

**STOCK MARKET PRICE PREDICTION USING LONG-SHORT
TERM MEMORY (LSTM)**

FOR THE AWARD OF DEGREE
OF

MASTER OF TECHNOLOGY

IN

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

[COMPUTER SCIENCE AND ENGINEERING]

Submitted By:

[WALAA MAKHLOUF]

(Roll No. 2K18/CSE/21)

Under the supervision of

[Dr. Ruchika Malhotra]

(Associate Head & Associate professor, Department of Computer science
and Engineering)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

OCTOBER, 2020

DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY DELHI COLLEGE OF ENGINEERING)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Walaa Makhlof, Roll No 2K18/CSE/21, student of M.Tech in Computer Science & Engineering, hereby declare that the project Report titled "Stock market price prediction using Long-Short Term Memory (LSTM)" which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirements for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Date: 29th Oct 2020



Walaa Makhlof

STUDENT

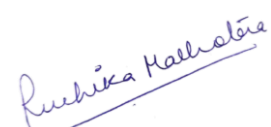
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Report titled " Stock market price prediction using Long-Short Term Memory (LSTM)" which is submitted by Walaa Makhlouf, Roll No 2K18/CSE/21 Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 29th Oct 2020



Dr. Ruchika Malhorta

SUPERVISOR

ABSTRACT

Stock index prices predicting is a tough task and, because of various reasons relating to many technological and non - tech reasons, share price knowledge is an extremely difficult, unpredictable and dynamic environment. In parallel to deep learning techniques, a variety of academic experiments from different disciplines to resolve this topic and machine learning techniques are one of the many technologies used.

Many machine learning techniques in this field were able to produce acceptable outcomes while it was used in this type of predictions.

This project studies stock market price prediction using LSTM model which is applied on Stock index prices historical data along with indications analysis which will be used to achieve more accurate results.

In this study, data sets of historical prices of common stock of Agilent Technology, Alcoa Corporation Common Stock and American Airlines Group Common Stock were gathered to achieve this objective, and several tests were carried out using LSTM, the findings were evaluated using RMSE and RMSPE values that guarantee better performance for the LSTM method used.

ACKNOWLEDGEMENT

First, I would like to extend my sincere gratitude to my mentor, Dr. Ruchika Malhorta for her constant encouragement and insight. I thank her for suggesting the initial idea of supervised Machine Learning, which eventually formed the core of this dissertation. I greatly appreciate her constant motivation in enabling me to understand the Machine Learning and Deep learning concepts which I was not familiar with. This work would not have been accomplished without her support.

The faculty members of Computer Science and Engineering played an extensive role in preparing me to pursue this research. Without, their knowledge of the various core courses they taught me, I would not have acquired the technical knowledge of undertaking such a project. On this note therefore, I sincerely thank them for preparing me for the journey that has successfully come to its completion.

I was fortunate enough to be part of a vibrant class of members who were a great pillar in my Masters studies. I am grateful for the intellectual support they rendered to me during the two years of vigorous training. Working with classmates from various backgrounds also enabled me to develop a true sense of awareness of equality.

I also take this honour to sincerely appreciate the Indian Government through Indian Council for Cultural Relations (ICCR), my sponsoring organization for funding my Masters Education.

Above all, I thank the Almighty God for enabling me to accomplish this project.

CONTENTS

CANDIDATE'S DECLARATION	ii
CERTIFICATE	iii
ABSTRACT	iv
ACKNOWLEDGEMENT	v
CONTENTS	vi
LIST OF TABLES	viii
LIST OF FIGURES	viii
CHAPTER 1: INTRODUCTION	1
1.0 Chapter Summary	1
1.1 Introduction	1
1.2 Research Questions and Objectives	2
1.3 Problem Formulation	2
1.4 Scope	3
1.5 Limitation	3
CHAPTER 2: BACKGROUND AND MOTIVATION	4
2.0 Chapter Summary	4
2.1 Background	4
2.2 Efficient-market theory	4
2.3 Artificial Neural Networks (ANN)	5
2.4 Recurrent Neural Networks (RNN)	7
2.5 Long Short-Term Memory	4
CHAPTER 3: LITERATURE REVIEW	11
3.0 Chapter Summary	11
3.1 Related Work	11
CHAPTER 4: METHODOLOGY	16

4.0 Chapter Summary	16
4.1 Data Gathering and Preprocessing	16
4.2 Technical Indicators and Features	16
4.3 LSTM Model	18
4.4 Evaluation/Performance Measurement	18
CHAPTER 5: RESULTS AND DISCUSSION	20
5.0 Chapter Summary	20
5.1 Modeling and Prediction	20
5.2 Results and Model Evaluation	20
CHAPTER 6: CONCLUSION AND FUTURE WORK	30
6.0 Chapter Summary	30
6.1 Conclusion	30
6.2 Future Work	30
REFERENCES	32

LIST OF TABLES

Table 1: Agilent Technologies stock LSTM performance measurement	28
Table 2: Alcoa Corporation Stock LSTM performance measurement	29
Table 3: American Airlines Stock LSTM performance measurement	29

LIST OF FIGURES

Figure 1: A simplified comparative illustration of an ANN and RNN	7
Figure 2: model structure of the RNN	8
Figure 3: RNN structure	8
Figure 4: the condensed version of the repeated module in an LSTM	9
Figure 5: LSTM Cell architecture	12
Figure 6: preview of used dataset	17
Figure 7: Agilent Technologies stock prediction using 5 Time steps	21
Figure 8: Agilent Technologies stock prediction using 10 time steps	21
Figure 9: Agilent Technologies stock prediction using 25time steps	22
Figure 10: Agilent Technologies stock prediction using 50 time steps	22
Figure 11: Agilent Technologies stock prediction using 100 time steps	23
Figure 12: Alcoa Corporation Stock prediction using 5 time steps	23
Figure 13: Alcoa Corporation Stock prediction using 10 time steps	24
Figure 14: Alcoa Corporation Stock prediction using 25 time steps	24
Figure 15: Alcoa Corporation Stock prediction using 50 time steps	25
Figure 16: Alcoa Corporation Stock prediction using 100 time steps	25
Figure 17: American Airlines Stock prediction using 5 time steps	26
Figure 18: American Airlines Stock prediction using 10 time steps	26
Figure 19: American Airlines Stock prediction using 25 time steps	27
Figure 20: American Airlines Stock prediction using 50 time steps	27
Figure 21: American Airlines Stock prediction using 100 time steps	28

CHAPTER 1: INTRODUCTION

1.0 Chapter Summary

The initial overview of the subject of study was added in this chapter. The analysis issue that influenced the thesis and the aims follows this. Over the next part, the issue being discussed has been conceived and nature and constraint of the study has been well established.

1.1 Introduction

Due to its value in trading in order to getting investors ready to make informed choices where to sell and where not to sell based on evaluating the available information of the portfolio that will show whether the portfolio price will increase or downward in the future, financial market projections have been very important study point over the past several years. The signal-to - noise ratio is immense and marginal in the set of variables and pieces of knowledge considered.

The task of predicting the stock share price rates are high tricky. For many years, research has been investigating the possibility of this achievement, and it is shocking in the published studies that most projection models fail in a general sense to make accurate predictions. Nonetheless, there seem to be an increasing amount of research from multiple areas that seek to solve this issue, providing a wide variety of approaches to that end.

So the use of AI and ML Techniques and models which able to be trained through historical market index to do the prediction of future market values is one appropriate approach. This research moves in that direction but studies the particular framework used by neural networks under deep learning type. The short-term memory advantage of this technique and the hypothesis to be measured here is that this function would benefit from more conventional approaches in the machine learning field than those in terms of efficiency.

In this project LSTM is the Selected chosen algorithm here. neural network is very strong on a number of topics and one of them is stock future price prediction, including its ability to distinguish between latest and old examples by assigning 1 point different sizes, while lacking the knowledge it finds to be meaningless to predict the next output. Nevertheless, similar to other deep learning neural networks which are only able to memorize a small list, it is more able of solving long sequence data with saving important information.

The main objective of this project is to find out the usefulness of the forecasting stock market volatility for recurring Deep learning neural networks, especially the LSTM networks. Evaluate their efficiency in terms of efficiency and other measurements by real - world data studies and evaluate whether they show any gain relative to various machine learning algorithms.

1.2 Research Questions and Objectives

Research Questions

- (1) Can we use deep learning approaches to achieve very detailed forecast stock values using different features?
- (2) Can the expected prices of the stocks in the investment sector be used and how can this affect the result of the investor?

Objectives

- (1) In order to build a model for stock market index forecasts, the approach based on the LSTM deep learning approach.
- (2) To train and test a LSTM in the field of stock market price prediction.
- (3) To assess the model based of RMSE principles and to carry out preparation and research on the basis of other evaluation techniques.

1.3 Problem Formulation

Predicting stock direction and stock price index are difficult due to uncertainties involved. There's many, nevertheless, two forms of analysis which traders have before making an investment. The basic analysis is first. In this, investors look at the valuation methods of inventories, business and economic results, political environment, etc. to determine to choose whether or not purchase. Technical analysis, and from the other hand, is the assessment of stocks by the study of researchers implement by economic trends, such as past volume and price. Technical analysts do not aim to quantify the inherent value of a security, but instead use trading strategies to detect insights and changes that can show how a stock will function in the future.

Skeptics might argue that, because markets seem to be efficient, there is little room for predictability. The truth is that tremendous investment and commitment remains to be made in discovering structure and guidance in financial system.

Therefore the problem becomes: for a given stock market history, determine the moment of buying/entry or selling/exit the good for generating profit. Machine learning (ML) is one of the main methods to address the problem of big data mining. ML will enable self-innovation and enhancement of the e-commerce system by acquiring previous knowledge. Big data generated by e-commerce companies' sales, engagement and evaluation will significantly include a decision-making service for marketing technique.

1.4 Scope

The analysis utilizes empirical data, i.e. historical stock market data obtained from 2005 to 2020 over 15 years, to train and validate an LSTM network for potential stock market index prediction. The educated model can only forecast fraud along with past data gathered from Kaggle. Instead, the framework on when such a model is to be implemented is being used in data collection for broader use, since the network state can be predicted by the learned model whose training dataset contains information from such a system.

1.5 Limitation

Noisy, unpredictable existence of information on stock markets make estimation impossible and the unforeseen fundamental variables such as the market combine or split, political condition, disasters, etc. that cannot be forecast and influence the course of markets.

CHAPTER 2: BACKGROUND AND MOTIVATION

2.0 Chapter Summary

This chapter is split up into different parts. The first following part offers a detailed study history and the key drivers behind this entire subject; the second part explores the different principles of using methods.

2.1 Background

Companies are divided into shares and exchanged for hundreds of years as stocks[19]. A major portion of our modern-day economy is the securities industry. Indices are used very often in the efficiency assessment of the overall stock market. For some examples, endorsed corporate size (size relating to market capitalization), geography or industries are often selected for stocks that are included in an index. Using the price index prices of three companies. as an example, the used indexes are typically weighted and most widely and constituted by the larger firms in chosen markets.

We want an adequate amount of information to build a correct ML approach to be ready to forecast stock exchange fluctuations, so we will target the index with the most significant amount of knowledge available.

2.2 Efficient-market theory

The efficient-market Theory (EMT) in economics is an important theory formalised and published by Eugene Fama in 1970 by extending the initial concept he had formulated just a year earlier [21]. It notes that markets are successful financially considering all stock information is completely expressed in stock values and that securities are always selling at reasonable value, making it impossible to underestimate and overestimate public corporations. It will then be impossible to outperform the market, suggesting that the only method to obtain outsized returns is to maximize investment risk. In his 1970s paper, Fama suggested three different iterations of his hypothesis. Semi-strong performance means that technical or fundamental analysis renders it impossible to achieve surplus returns, because this sort of market behavior will render policies unable to take advantage of the skewed and inefficient nature of this market.

Strong business effectiveness means market is actually efficient and that all equity information, both private and public, is reflected by historical data set. This degree of success means that consistent excess returns are difficult to produce.

In all other phrases, through applying excess pressure, the only key to beating the work account for a long time period is. In recent years, the EMT and its Key inference that the rivalry is unbeatable have both been praised and steadily rejected since its inception. Such influential owners As Warren Buffett debunked the hypothesis, and analytical papers such as the 1995 paper on low P / E stocks with outsized returns by Dremen and Berrys dismissed at least the high business performance version of the EMH[24][25]. In comparison, a popular reflection of varying degrees of productivity in the financial instrument: weak, semi-solid, and strong form. Loose-shape performance states that future values are hard to predict by following past patterns.

In comparison, in 1999, Andrew Lo and Craig MacKinlay published their EMH and RMH study paper collection, A Non-Random Walk Down Wall Street, basically saying that there is no random walk [26]. In several cases, the EMT has thus been dismissed, adding to the economic value of this paper, including adding to the study entity discussing the existence of Markets that are competitive as described by Fama, whether or not the used stock market indexes relates to random walking. In addition to contributions in computer science and machine learning, the study of literature conducted appears to show an economic importance for this project, with various cited papers and works refuting the EMH, such as the whole area of cognitive finance.

2.3 Artificial Neural Networks (ANN)

Essentially, as an algorithm, the ANN is brain-inspired, where endless neurons have provided signals between each other. They are referred to as neural [27] due to their origins in a simplified human neuron type, the McCulloch-Pitts neuron. However, the new execution of ANNs no longer draws on these biological inspirations. Instead, an ANN is a small network of computing units, with an input value vector being taken by each unit and a single computational output value being generated. Applications in current ANNs are automation, computer vision, image classification and stock price analysis.

In general, ANNs can be described in a simplified way as graph theory with weighted edges. The feed-forward network, a multi - layer network where nodes are connected with no

circles, is the simplest type of machine learning; outputs are transmitted within each layer of processing layers without moving anything towards the next upper layer [27].

This method, before a real-valued output is obtained, the input image is directed through the Center Surfaces. Output vector and activation function consist of single levels, because the opaque two layers comprise of one or two layers, depending on the design.

$$y = \sigma(z) = \frac{1}{(1 + e^{-z})}$$

Another activation function, arguably the most prevalent in the rest as the positive part of its declaration, it is specified feature [27] is a deep neural network. As illustrated below, the performance, [28]: is defined as the positive section of the statement.

$$f(x) = x^+ = \max(0, x)$$

Linear activation units are related to units that use the connector function, and ReLUs are seen to enable deep neural structure learning to be easier and faster. Using structured details [29]. In line with the structure of the network in learning, the parameters identified by the weighted edges and the bias term are recursively calibrated for each level and ANN in reach of a performance closest to the real data [27]. The process of optimizing is according to the similar regression methods analysis methods, like gradient - based optimization of the loss function. In the network, the replication function calculates the effects of each level, where w specifies the weights and x specifies the data[30]. The mathematical reasoning functions as follows.

The propagation method be h_i , x_i be the input variables, and w_{ij} the corresponding weights are the weights for each edge in the graph. Then one can be used to activate the i : th node activation function on the j :th layer to be shown as:

$$h_i = \sum_{j=0}^n x_j w_{ij} \quad i = 1, 2, 3, \dots, m$$

As stated, the propagation function generates the final the secret layer output, in accordance with activation, Function, as shown below:

$$z_i = fh(h_i)$$

$$fh = \frac{1}{1 + e^{-x}} \quad \text{if sigmoid}$$

$$fh = \max(0, x) \quad \text{if ReLU}$$

The artificial neural network's final outputs will then be able to describe to be:

$$hi = \sum_{i=0}^m z_i w_{ij} \quad i = 1, 2, 3, \dots, m$$

2.4 Recurrent Neural Networks (RNN)

Any variant of the neural is a RNN algorithm that includes a cycle nodes with links [27]. Instead, this implies the importance of a network main processor that depends on its own performance as a source, explicitly or implicitly. RNNs therefore differ from of the consume-forward ANNs described in the previous chapter in that their internal state or storage is used to filter input data where critical background is given by the sequence [31]. They thus form an excellent basis for financial time series, where historical transactions of the variable analyzed provide valuable details, such as volatility, for instance.

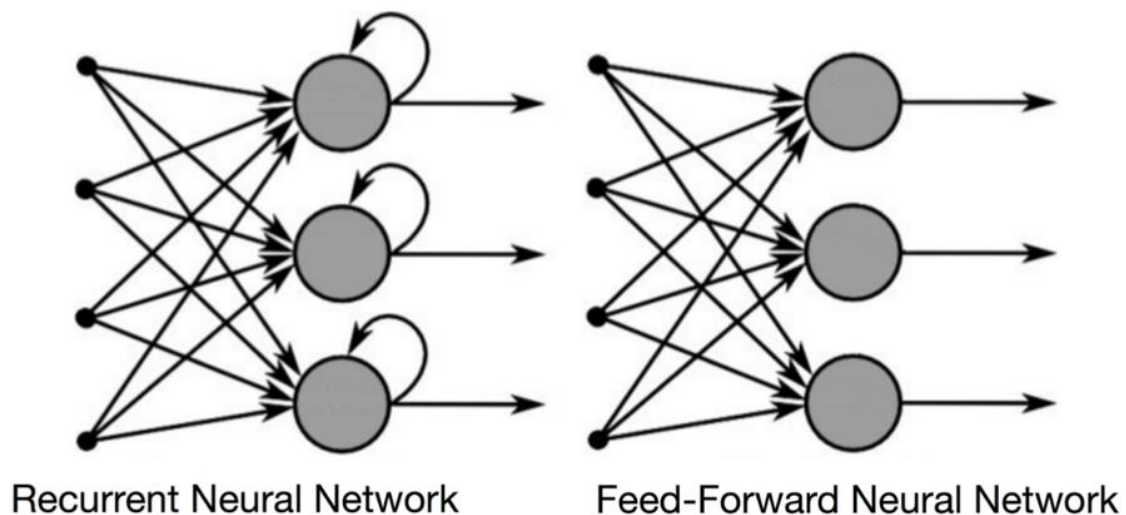


Figure 1 A simplified comparative illustration of an ANN and RNN

Making the rounds neural network (RNN) is almost like a normal neural network that deals primarily with issues of dataset. This can acquire time-series, enable data persistence, and then use previous data to produce trends of follow-up [40]. Figure 2 is a model structure of the RNN with a condensed structure of the left hand on the right. The RNN structure is seen in Figure 3.

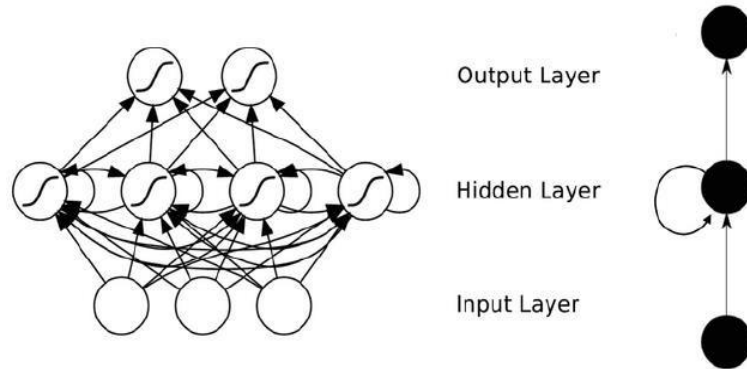


Figure 2 model structure of the RNN

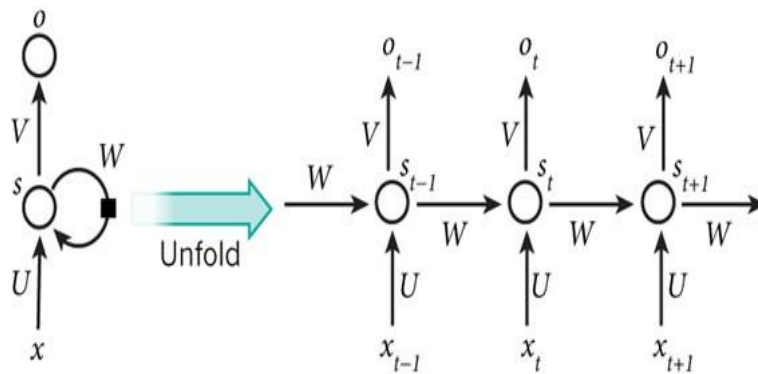


Figure 3 RNN structure

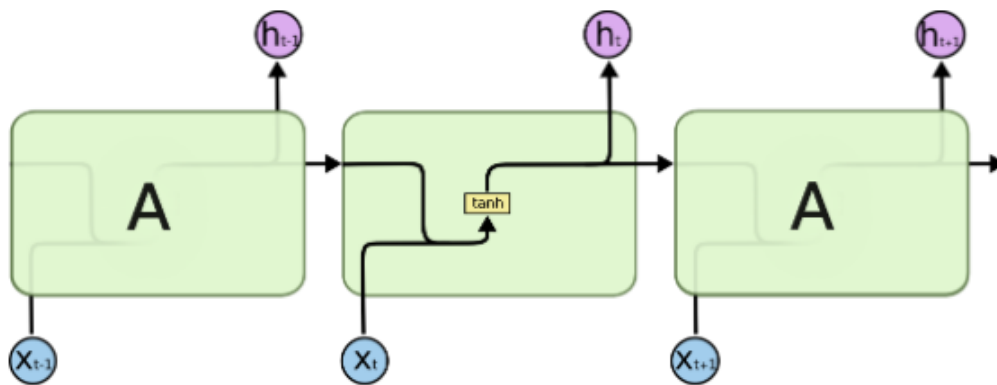
RNN's key feature is that it can work with unpredictable inputs and get those outputs.

Even then, because of the growth of the knowledge alliance, RNN will not be able to learn the relation between knowledge, and then lose all its strength. The system of LSTM was suggested and developed in terms of addressing the memory deficit of RNN.

2.5 Long Short-Term Memory

The word of Long short memory (LSTM) for this study refer to a newly developed artificial RNN architecture to resolve the overflowing and vanishing gradient problems inherent in traditional RNN preparation [32]. In short, these are all the constraints in long term that cause issues when a cell is an issue. Everything must be removed for a lengthy period of time. Where the gradient can, in many situations, train a neural network using gradient-based learning methods and back propagation. Either becomes vanishingly thin, thus escaping weight from altering its worth or starting to tend towards infinity. It becomes a concern for finite-precision numbers as the computations are used in the approach. This is partially solved by LSTMs, however. The point of having unchanged flow gradients is a problem.

A cell, an entry gate and the entrance gate make up the common LSTM unit [33]. The condensed version of the repeated module in an LSTM is shown in Fig 4 below, where each module comprises four interacting layers instead of one (activation function) layer in a regular RNN. A step by step walkthrough of how data passes through a cell in the architecture is provided in the following paragraph.



The repeating module in a standard RNN contains a single layer.

Figure 4 the condensed version of the repeated module in an LSTM

The horizontal line that passes along the figure's top is Cell condition that is altered by the layers of the LSTM unit. The original sigmoid layer represented is what is generally referred to as the "forget gate layer," which looks at the external input in the repeated modules chain as well as the input from the previous module and produces a variety of produces.

From 0 and 1 for every level, in the network. This, In essence, the sigmoid layer defines which numbers remain in the in conjunction with the following equation, the cell state and those that are eliminated:

$$ft = \sigma(Wf * [ht - 1, xt] + bf)$$

The next move is to determine what new information is available.

The latest input needs to be analysed. Next, another sigmoid layer is the "data gate layer" that determines which variables to modify. Next, a new matrix is created which, constructed of new party numbers, could be added to the cell state. This vector is generated in Fig. through a layer of tanh. 2. The cell state can be changed by integrating these two layers. As seen, this takes place through three distinct mathematical steps. Listed below:

$$it = \sigma(Wi * [ht - 1, xt] + bi)$$

After the values to update have been decided, the candidate Vector is created:

$$Cnew = \tanh(WC * [ht - 1, xt] + bC)$$

Lately, the cell state indicated by the horizontal line would be able to update:

$$Ct = ft * Ct - 1 + it * Cnewt$$

The final layer is consistent with the choice of what gets from the cell, output. Firstly, once again, the sigmoid layer dictates which beliefs continue and which are discarded.

$$ot = \sigma(Wo[ht - 1, xt] + b0)$$

Then the state of the cell passes through the layer of tanh (or ReLU in This study) prior to multiplying it by the sigmoid production Layer such that the production stays with only the planned pieces.

$$ht = ot * \tanh(Ct)$$

Then the output value is fed into the next recurring value, The module in which the procedure is replicated and to the next layer In architecture, of course.

CHAPTER 3: LITERATURE REVIEW

3.0 Chapter Summary

This chapter includes a description of the associated work and the multiple uses already being developed to the estimation of stock market values.

3.1 Related Work

In addition to its intrinsic uncertainty and dynamism, the predictability of investment returns when it comes to stock markets has been continually discussed. [1] Introduced the Efficient-Market hypothesis that determines that an asset's current price often represents all previous knowledge immediately available to it. The Random-Walk hypothesis [2] also suggests that market price rises regardless of their past, i.e. that the price of tomorrow will rely only on the knowledge of tomorrow, regardless of the price of today.

These two hypotheses determine there were no means of correctly predicting a market index. In addition,[3] has carried out a series of experiments showing that some of the most classic technique trading techniques, such as Moving Average Convergence Divergence (MACD) and Relative Strength Index (RSI), can be out-performed by a random strategy.

From another end, there are other scientists who believe that, to some degree at least, market indexes should potentially be forecast [4]. But a variety of methods to predicting and modeling stock behavior have been explored in many different areas, such as economics, statistics, genetics and computer engineering. It is important to remember that it was calculated that roughly 85% of trades inside the financial markets of the United States were conducted by algorithms in 2012. [5].

A very popular way of modeling and forecasting the financial markets is technical analysis that represents a method focused on historic market data, mainly price and volume. Some theories follow: (1) prices are determined solely by the interaction between production and consumption; (2) prices adjust following trends; (3) production and consumption shifts appear to offset patterns; (4) production and change policies may be found in the graphs; and (5) graphs appear to replicate patterns [6].In other words, no external factors, such as political, social or macro-economic factors, are taken into account in technical analysis.

Due to the fact that to obtain reliable stock market several researchers have compared various approaches with respect to computational intelligence. They go from genetic algorithms to evolutionary computation, as shown in [7],

Statistical learning using algorithms such as (SVM) [8] and a range of others, such as neural

networks, item model construction, and content analysis based Media info, which is also clarified in [9]. A modern approach-based collaborative wisdom. Taking a deeper look at tasks related to deep learning in financial markets there are several cases like [10] in which a research is made on the use of a Deep Belief Network (DBN), that is built of stacked Restricted Boltzmann Machines, bound to a Multi-level Perceptron (MLP) and uses long - distance log prices from share prices to forecast above-median rates of return for each day. [11] The use of DBN, however this time utilizing historical prices as feedback in addition to functional factors, in such a similar strategy to this initiative. Each of these tasks present positive performance especially in comparison to their baseline methods, and also in [12] in which a review of deep learning techniques introduced to financial is carried out and their enhancements addressed.

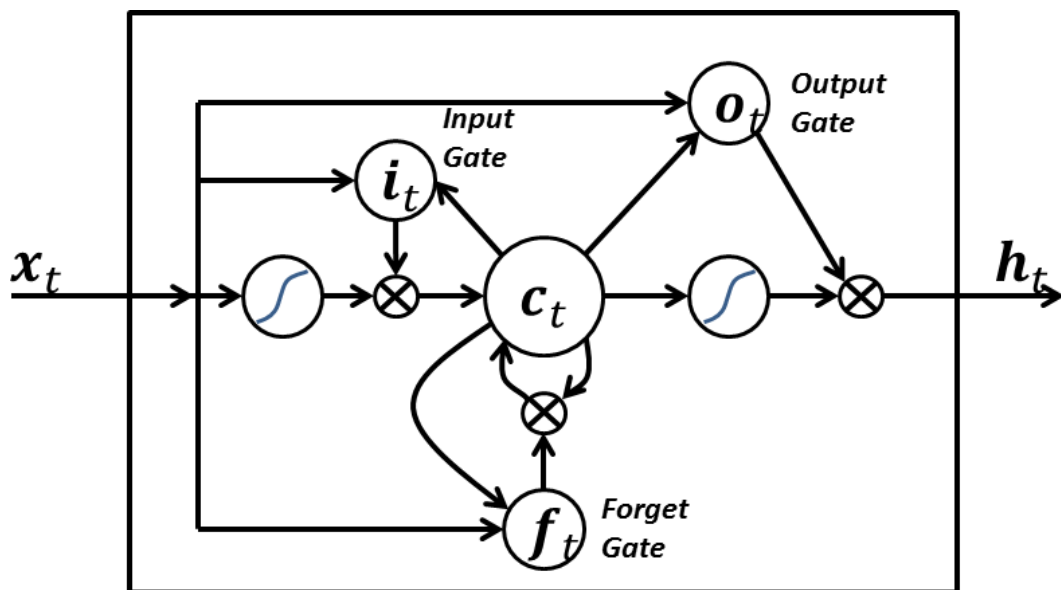


Figure 5 LSTM Cell architecture

A deep and recurrent model of neural networks is the (LSTM) Networks (Figure 1) used for this experiment. In many phrases, neurons can transmit data to a prior as well as the same layer. RNNs vary from conventional networks also in fact that they mostly have neural ties in a single direction. In either case, in addition to the LSTM that ANNs now have as a function of preparation, data does not flow in a particular direction, and the functional consequences are the presence of term memory (short). LSTM was applied by [14] and it targeted for a

better outcome by addressing the vanishing gradient problem that RNNs will struggle when facing adversity data sequences. It does so by holding the error flow via support units called "gates" constant, allowing weight changes as well as gradient truncation when its data is not needed.

These technologies are typically used, but comparison to many other methods [15], especially speech recognition Processing (NLP), they have indeed been able to deliver a few of the optimal outcomes and are called state-of-the-art, particularly for the identification of handwriting [16]. And it segmented into number of versions which checked by [13] towards their original implementation since its development, but so far do not appear to represent any significant modification.

Given the notorious success LSTM networks have used in NLP for market price prediction, it has been mainly used to forecast price patterns by breaking news content data and use it as the input data. However, there is also some analysis using market data to forecast market movement [17], using historical price data in addition to stock indices to forecast whether stock prices will increase, decline or remain the same on a regular basis. [18] Evaluates the accuracy of LSTM and MLP with their own recommended solution based on a mixture of wavelet coefficients and deep CNNs, both of which perform better the LSTM and both have effects similar.

In particular, since the turn of the century, neural networks have been extensively Applied to financial time series data through numerous experiments, as defined in the Introduction of this document and in the following subsections. In fact, LSTM models were evaluated in several studies in same method and in the target of outclassing background experiences ML techniques. Various reports have varying time periods for all the data from the data sets, along with different features used in model development. The previous research most applicable to this study, including LSTM models applied on the results of the financial data collection, is outlined in the following section.

The LSTM model significantly outperformed a standard RNN in the comparable assessment of forecasting using machine learning methods by Gopalakrishnan et al., although significantly surpassing a linear model for analysis of data series [34].

For the time duration from July 2014 to June 2015, for companies in the pharmacy and IT sectors, the data gathering used composed of market values on minute wise. They arrived at the fact that deep learning strategies are able to capture and forecast latent realities. Yue, Bao and Rao 's rigorous comparison of various LSTM-based models and RNNs, when extended to

six different stock indexes, enforces this conclusion[35].

They conclude that LSTM-based models typically, they appear to be reliably outperforming RNNs and that the choice of indices should rely on the maturity of the indices for comparative purposes. For the intent of this paper, with Yue, Bao and Rao 's conclusion in mind, the decision to use only one broad index was taken in order to minimise any data-related noise rather than simply illustrate the effect of the models selected.

Unlike the above reports, Dai, Chen and Zhou have not reached an especially high degree of precision[36]. Their research, however, demonstrated how the collection of features significantly impacts the accuracy of LSTM models, managing to improve accuracy by a factor of two by increasing the amount of features often used in financial trading datasets by readily available data points. They also agree that the use of a relatively more volatile Chinese stock index may have changed the outcome of the trial, further cementing the decision for this paper to use the S&P 500 index. Through using 180 features in the input layer and seven years of trading data for their binary classification LSTM model, Pereira et al. took the feature selection further [37].

The LSTM model again outperformed the other models (Random Forests, multi-layer perceptron, pseudo-random model) that were used in the comparison. In addition to the previously reached conclusions, they observed that when comparing possible losses from each model in a simulated buy-and - sell process, the LSTM model seemed Less unpredictable than any other approach. However, when considering possible real-life outcomes rather than as a framework for prospective future studies, they should not mention acquisition costs and market corrections, which is why this Project would incorporate these real-life considerations in the consideration of experimental outcomes.

In earlier studies where it has yet to be extensively explored, the choice of research issue revolving around quantitative difference in the choice of time steps in an LSTM model appears well established. The final areas of the preceding thesis-related research begin and end around the time span of the data sets used. As well as Zhang, Xu and Xue 's extensive analysis including social media sentiment that essentially further enforces the power of LSTMs in time series forecasting, previously stated studies use data ranging from 63 days to close to ten years[38].

Analytical analysis by Steven Walczak on data parameters for the forecasting of data set using ANNs, however, has shown that the use of data from one to 2 decades is of the greatest precision [39]. The Time-Series Recency Effect, this phenomenon, claims that model

building data that is closer in time to the expected values creates more precise forecasting models. It is interesting that most researchers obviously do not pursue the insights of Walczak, but it does not seem to have more effect on related experiments than on recommendations that have been tried and tested. In conclusion, given that very limited research has been conducted out on the influence of various time steps on data from financial time series using LSTM, previous research suggests that there is some innovation in this area, hence the choice of study objectives for this analysis.

CHAPTER 4: METHODOLOGY

4.0 Chapter Summary

This Chapter entails the procedures, techniques and tools that will be employed to attain the goals of the project. A detailed definition of the tools and criteria in which they will be applied, the procedural approach of the study and data collection techniques is proposed. Not that finding the appropriate research methodologies is very critical in drawing up model solutions/systems to identified problems.

In this project, we will work with historical data about the stock prices of a publicly listed company.

Using the LSTM Technique, we could apply stock index forecasting.

4.1 Data Gathering and Preprocessing

Due to its role as a surrogate for financial stocks, the selected stock index for the Deep learning model is the Agilent Technologies and Alcoa Corporation Common Stock and American Airlines Group Common Stock. LSTM model was built in Python, using Colab environment which is an online compiler allows the user to import specified key figures and export them directly as CSV (comma-separated values) files. the period of collected data varies of 15 years American airlines common stock and 26 years for Alcoa Corporation Common Stock and 56 years for Agilent Technologies, The dataset includes one-day data points, each of which includes metrics such as stock daily opening price, stock daily closing price, stock daily high price, stock daily low price, and the volume of stock trading.

4.2 Technical Indicators and Features

From Kaggle, the following data points and technical metrics were gathered

- * Stock daily open price
- * Stock daily High price
- * Stock daily Low price
- * Stock daily Close price
- * Upper Bollinger Band (UBB)
- * Lower Bollinger Band (LBB)

The technological research component was our priority. The used data sources are Kaggle website (there is endless data for various stocks here anyway) and we had to use the data sets

of the previous companies that were listed above for this particular project.

There are several factors in the dataset, such as date, open, high, low, close, and trading rate.

The Open price and Close price columns represent the original and final price. In reality, the stock exchange took place on that day.

The low price, high price columns represent minimum and maximum share price for a given day.

And the last segment is Complete Trading Value, the number of shares registered on the day of purchase / sale.

The very important thing here is to remember that the shops is not operating on weekends and public holidays therefore the data of those days are blank so it is going to affect the model in the accuracy since blank records affect any model negatively on wise of accuracy so The measurement of benefit or loss is basically being calculated by the selling price of a stock for the day; the closing price was therefore be known as the target variable.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2005-09-27	21.049999	21.400000	19.100000	19.299999	18.194910	961200
1	2005-09-28	19.299999	20.530001	19.200001	20.500000	19.326199	5747900
2	2005-09-29	20.400000	20.580000	20.100000	20.209999	19.052801	1078200
3	2005-09-30	20.260000	21.049999	20.180000	21.010000	19.806999	3123300
4	2005-10-03	20.900000	21.750000	20.900000	21.500000	20.268938	1057900
5	2005-10-04	21.440001	22.500000	21.440001	22.160000	20.891151	1768800
6	2005-10-05	22.100000	22.309999	21.750000	22.200001	20.928862	904300

Figure 6 preview of used dataset

Many experiments were executed to train the model using open, high, low, and close price as the input for the network and we have found that the best results were obtained by using the closing price as primary factor cause by developing the algorithm, the primary evidence points indicator (row) in the time sequence.

4.3 LSTM Model

Using the Keras library, a high-level neural networks API carried out on behalf of the TensorFlow library, which is an open source code library created by Google, the LSTM model was carried out in Python. It uses the aforementioned dataset, with Agilent Technologies Common Stock data points over the time period of 1999-11-18 to 2020-04-01 and Alcoa Corporation Common Stock data points over the time period of 1962-01-02 to 2020-04-01 and American Airlines Common stock data points over the time period of 2005-09-27 to 2020-04-01, and in the built model the data set of each company splits up in 80% training data and 20% testing data. Close price of the following day is expected as output (y) with the selected indicator(s) of present day and past days (equal to time step) as input (X). The methodology was designed using the LSTM network, which contains two layers of LSTM along with a dense layer. 50 units each are used by the LSTM layers and the hyperbolic tangent is the activation function.

The Adam optimization algorithm was used in the network and it is a modern alternative algorithm to the SGD algorithm which its main property is to update the network weights in training data iteratively. On the other hand the SGD algorithm keeps a single training rate for all weight updates. The previous studies have shown that Adam performs in good way in practicing, even in the comparison to other optimization methods.

In this study various numbers of time steps were tested checking how they influence the model's reliability. The time steps selected to be tested are 5, 10, 25, 50 and 100. The network was learned and evaluated during each of those phases duration to determine their relative efficiency.

4.4 Evaluation/Performance Measurement

The key objective of used LSTM approach is to forecast the stock price more than forecast the nominal company's performance; output precision was calculated by measuring the difference between the expected values of the dataset (predicted closing prices) and the true values of the dataset (actual closing prices). Hence, we used the famous two metrics which are root means square error (RMSE) and Root mean square percentage error (RMSPE). For this approach, they compiled the first evaluative calculation. RMSE and RMSPE represent the accuracy by measuring the average distance which differs between the predicted prices and actual prices, as illustrated below:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (P_i - O_i)^2}$$

$$RMSPE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{P_i - O_i}{P_i}\right)^2} * 100$$

CHAPTER 5: RESULTS AND DISCUSSION

5.0 Chapter Summary

The results of executed experiments have been discussed in this chapter. Some introductory theory on the various steps taken obtaining the results has also been elaborated.

5.1 Modeling and Prediction

In this section, a summary of various steps have been taken to train and test the prediction models has been presented.

The libraries that were used in the experiment included pandas, matplotlib, numpy, and seaborn. These were imported in Google colab (online Python compiler), the integrated IDE working with Python 2 after the relevant dependencies were installed.

The datasets were set into CSV files and uploaded online using online compiler reader. The vital step in data analysis is Exploring and visualizing data. So we have done this to obtain acquainted of the patterns that exist in the data. The next step was feature choosing which was done getting better the prediction accuracy and performance.

Extraction the Feature requires expert knowledge to mine the different features from data using various techniques. The dataset was divided into training and testing data in ratio of 4:1 respectively. The LSTM algorithm used to construct the respective predictive models. On the 20 percent data, model testing was performed and their performance depending on the RMSE and RMSPE was measured.

5.2 Results and Model Evaluation

The diagram which is shown down explains the approach performance using different numbers of time steps along the used companies' data stock

LSTM performance Plots with Agilent Technologies stock among the set of time steps:

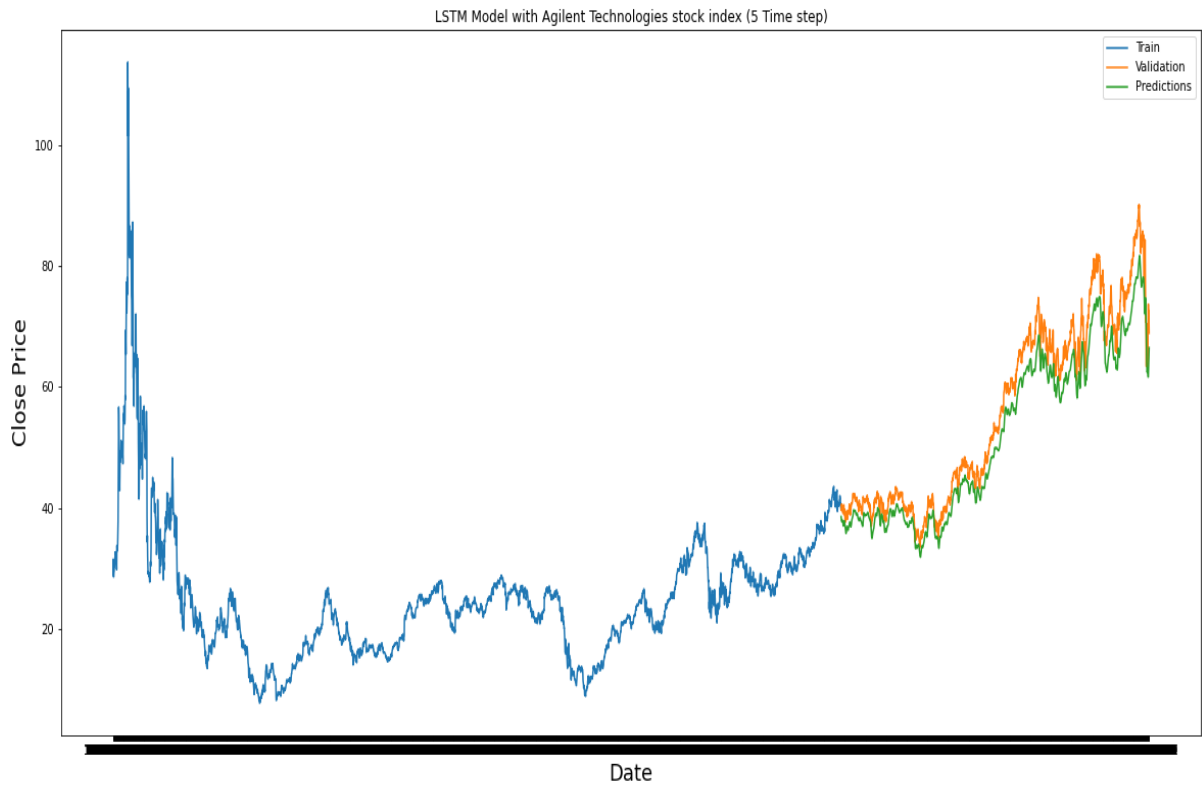


Figure 7 Agilent Technologies stock prediction using 5 Time steps

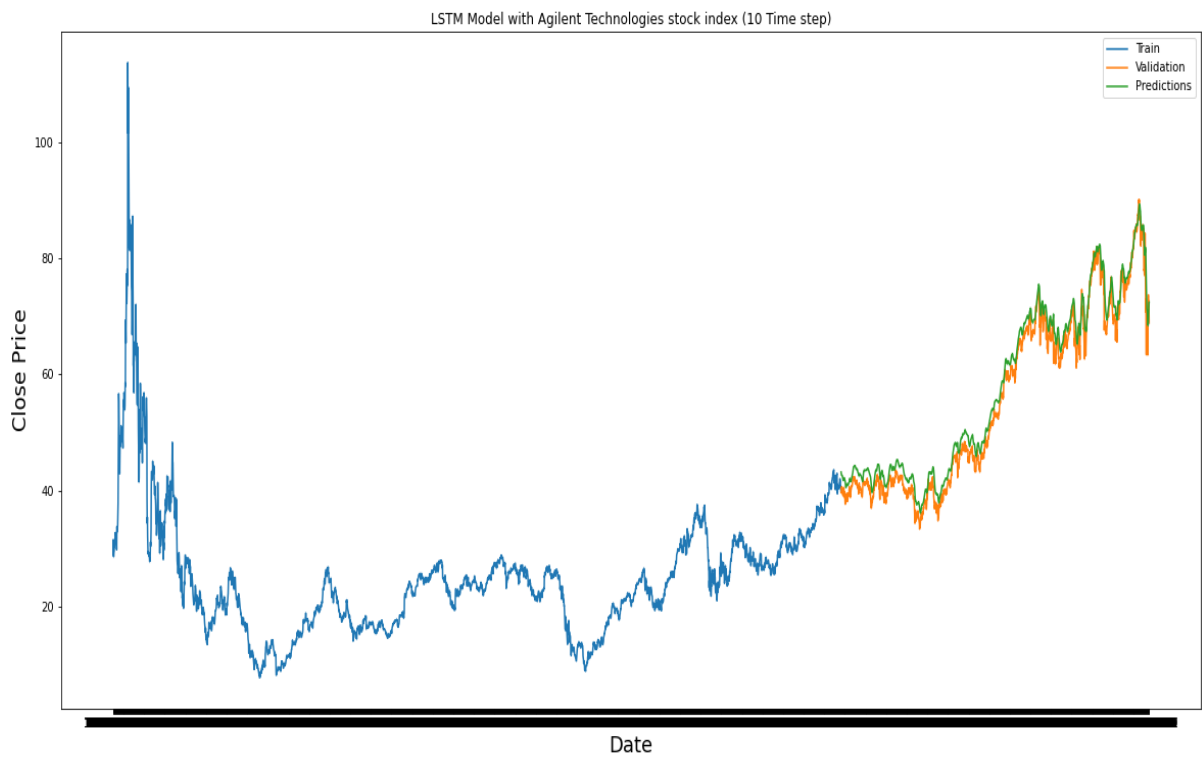


Figure 8 Agilent Technologies stock prediction using 10 time steps

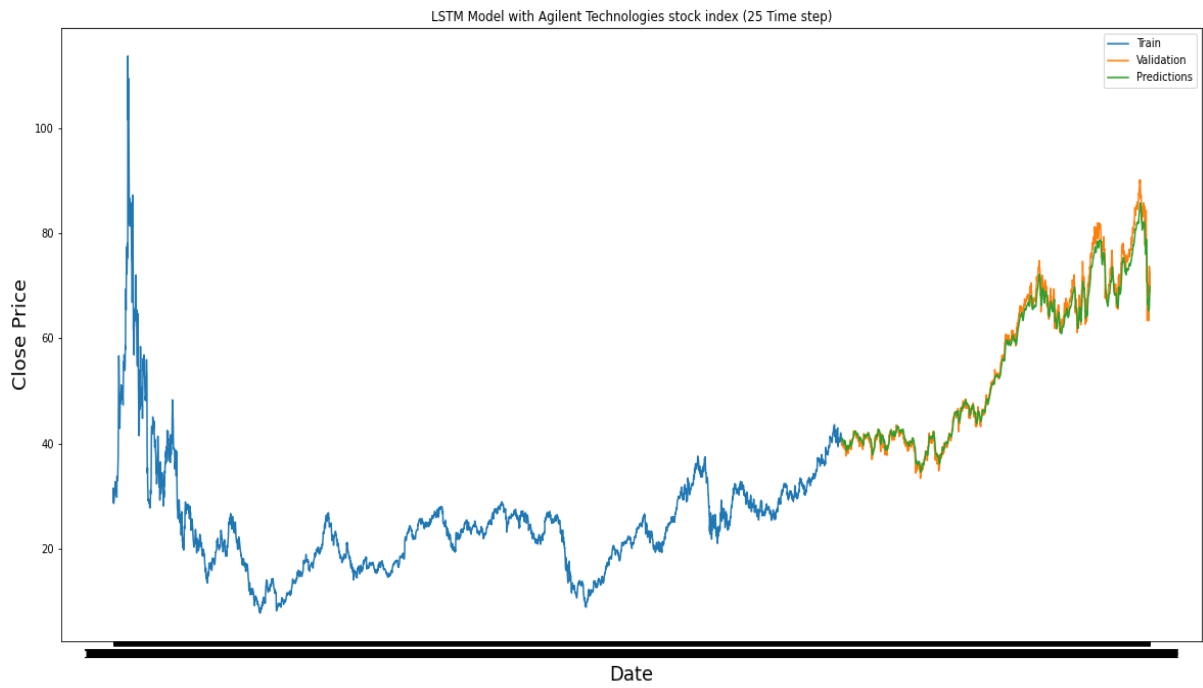


Figure 9 Agilent Technologies stock prediction using 25time steps

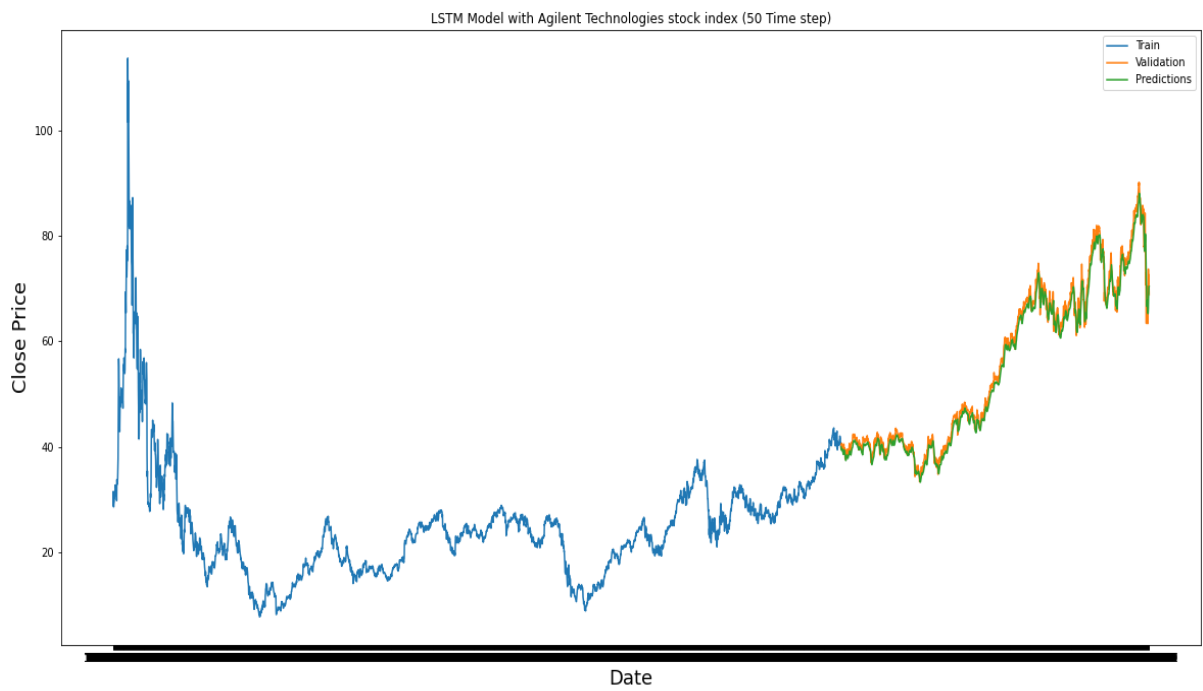


Figure 10 Agilent Technologies stock prediction using 50 time steps

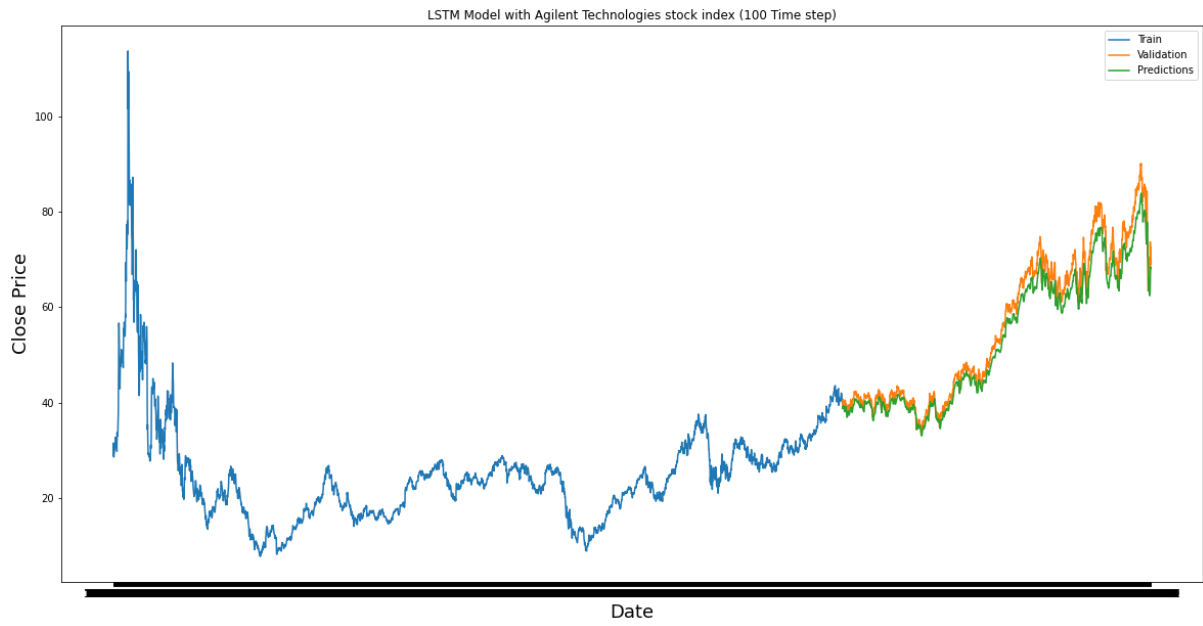


Figure 11 Agilent Technologies stock prediction using 100 time steps

LSTM performance Plot with Alcoa Corporation Stock among the selected set of time steps

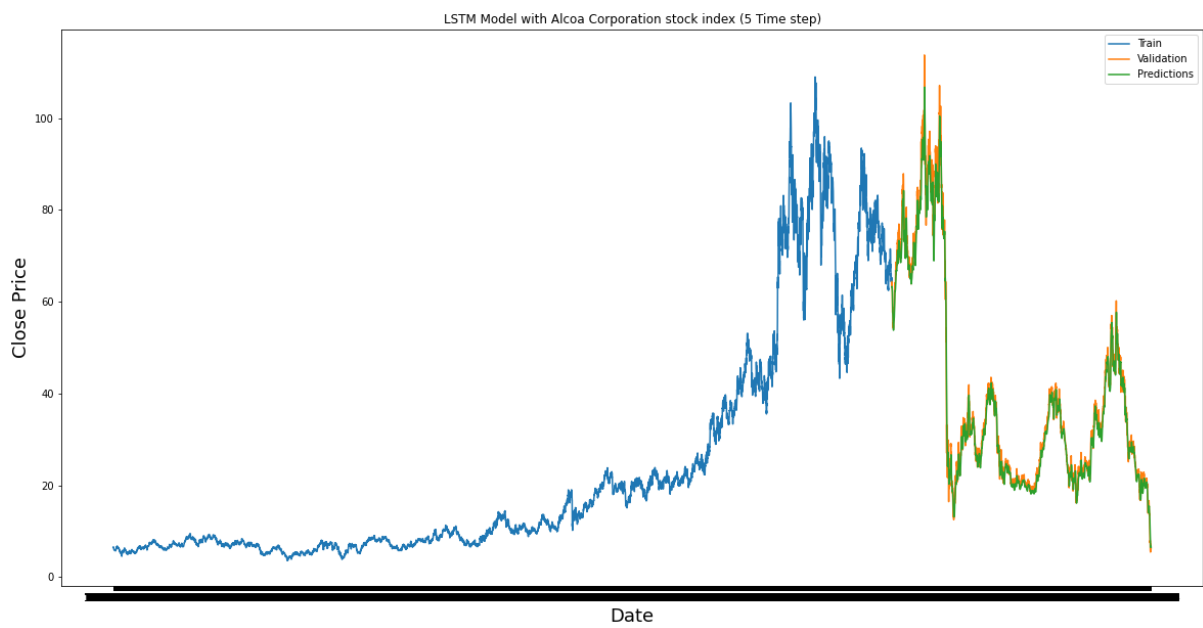


Figure 12 Alcoa Corporation Stock prediction using 5 time steps

ts

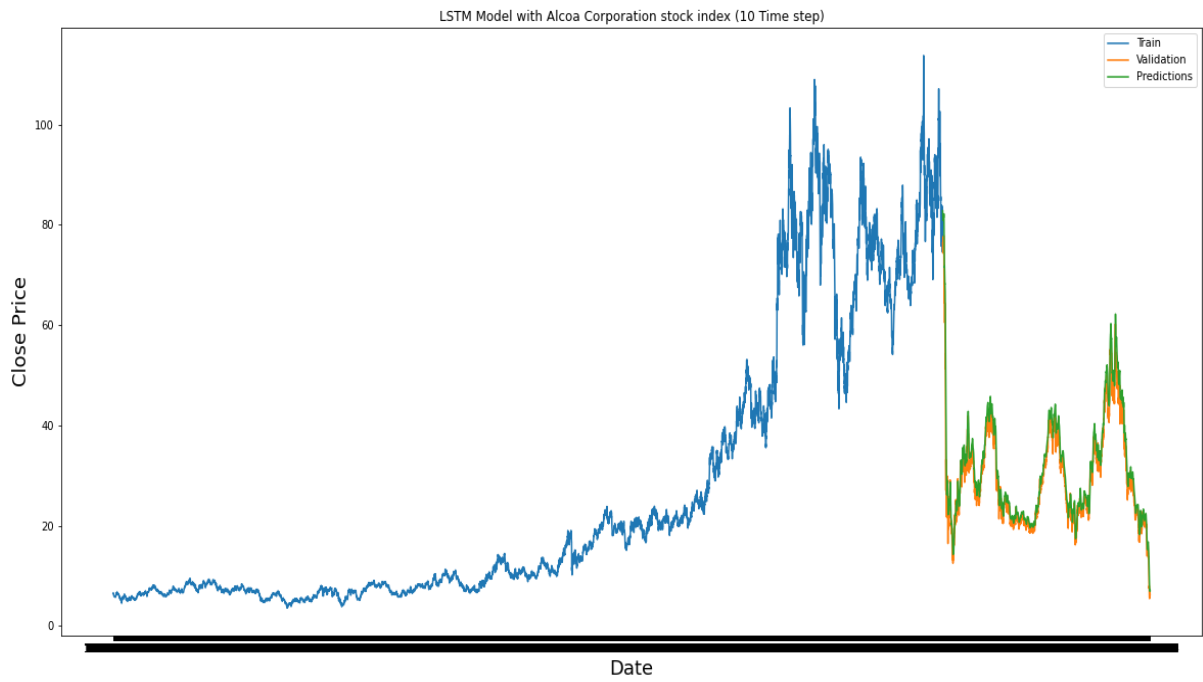


Figure 13 Alcoa Corporation Stock prediction using 10 time steps

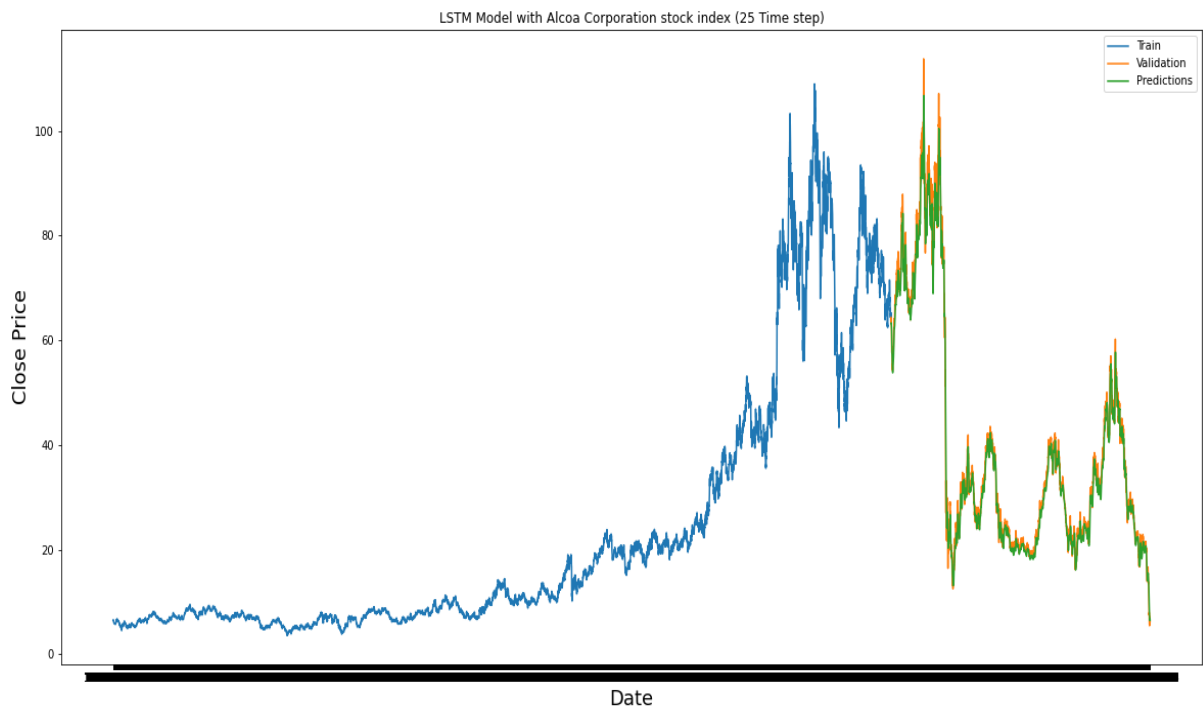


Figure 14 Alcoa Corporation Stock prediction using 25 time steps

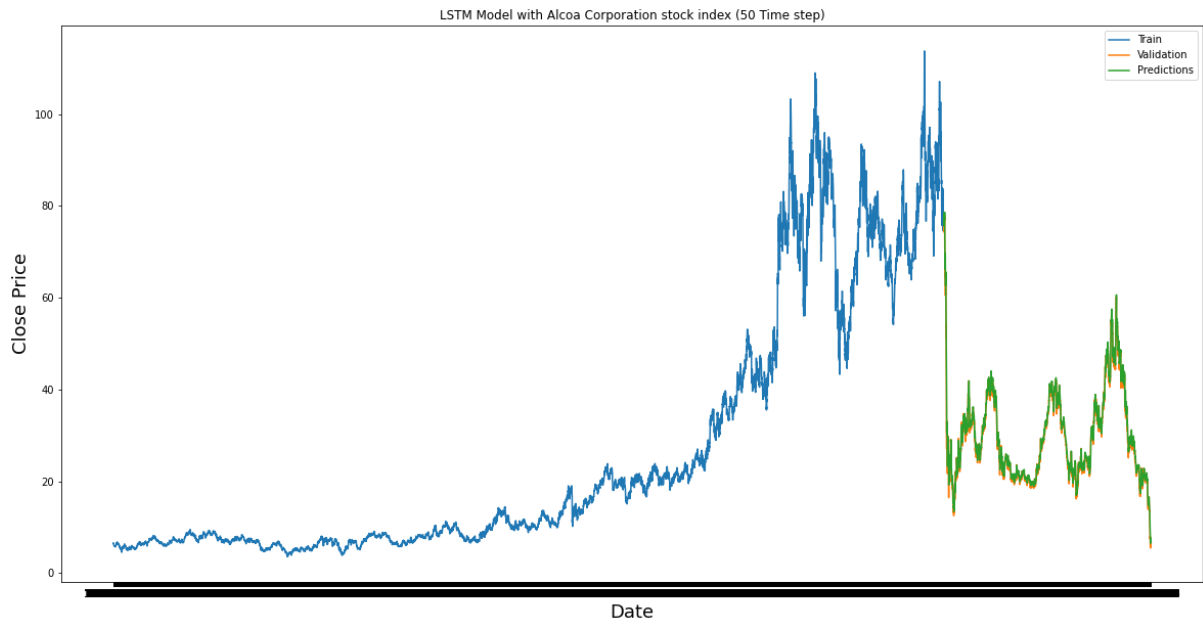


Figure 15 Alcoa Corporation Stock prediction using 50 time steps

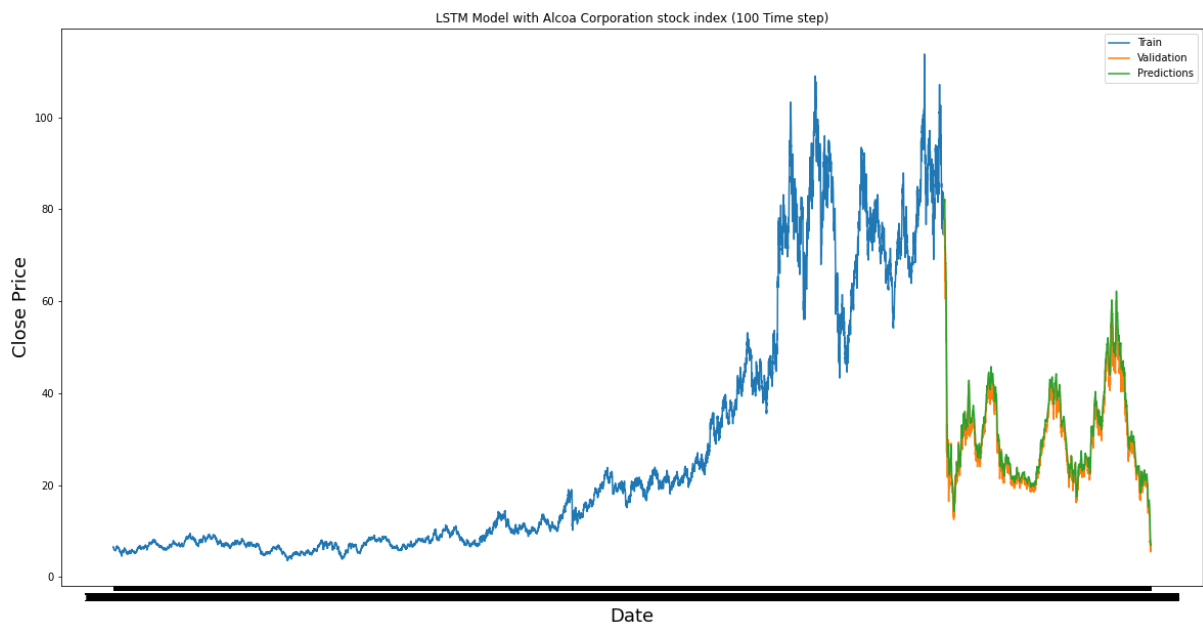


Figure 16 Alcoa Corporation Stock prediction using 100 time steps

LSTM performance with American Airlines Stock among the selected set of time steps

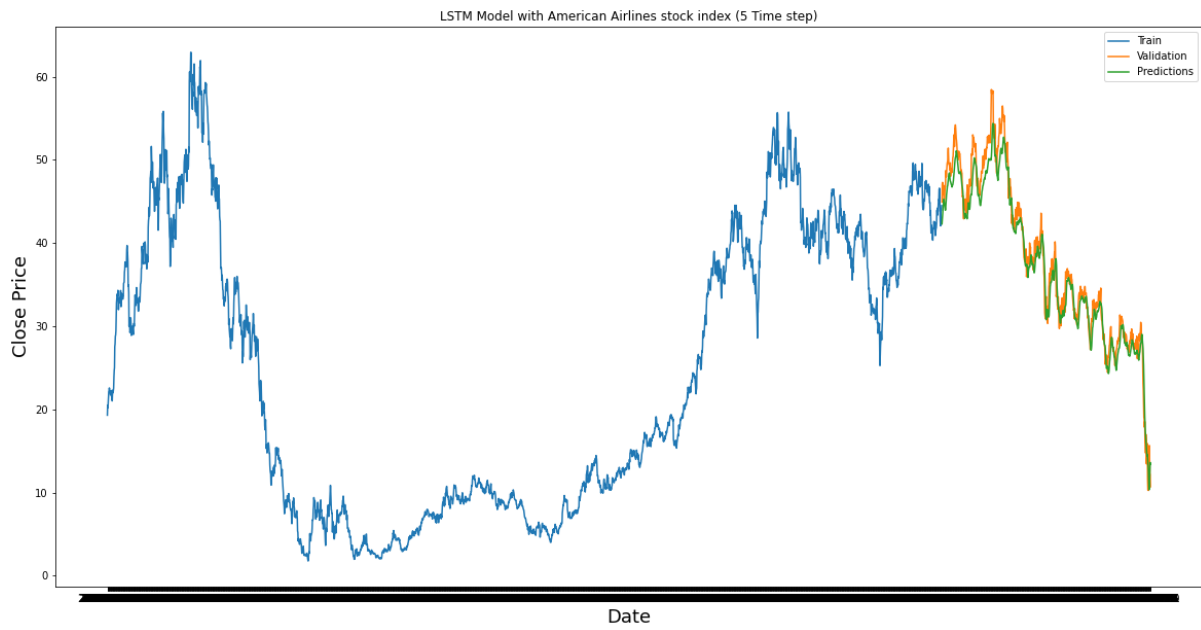


Figure 17 American Airlines Stock prediction using 5 time steps

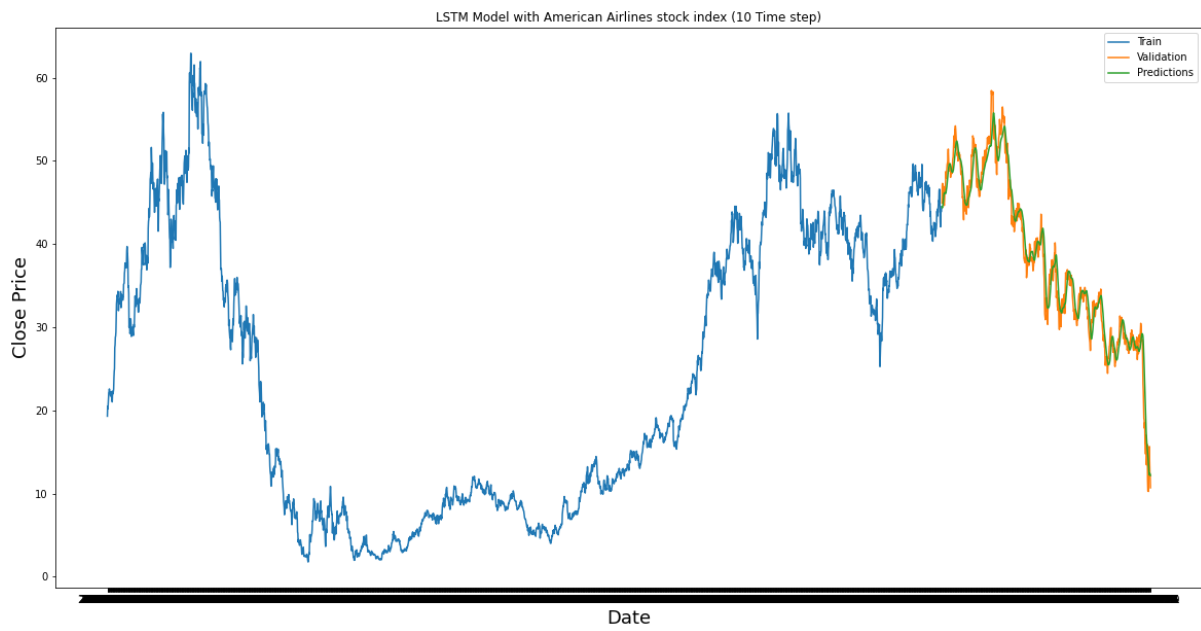


Figure 18 American Airlines Stock prediction using 10 time steps

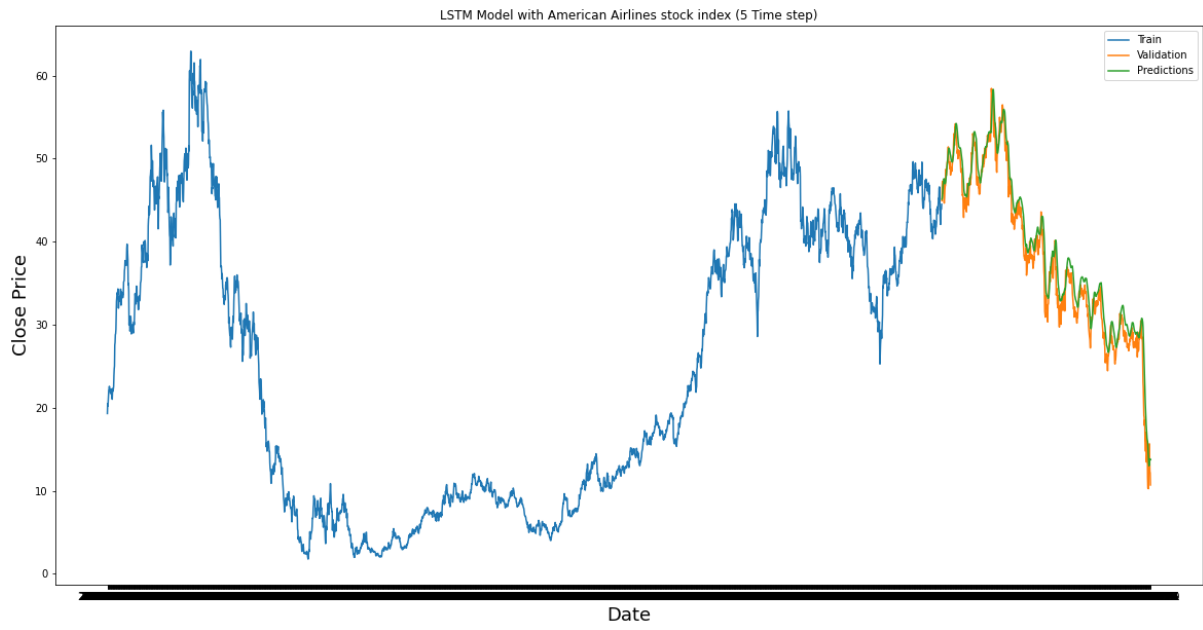


Figure 19 American Airlines Stock prediction using 25 time steps

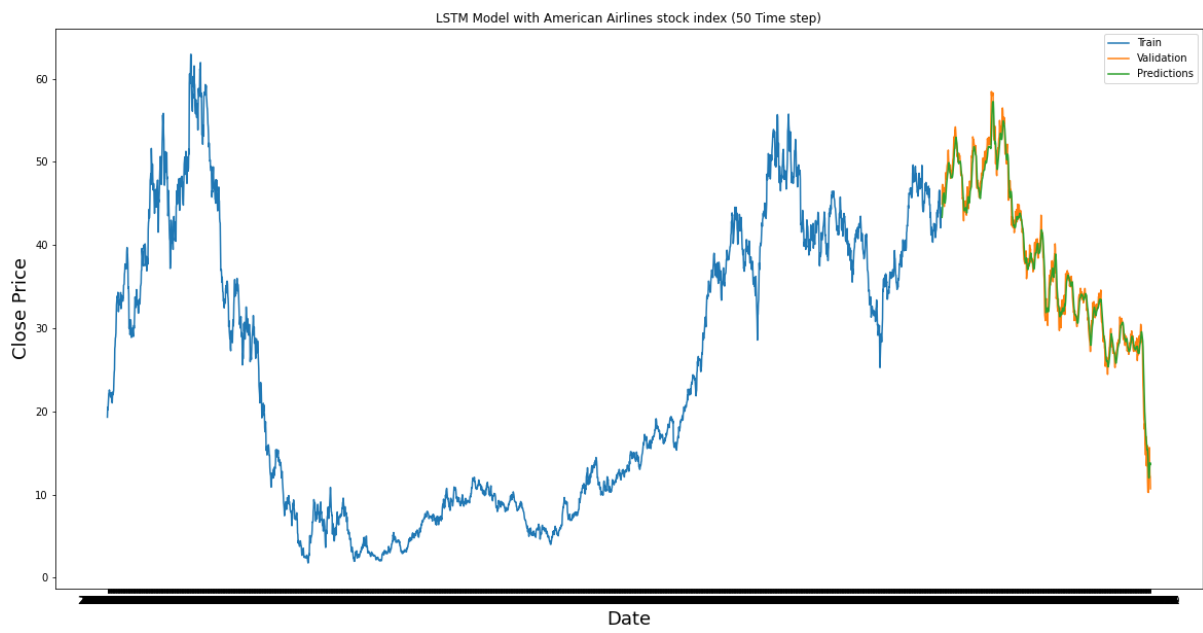


Figure 20 American Airlines Stock prediction using 50 time steps

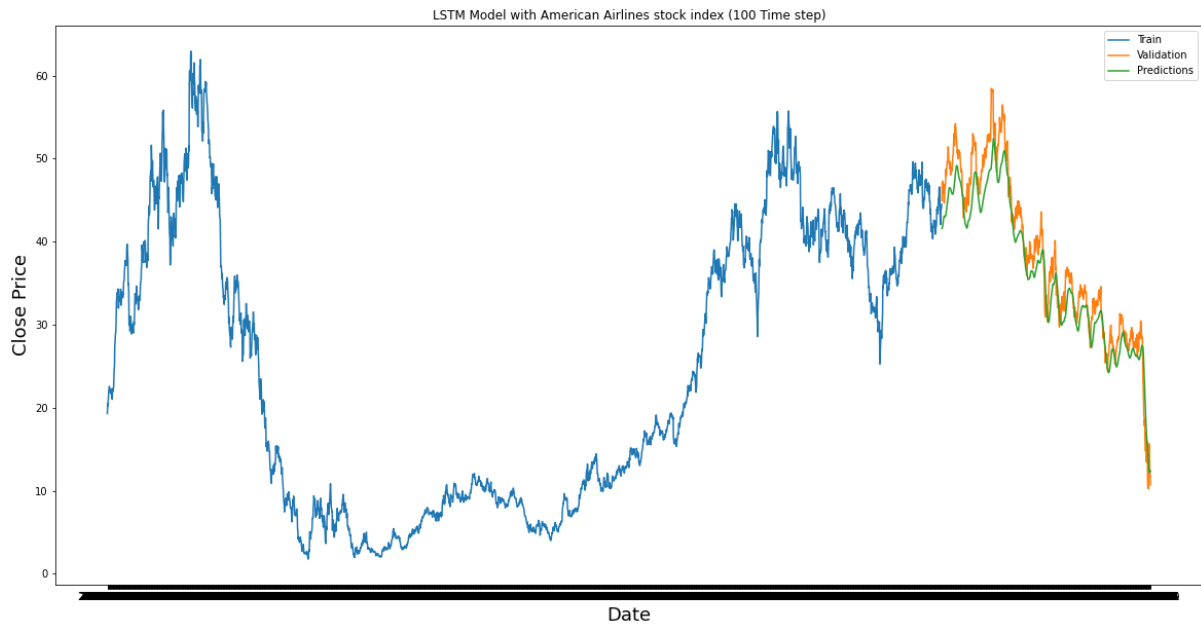


Figure 21 American Airlines Stock prediction using 100 time steps

The predictions proved the best iterations on every time step setting.

The following tables summarize the used LSTM network performance based on RMSE and RMSPE values.

Agilent Technologies stock LSTM performance measurement table:

Number of Time Steps					
	5	10	25	50	100
RMSE	1.997	1.65	1.569	1.482	5.635
RMSPE	4.02	2.647	2.918	2.31	8.0677

Table 1 Agilent Technologies stock LSTM performance measurement

Alcoa Corporation Stock LSTM performance measurement table:

Number of Time Steps					
	5	10	25	50	100
RMSE	1.431	1.61	1.45	1.20	1.75
RMSPE	5	5.2	5.5	4.8	5.3

Table 2 Alcoa Corporation Stock LSTM performance measurement

American Airlines Stock LSTM performance measurement table:

Number of Time Steps					
	5	10	25	50	100
RMSE	1.7	1.9	2.1	1.5	2.4
RMSPE	5.4	6.27	7.8	5	9.6

Table 3 American Airlines Stock LSTM performance measurement

As mentioned in the introduction part of this study, time series forecasting is an extremely complicated mission which was studied over many researches in the history, in special, the stock price varies.

The results shown in Graphs for each company as well as in the performance measurement tables for different time steps have claimed that for the different numbers of time steps measured in this study, the optimal value was recorded for the 50 time step. This function was scored approximately an RMSE of 1.488, compared to the worst-performing setting of 100 steps resulting in an RMSE of 6 for Agilent Technologies stock, in addition to this 50 time steps recorded approximately for Alcoa Corporation Stock 1.2 and 1.7 for 100 Time steps, and for the same time steps for American Airlines stock the values were recorded as 1.5 for 50 Time steps and 2.4 for 100 Time steps.

50 time steps seem to be the stellar setting proved in both the RMSE evaluation as well as RMSPE.

The answer to the study question concluded from the study is LSTM proved its capability in stock market price prediction with high level of accuracy, on the top of that this study could answer the second question of the study and proved that the time step settings effect on the LSTM performance using various datasets with different time periods. Overall the built model did quite well performance using LSTM networks.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.0 Chapter Summary

A description for the work performed was given here in the conclusion section and, ultimately, in the last section, the potential places for future work were highlighted.

6.1 Conclusion

The main target of this research project was to make the LSTM model get better performance in forecasting stock market future price and to analyze the influence of time steps settings on an LSTM model when selecting the sample from the stock prices, in particular stock of Agilent Technology, Alcoa Company Stock, and stock of American airlines.

The project proved that LSTM could predict the future price with high accuracy and claimed that time steps varying does impact the predictive power but not in large extent, however based on the results which was obtained 50 time steps to be the optimal number of time steps among the various numbers used in this model. At a higher level, the model did quite accurate predicting for closing prices of the used prices index.

6.2 Future Work

First, the obvious disadvantages to the study in that it depends entirely on the architecture of the LSTM and the setup of one factor of the methodology. Though driven by the scope and ultimate objective of forecasting share price fluctuations through data from the time series data, the research was limited to examining how shifts in time steps influence the LSTM model's performance. More distinct models or a distinction between neural networks and other types of models, such as support vector machines and supervised models, may be used in future studies to forecast stock market movements. Hybrid models, merging models in search of even improved prediction, would be another aspect which could be applied to the comparative analysis. All the implications of the nature of this study, which was agreed on due to resource limits, were these restrictions.

In addition, as a result of time constraints and a narrow scope of study, there are several potential areas of improvement relevant to the fine-tuning of the methodology as well as its components. For e.g., as prior research contains as little as one technical indicator as well as 180 technical indicators as features, more features should be added. Significant scopes for future study on the basis of this study may also be to explore which numbers of features have the best precision. More independent indices should be discussed in addition to technological measures and compared to see how good on distinct indices, the models work. Different models have previously shown that when applied to various indices where variance and other variables vary, they behave differently.

Because of the above factors that culminated in the used data being the best proxy and satisfactory to address the thesis question, this was omitted from this research. In addition, it

is definitely possible to explore more detailed data processing, model optimization, and data split between instruction, research, and assessment. However, this was excluded from this analysis due to the above-mentioned limitations, as well as major iterations of configurations that would render such an analysis highly time-consuming without assurances of improvements, as well as major mixes of configurations that would render such a study excessively time-consuming with no assurances of improvements other than further knowledge of the study.

The authors expect the thesis to support future studies by focusing on LSTM and the impact of different time periods to get an extended to the model-wise field. While the study performed itself has many ways of improving, it has ideally given a clue for continued study on recurrent neural networks and share market forecasting.

REFERENCES

- [1] Fama, Eugene F. "Efficient capital markets: A review of theory and empirical work." *The journal of Finance* 25.2 (1970): 383-417.
- [2] Malkiel, Burton Gordon. *A random walk down Wall Street: including a life-cycle guide to personal investing*. WW Norton & Company, 1999.
- [3] Biondo, Alessio Emanuele, et al. "Are random trading strategies more successful than technical ones?." *PloS one* 8.7 (2013): e68344.
- [4] Lo, Andrew W., and A. Craig MacKinlay. *A non-random walk down Wall Street*. Princeton University Press, 2011.
- [5] Glantz, Morton, and Robert Kissell. *Multi-asset risk modeling: techniques for a global economy in an electronic and algorithmic trading era*. Academic Press, 2013.
- [6] Kirkpatrick II, Charles D., and Julie A. Dahlquist. *Technical analysis: the complete resource for financial market technicians*. FT press, 2010.
- [7] Nelson, David MQ, Adriano CM Pereira, and Renato A. de Oliveira. "Stock market's price movement prediction with LSTM neural networks." 2017 International joint conference on neural networks (IJCNN). IEEE, 2017.
- [8] Kim, Kyoung-jae. "Financial time series forecasting using support vector machines." *Neurocomputing* 55.1-2 (2003): 307-319.
- [9] Melo, Bruno. "Considerações cognitivas nas técnicas de previsão no mercado financeiro." Universidade Estadual de Campinas (2012).
- [10] Batres-Estrada, Bilberto. "Deep learning for multivariate financial time series." (2015).
- [11] Sharang, Abhijit, and Chetan Rao. "Using machine learning for medium frequency derivative portfolio trading." arXiv preprint arXiv:1512.06228 (2015).
- [12] Heaton, J. B., Nicholas G. Polson, and Jan Hendrik Witte. "Deep learning in finance." arXiv preprint arXiv:1602.06561 (2016).
- [13] Greff, Klaus, et al. "LSTM: A search space odyssey." *IEEE transactions on neural networks and learning systems* 28.10 (2016): 2222-2232.
- [14] Hochreiter, Sepp. "JA1 4 rgen Schmidhuber (1997). "Long Short-Term Memory".
Neural Computation 9.8.
- [15] Graves, Alex. "Connectionist temporal classification." *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, Berlin, Heidelberg, 2012. 61-93.
- [16] Graves, Alex, et al. "A novel connectionist system for unconstrained handwriting recognition." *IEEE transactions on pattern analysis and machine intelligence* 31.5 (2008): 855-868.
- [17] Chen, Kai, Yi Zhou, and Fangyan Dai. "A LSTM-based method for stock returns prediction: A case study of China stock market." 2015 IEEE international conference on big data (big data). IEEE, 2015.
- [18] Di Persio, Luca, and Oleksandr Honchar. "Artificial neural networks approach to the forecast of stock market price movements." *International Journal of Economics and Management Systems* 1 (2016).
- [19] Stringham, Edward. "The extralegal development of securities trading in seventeenth-

- century Amsterdam." *The Quarterly Review of Economics and Finance* 43.2 (2003): 321-344.
- [20] Bergström, Carl, and Oscar Hjelm. "Impact of Time Steps on Stock Market Prediction with LSTM." (2019).
- [21] Ball, Ray, S. P. Kothari, and Charles E. Wasley. "Can we implement research on stock trading rules?." *Journal of Portfolio Management* 21.2 (1995): 54-63.
- [22] Kirman, Alan. "Economic theory and the crisis." *real-world economics review* 51 (2009): 80-83.
- [23] Shen, Peter. "The P/E ratio and stock market performance." *Economic review-Federal reserve bank of Kansas city* 85.4 (2000): 23-36.
- [24] Zell, Andreas. "Chapter 5.2." *Simulation Neuronaler Netze [Simulation of Neural Networks](in German)(1st ed.)*. Addison-Wesley (1994).
- [25] Dreman, David N., and Michael A. Berry. "Overreaction, underreaction, and the low-P/E effect." *Financial Analysts Journal* 51.4 (1995): 21-30.
- [26] Lo, Andrew W., and A. Craig MacKinlay. *A non-random walk down Wall Street*. Princeton University Press, 2011.
- [27] Al-Hnaity, Bashar, and Maysam Abbod. "A novel hybrid ensemble model to predict FTSE100 index by combining neural network and EEMD." *2015 European Control Conference (ECC)*. IEEE, 2015.
- [28] G. Hinton, Y. Bengio and Y. LeCun, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May 2015.
- [29] Maas, Andrew L., Awni Y. Hannun, and Andrew Y. Ng. "Rectifier nonlinearities improve neural network acoustic models." *Proc. icml*. Vol. 30. No. 1. 2013.
- [30] Zell, Andreas. "Chapter 5.2." *Simulation Neuronaler Netze [Simulation of Neural Networks](in German)(1st ed.)*. Addison-Wesley (1994).
- [31] A. Senior, F. Beaufays and H. Sak, "Long Short-Term Memory Recurrent Neural Network Architectures For Large Scale Acoustic Modeling," Google, USA. Feb. 2014.
- [32] P. Frasconi, P. Simard and Y. Bengio, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157-166, Mar. 1994.
- [33] Bonassi, Fabio, et al. "LSTM neural networks: Input to state stability and probabilistic safety verification." *Learning for Dynamics and Control*. 2020.
- [34] Selvin, Sreelekshmy, et al. "Stock price prediction using LSTM, RNN and CNN-sliding window model." *2017 international conference on advances in computing, communications and informatics (icacci)*. IEEE, 2017.
- [35] Bao, Wei, Jun Yue, and Yulei Rao. "A deep learning framework for financial time series using stacked autoencoders and long-short term memory." *PloS one* 12.7 (2017): e0180944.
- [36] Chen, Kai, Yi Zhou, and Fangyan Dai. "A LSTM-based method for stock returns prediction: A case study of China stock market." *2015 IEEE international conference on big data (big data)*. IEEE, 2015.
- [37] Rueda Rojas, Edwin Jahir. *Prediccion De Series Financieras Con Redes Neuronales Recurrentes*. Diss. Universidad Industrial de Santander, Escuela De Ing. De Sistemas,

2018.

- [38] Zhang, Gaowei, Lingyu Xu, and Yunlan Xue. "Model and forecast stock market behavior integrating investor sentiment analysis and transaction data." *Cluster Computing* 20.1 (2017): 789-803.
- [39] Walczak, Steven. "An empirical analysis of data requirements for financial forecasting with neural networks." *Journal of management information systems* 17.4 (2001): 203-222.
- [40] Qu, Yaxin, and Xue Zhao. "Application of LSTM Neural Network in Forecasting Foreign Exchange Price." *Journal of Physics: Conference Series*. Vol. 1237. No. 4. IOP Publishing, 2019.