


Fig 6.17 Actual VS Predicted graph.

This graph (Figure 6.17) shows the relation between the actual price of bitcoin and the model predicted price of the bitcoin i.e. we can see from the graph that the error in the predicted price is less.

8.2 Summary

This chapter clearly shows the working of our web application through the screen shots and by including some brief discussion to that. For better understanding of variations of water parameter values screen shot of graphs is also included.

CHAPTER 9

CONCLUSION AND FUTURE WORK

Deep learning models such as the RNN and LSTM are evidently effective learners on training data with the LSTM more capable for recognising longer-term dependencies. However, a high variance task of this nature makes it difficult to transpire this into impressive validation results. As a result, it remains a difficult task. There is a fine line to balance between overfitting a model and preventing it from learning sufficiently. Dropout is a valuable feature to assist in improving this. However, despite using Bayesian optimisation to optimize the selection of dropout it still couldn't guarantee good validation results. Despite the metrics of sensitivity, specificity and precision indicating good performance, the actual performance of the ARIMA forecast based on error was significantly worse than the neural network models. The LSTM outperformed the RNN marginally, but there was not significant difference in the results of both. However, the LSTM takes considerably longer to train. The performance benefits gained from the parallelisation of machine learning algorithms on a GPU are evident with a 70.7% performance improvement for training the LSTM model on the GPU as opposed to the CPU. This confirmed the findings indicated by the related work.

Previous efforts to predict cryptocurrency fluctuations relied on Twitter sentiment analysis to serve as a proxy for future cryptocurrency demand which would result in increasing or decreasing prices. We have shown that these results were in part due to the study occurring at a time when cryptocurrency prices were always going up. Additionally, Twitter sentiment with respect to cryptocurrencies tend to be positive regardless of future price changes. People who tweet about cryptocurrencies even when their prices drop have an interest in them beyond investment opportunity making the tweets biased towards positive.

A more robust model would incorporate a measure of overall interest in terms of volume. This paper's recommendation is to use proxies for general interest such as

Google Trends or tweet volumes. We have shown that the search volume index is highly correlated with cryptocurrency prices both when prices rise and when they fall, as are tweet volumes. With these inputs a multiple linear regression model, with the addition of lagged variables, accurately predicted future price changes. Future work should determine if these results continue to hold in varying pricing environments. Additionally, more complex models, and not just linear ones like we used, could be fit using Google Trends and tweet volumes as inputs to see if results could be improved further.

REFERENCES

[1] V. Y. Naimy, and M. R. Hayek, "Modelling and predicting the Bitcoin volatility using GARCH models," International Journal of Mathematical Modelling and Numerical Optimisation, vol. 8, no. 3, pp. 197-215, 2018.

[2] Y. B. Kim, J. G. Kim, W. Kim, J. H. Im, I. H. Kim, S. J. Kang, and C. H. Kim, "Predicting fluctuations in cryptocurrency transactions based on user comments and replies," PloS one, vol. 11, no. 8, e0161197, 2016.

[3] S. Ahamad, M. Nair, and B. Varghese, "A survey on crypto currencies," In 4th International Conference on Advances in Computer Science, AETACS, pp. 42-48, Citeseer, 2013.

[4] D. Garcia, C. J. Tessone, P. Mavrodiev, and N. Perony, "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy," Journal of the Royal Society Interface, vol. 11, no. 99, 20140623, 2014.

[5] M. Matta, I. Lunesu, and M. Marchesi, "Bitcoin spread prediction using social and web search media," In UMAP Workshops, 2015.

[6] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[7] R. Böhme, N. Christin, B. Edelman, and T. Moore, "Bitcoin: economics, technology, and governance," Journal of Economic Perspectives, vol. 29, no. 2, pp. 213-238, 2015.

[8] R. Grinberg, "Bitcoin: An innovative alternative digital currency," Hastings Science & Technology, LJ, 4, 159, 2012.

[9] N. I. Indera, I. M. Yassin, A. Zabidi, and Z. I. Rizman, "Non-linear autoregressive with exogeneous input (NARX) Bitcoin price prediction model using PSO-optimized parameters and moving average technical indicators," Journal of Fundamental and Applied Sciences, vol. 9, no. 3S, pp. 791-808,

2017.

[10] 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.

[11] M. H. Miraz, and M. Ali, “Applications of blockchain technology beyond cryptocurrency,” arXiv preprint arXiv:1801.03528 (2018).

[12] J. H. Ziegeldorf, R. Matzutt, M. Henze, F. Grossmann, and K. Wehrle, “Secure and anonymous decentralized Bitcoin mixing,” Future Generation Computer Systems, 80, pp. 448-466, 2018.

[13] B. Biais, C. Bisiere, M. Bouvard, and C. Casamatta, “The blockchain folk theorem,” 2018.

[14] J. Ma, J. S. Gans, and R. Tourky, “Market structure in Bitcoin mining, No. w24242, National Bureau of Economic Research, 2018.

[15] C. Decker, and R. Wattenhofer, “Information propagation in the Bitcoin network,” In 13th International Conference on Peer-to-Peer Computing (P2P), pp. 1-10, IEEE, 2013.

[16] A. Jain, A. Gupta, N. Sharma, S. Joshi, and D. Yadav, “Mininig application on analyzing users’ interests from Twitter,” In Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), IEEE, March, 2018.

[17] G. Hileman, and M. Rauchs, “2017 global cryptocurrency benchmarking study,” 2017.

[18] S. Eskandari, J. Clark, D. Barrera, and E. Stobert, “A first look at the usability of bitcoin key management,” arXiv preprint arXiv:1802.04351, 2018.

[19] R. Sidel, “Overstock CEO sees Bitcoin sales rising more than expected,” Wall Street, 2014.

[20] R. Grinberg, “Bitcoin: An innovative alternative digital currency,” Hastings Science & Techchnology, LJ, 4, 159, 2012.

[21] I. Alqassem, and D. Svetinovic, “Towards reference architecture for cryptocurrencies: Bitcoin architectural analysis,” In International Conference on Internet of Things (iThings), and Green Computing and Communications (GreenCom), and Cyber, Physical and Social Computing (CPSCom), pp. 436-443. IEEE, 2014.

[22] F. Armknecht, G. O. Karame, A. Mandal, F. Youssef, and E. Zenner, “Ripple: overview and outlook,” In International Conference on Trust and Trustworthy Computing, pp. 163-180. Springer, Cham, 2015.

[23] D. Bradbury, “Bitcoin’s successors: from Litecoin to Freicoin and onwards,” The Guardian, 25, 2013.

[24] N. Gandal, and H. Halaburda, “Competition in the cryptocurrency market,” 2014.

[25] S. McNally, J. Roche, and S. Caton, “Predicting the price of Bitcoin using Machine Learning,” In 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), pp. 339- 343, IEEE, 2018.

[26] A. Greaves, and B. Au, “Using the Bitcoin transaction graph to predict the price of Bitcoin,” No Data, 2015.

[27] S. Asadi, S. S. Amiri, and M. Mottahedi, “On the development of multi-linear regression analysis to assess energy consumption in the early stages of building design,” Energy and Buildings, 85, 246-255, 2014.