IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Stock Market Prediction Using Machine Learning

Dikshant Lade[1], Abhishek Patil[2], Pratik Yenkar[3], Shubham Alone[4], Prof. Sachin Dhawas[5]
Students, Department of Information Technology[1,2,3,4]
Professor, Department of Information Technology[5]
Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur, Maharashtra, India

Abstract: *In the finance world stock trading is one ofthe most important activities. Stock market prediction is an act of trying to determine the future value of a stock other financial instrument traded on a financial exchange. This paper explains the prediction of a stock using Machine Learning. The technical and fundamental or the time series analysis is used by the most of the stockbrokers while making the stock predictions. The programming language is used to predict the stock market using machine learning is Python. In this paper we propose a Machine Learning (ML) approach that will be trained from the available stocks data and gain intelligence and then uses the acquired knowledge for an accurate prediction. In this context this study uses a machine learning technique called Support Vector Machine (SVM) to predict stock prices for the large and small capitalizations and in the three different markets, employing prices with both daily and up-to-the-minute frequencies*

Keywords: Stock Prediction, Machine Learning, Neural Networks, Deep Learning, Recurrent Neural Network

I. INTRODUCTION

Artificial Intelligence is changing virtually every aspect of our lives. Today's algorithmsaccomplish tasks that until recently only expert humans could perform. As it relates tofinance, this is an exciting time to adopt a disruptive technology that will transform howinvestment decisions are made on a broad scale.Models that explain the returns of individual stocks generally use company and stockcharacteristics, e.g., the market prices of financial instruments and companies' accountingdata. These characteristics can also be used to predict expected stock returns out-of-sample.Most studies use simple linear models to form these predictions. An increasing bodyof academic literature documents that more sophisticated tools from the Machine Learning(ML) and Deep Learning (DL) repertoire, which allow for nonlinear predictor interactions,can improve the stock return forecasts.The main goal of this project is to investigate whether modern DL techniques can be utilizedto more efficiently predict the movements of the stock market. Specifically, we train LSTM networks with time series price-volume data and compare their out-of-sample returnpredictability with the performance of simple logistic regressions (our baseline models).

II. LITERATURE REVIEW

Traditional Machine Learning Techniques The authors of [8] studied the behavior of the stock market and determine the best fit model from the several traditional machine learning algorithms which included Random Forest (RF), Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbor (KNN), and Softmax for stock market prediction. The authors conducted a comparative study of these approaches, several technical indicators were applied to the data that was gathered from different data sources including Yahoo and NSE-India. The accuracy of each model was compared and it was observed that RF gave the most satisfying results for large datasets whereas for small datasets Naive Bayesian revealed the highest accuracy. Another observation made was, as the count of technical indicators was reduced the accuracy of the models decreased. The paper [9] used various TF-IDF features to forecast the prices of the stocks of the next day based on the data that was gathered from different news channels. The authors computed TF-IDF weights to count the word score. Finally, an HMM model was generated to calculate the probability of a sequence and contained the probabilities of switching values. From this model the authors observed a trend of positive and negative

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

predictions which were partially matching and showed an error of 0.2 to 4%, however increasing the size of the dataset, employing various machine learning algorithms or increasing the number of technical indicators and input features can lead to higher accuracy. Traditionally, only historical data was applied for forecasting share prices. However, analysts now recognize that relying purely on historical data isn't accurate because a lot of other factors are key to determining the stock price. In the paper [10] the authors study and apply different methods to predict stock prices but a high rate of accuracy is still not achieved even after analyzing major factors affecting the stock price. The authors have reviewed major techniques such as SVM, Regression, Random Forest, etc. and also analyzed hybrid models by combining two or more techniques. According to the authors, some models work better with historical data than with sentiment data. Fusion algorithms yielded results with higher predictions. The paper [11] by Kunal Pahwa et al uses Linear Regression, the supervised learning approach to predict stock prices. The proposed research work basically outlines the entire process of using a given AICECS 2021 Journal of Physics: Conference Series2161 (2022) 012065 IOP Publishing doi:10.1088/1742-6596/2161/1/012065 3 dataset to forecast the closing value, by studying the GOOGL stock and extracting approximately 14 years of data. The paper [12] by Meghna Misra et al concludes that predictions made using the Linear Regression Model have an enhanced accuracy rate after applying the Principal Component Analysis (PCA) on the data for picking out the most relevant components. SVM demonstrates high accuracy on non-linear classification data, Linear regression is preferred for linear data because of its high confidence value, a high accuracy rate was observed on a binary classification model using Random Forest Approach and the Multilayer Perceptron (MLP) yielded the least amount of error while making predictions. Many of the aforementioned techniques are not just limited to stock price prediction but can also be used broadly in the financial markets as the authors in the paper [13] conclude by studying the application of machine learning models to analyze financial trading and to design optimal strategies for the same. After performing a quantitative analysis of different techniques, the authors recommend delving further into behavioural finance to evaluate market or investor psychology to understand market fluctuations. The authors propose to make use of text mining and machine learning methods to monitor public interaction on digital financial trading platforms. 2.2. Deep Learning and Neural Networks Yoojeong Song and Jongwoo Lee from Sookmyung Women's University observed that from a large set of Input Features only a few actually affect the stock price, they hence studied these input features and wished to determine the ones which can be employed for the best prediction of stock value. The paper [14] proposes three different Artificial Neural Network models which include the use of multiple-input features, binary features and technical features to find the best approach to achieve the aim. The accuracy of the models was computed and revealed that the model with binary features showed the best accuracy and concluded that binary features are lightweight and are most suitable for stock prediction. However, the study has some limitations in that converting the features to binary eliminates some of the relevant information for prediction. Delving into specific techniques methods such as the Multi-Layer Perceptron Model (MLP), Sequential Minimal Optimizations and the Partial Least Square Classifier (PLS) have been studied and applied on the Stock Exchange of Thailand Data in the paper 'Stock Closing Price Prediction Using Machine Learning [15] by PaweeWerawithayaset where SET100 stocks were used by using 12 months' worth of data. Although the paper doesn't focus on long term investment decisions, it does present conclusive evidence that the Partial Least Square method yielded minimum error value followed by Sequential Minimal Optimization and the Multilayer Perceptron showed the maximum error value out of the three algorithms chosen for the particular dataset. [16] focuses on the effect of the indices in the stock price prediction. The model identifies the variables and relationship between the indices and overcomes the limitations of the traditional linear model and uses LSTM to understand the dynamics of the S&P 500 Index. The paper also analyses the sensitivity of internal memory of LSTM modelling. However, the study has some limitations, the difference between the predictive value and actual value becomes large after a certain point and thus cannot be used to develop a system to give a profitable trading strategy. [17] proposes a system that would recommend stock purchases to the buyers. The approach opted by the authors combines the prediction from historical and real-time data using LSTM for predicting. In the RNN model, latest trading data and technical indicators are given as input in the first layer, followed by the LSTM, a compact layer and finally the output layer gives the predicted value. These predicted values are further integrated with the summarized data which is collected from the news analytics to generate a report showing the percentage in change. AICECS 2021 Journal of Physics: Conference Series2161 (2022) 012065 IOP Publishing doi:10.1088/1742-6596/2161/1/012065 4 2.3. Time Series Analysis The paper "Share Price Prediction using Machine Learning Technique" [18] represented the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

stock price in the form of a time series and avoided the complications endured by the model in the training process. The paper used normalised data and a Recurrent Neural Network model for making the predictions that predicted values that were very close to the actual ones and thus, the author's considered machine learning algorithms best for forecasting the stock prices. The authors of [19] noticed an impact of daily sentiment scores of various companies on the values of their stock prices. As the information or news that gets posted on various social media platforms about/by an organisation can influence the investors to buy/sell the stocks of the company thus affecting its stock value. The authors thus proposed a model for stock market prediction that employed sentimental analysis as one of the indicators. The algorithm made use of data collected from various online platforms such as Yahoo Finance and positive/negative/neutral tweets as features for the prediction and computed the stock price movement using opening and closing price of stock for the respective company. Another interesting aspect noted by the authors was the effect of holidays, seasonality, trends and non-periodic data and designed a curve time series model which took all these components into account. This culminated in the authors employing the Generalised Additive Model for maximizing prediction quality and to accommodate new components. Finally, Multiple Linear Regression was used to train the model and predict the prices of stocks for the next 10 days. 2.4. Graph-Based Approach A rather interesting approach has been adopted by Pratik Patil et al in their paper [20] which visualizes the stock market as a graphical network in a rather unique way and the authors have included both correlation and causation using historical price data as well as applying sentiment analysis which is highly useful in taking into account different factors that determine the stock price. The Graph Convolutional Network model proposed in this paper is vulnerable to the detonating inclination issue as nodes with more significant levels will have bigger worth in their convolved feature portrayal, while nodes with a more modest degree will have more modest worth in feature representation. An answer for this issue can diminish the intricacy of the model training. It will likewise be intriguing to check the exhibition of GCN on more conventional time series estimating issues. Raehyun Kim et al [21] proposed a Hierarchical Attention Network for Stock Prediction (HATS) to forecast share prices and stock index market movement by applying the concept of Graph Theory and Graph Neural Networks. The authors proposed this new method to selectively cluster the available data on the different relations and add that information to the representation. The Hierarchical Attention Network is key to improving the performance and is used to assign different weight values for selection of information based on its importance and relevance. Another important work in this direction is done by researchers Yang Lieu et al [22] in which they have used information characteristics of tuples in building a knowledge graph which later on is used for feature selection. In the proposed work the authors have used the CNN to extract features and build the semantic information of the news related to the stock. The combination of deep learning and Knowledge graph have proven to be useful for effective feature extraction retaining semantics. However, due to the limited training sets of financial information, knowledge graph extraction seems to be challenging.

III. FUNDAMENTAL AND TECHNICAL ANALYSIS

The stock market's movements are analyzed and predicted using fundamental and technical analysis. The fundamental analysis includes an in-depth analysis of the performance of a company. This method is ideally suited for forecasting for the long term. Moreover, it has been occasionally proved to be a strong indicator of the movement of stock prices in the works of (Checkley, Higón, and Alles 2017, Tsai and Wang 2017) and (Zhang et al. 2018). Whereas, technical analysis is a method of assessing stocks by examining market behavior, past prices and volume produced statistics. It's looking peaks, bottoms, trends, and other factors that influence the price movement of a stock. This method supports short-term forecasts (Drakopoulou 2016). The following research forecast the future movement of stock-prices based on technical analysis (Putriningtiyas and Mochammad 2017, Nti, Felix, and Asubam 2020). Both these methods are having their limitations and fail to give expected results. Moreover, the results produced by these tools can be interpreted by the experts only and also these tools require a lot of time in a modern dynamic trading environment.

IV. CHALLENGES AND OPEN PROBLEMS

Stock market analysis and prediction continue to be an interesting and challenging problem. As more data are becoming available, we face new challenges in acquiring and processing the data to extract knowledge and analyse the effect on

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

stock prices. These challenges include issues of live testing, algorithmic trading, self-defeating, long-term predictions, and sentiment analysis on company filings.

Regarding live testing, most of the literature on stock analysis and prediction claim that the proposed techniques can be used in real time to make profits in the stock market. It is a big claim to make because an algorithm may work fine on backtesting in controlled environments, but the main challenge is live testing, because a lot of factors like price variations, and uneventful news and noise exist. One such example is the Knight Capital Tragedy1 where the company suffered a loss of 440 million. Hence, a viable research direction would be to understand how some of the popular stock analysis techniques work in live or simulated environments.

Algorithmic trading systems have changed the way stock markets function. Most of the trading volumes in equity futures are generated by algorithms and not by humans. While algorithmic trading gives benefits like reduced cost, reduced latency, and no dependence on sentiments, it also brings up challenges for retail investors who do not have the necessary technology to build such systems. Today, it is common to see events where panic selling is triggered due to these systems and hence the markets overreact. As a result, it becomes more difficult to evaluate market behaviour. With new algorithms continuing to flood the markets every day, comparison of the efficacy and accuracy of these algorithms pose yet another challenge.

An interesting aspect of this research area on stock market prediction is its self-defeating nature. In simple words, if an algorithm can use a novel approach to generate high profits, then sharing it in any way to the market participants will render the novel approach useless. Thus, state of the art algorithms which are trading out there in the markets is proprietary and confidential. The research or methodology behind such algorithms is generally never published.

Researchers, analysts, and traders mostly focus on short term prediction of stock prices compared to longer term, i.e., weekly or monthly predictions based on historical data. Some good approaches to long term price prediction already exist such as the ARIMA. Stock markets are generally more predictable in the longer term. Several newer ANN approaches such as the LSTM and RNN are now being explored and compared against existing approaches in predicting long term dependencies in the data and the stock prices, which are equally valuable to the investors and data scientists.

Recently, due to the rising influence of social media on many aspects of our lives, a lot of attention is being given to sentiment analysis based on Twitter or news data. Social media data can be unreliable and difficult to process, and fake news is being posted on the web by multiple sources. A good alternative to these or additional resource would be the quarterly or annual reports filed by the companies (e.g., 10-Q and 10-K) for stock prediction to apply sentiment analysis. These documents, if decoded correctly, give a major insight into a company's status, which can help to understand the future trend of the stock.

V. MACHINE LEARNING MODELS

Machine learning is a field of Artificial Intelligence (AI). It has been widely studied and explored for the prediction of stock price direction. Machine learning tasks are usually classified as supervised and unsupervised learning. Supervised learning techniques are used to collect training data in which instances are marked with labels and each label shows a specific instance's class (Ashfaq et al. 2017). Supervised learning techniques like Linear Regression, Support Vector Machine (SVM), Random Forest, Nearest Neighbor, and Decision Trees can attempt to forecast historical data-based stock market prices and patterns as well as provide useful historical price analysis. Analysts in (Milosevic 2016) conducted a manual collection of features, picked 11 related fundamental ratios, and implemented various machine learning algorithms to stock forecasting. It suggests that against approaches like SVM and Naïve Bayes, Random Forest obtained the highest F-Score of 0.751. Researchers in (Chou and Thi-Kha 2018) proposed a novel approach that develop a stock price forecasting expert system for Taiwan construction companies based on a metaheuristic firefly algorithm and least squares support vector regression (MetaFA-LSSVR), to enhance predicting accuracy. The established hybrid system performed exceptionally well in terms of the one-day prediction of 2597.TW stock prices were better than that of any construction company stock prices with a MAPE of 1.372% and an R2 of 0.973. So, the suggested system can be utilized as a decisive tool for short-term stock price forecasting. Also, researchers in (Zhang et al. 2018) establish a system of prediction of stock price movements that can forecast both the movement of stock prices and their rate of growth (or decline) within a predetermined duration of the prediction. Using historical data from the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Shenzhen Growth Enterprise Market (China), they trained a random forest model to divide different stock clips through four classes by gradient of their close prices. Analysts in (Lv et al. 2019) have recently synthetically tested various ML algorithms and recorded regular stock trading success within transaction costs and no transaction costs. Between 2010 and 2017, they used 424 S&P 500 Index Component Stocks (SPICS) and 185 CSI 300 Index Component Stocks (CSICS), comparing traditional machine learning algorithms with advanced DNN models. Traditional ML algorithms involve Logistic Regression, Random Forest, Classification and Regression Tree (CART), SVM, and eXtreme Gradient Boosting, whereas DNN architectures contain MLP, Deep Belief Network (DBN), GRU, RNN and LSTM. Their results indicate that in many of the directional assessment indicators conventional ML algorithms have higher performance without knowing transaction costs, but DNN models show better results given transaction costs.

VI. DEEP LEARNING MODELS

A new study area of ML has been brought into existence since 2006, called deep structured learning or deep learning. The developer and scientist do not need to select features manually compared to traditional machine learning. Alternatively, using deep learning, these features can be generated automatically. Deep learning involves learning different levels of description and interpretation that give a better idea of data like images, sound, and text (Xiang et al. 2016). Classification techniques such as the k-nearest neighbor, Naïve Bayes algorithms, are commonly used to predict the stock price trend (Khedr, Yaseen, and Yaseen 2017) rise or fall but this paper aims to solve the problem as a simulation by using Deep Neural Networks. Also, the neural network methods were examined for India's stock markets in (Kumar and Murugan 2013) where performance analysis is required. It had used metrics such as RMSE, MAE, MAPE. An analysis of the interdependence between stock price and stock volume was conducted in (Abinaya et al. 2016) for 29 selected companies listed in NIFTY 50. The proposed research focuses on applying deep learning algorithms for prediction of stock prices such in (Heaton, Polson, and Witte 2017) and (Jain, Gupta, and Moghe2018). Deep neural networks can be defined as non-linear algorithms that can map non-linear functions. Various kinds of deep neural network architectures are used, depending on the form of application such as Recurrent Neural Networks (RNN), Long Short Term Memory(LSTM), CNN(Convolutional Neural Network). They were deployed in multiple fields mainly image recognition, natural language processing and analyzing time series. Studies in the field of analyzing financial time series utilizing neural network (Heaton, Polson, and Witte 2017) methods used various input variables to predict the return of the stock. Authors in (Vargas, De Lima, and Evsukoff2017) used RNN architecture to predict the S&P 500 index intraday inertial movements which use a set of technical indicators as input and financial news titles. Analysts in (Roondiwala, Patel, and Varma 2017) used one of the most accurate predicting techniques to forecast stock returns of NIFTY 50 using the LSTM, which allows investors to have a good understanding of the potential stock market situation. They collected 5 years of historical data and used it to train and validate their model. Also, (Chen and He 2018) implemented a CNN model to make the share price prediction and used a Conv1D feature to handle the 1D data in the convolution layer. Different stock data examined the result and finally indicated that the CNN model is reliable and can be used to make accurate predictions even if the original data is 1D sequential. In (Hiransha et al. 2018), researchers used four Deep learning architectures predict stock prices of NSE and NYSE. They trained four RNN, LSTM, CNN, and Multilayer Perceptron (MLP) networks with NSE's TATA MOTORS stock price. CNN has worked better than the other three networks because it is capable of detecting the system's sudden changes while the next instant is predicted using a specific window. Also, three different deep learning architectures (RNN, LSTM, and CNN) were used in (Selvin et al. 2017) to predict the price of NSE listed companies then compare their performance. They have used a sliding window model to forecast future market values on a short-term basis. Models' performance was measured using a percentage error. A comparative analysis of different deep neural network methods introduced for stock price forecasting application is done in (Jain, Gupta, and Moghe2018). The models are trained on daily and historical stock price data which involves values of Open, High, Low, and Close price. Also, the researchers in (Jain, Gupta, and Moghe2018) proposed an approach based on the combining of layers of LSTM and CNN techniques called Conv1D-LSTM. The model's performance is measured using RMSE, MAE, and MAPE. Such errors are found to be very small in the Conv1D-LSTM model as compared to CNN and LSTM. The researchers in (Soui et al. 2020) introduce a novel deep learning-based approach for predicting financial firm bankruptcy that combines the feature extraction and classification phases into a single model. The results show that the Stacked Auto-Encoders (SAE) with

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

softmax classifier is more effective than other methods for accurately predicting corporate bankruptcy. Moreover, in the work of (Lee et al. 2020) a Multi-Agent Reinforcement Learning-based Portfolio Management System (MAPS) was developed. It used historical daily closing price data of the Russell 3000 index from the US stock market. Experiment results showed that MAPS exceeded all baselines in terms of annualized return and Sharpe ratio.

The research in (Zhang et al. 2021) offers a model based on deep learning for predicting stock price fluctuations. The model's forecast target is the next day's stock close price direction on the Shanghai Stock Exchange (SSE) and the Shenzhen Stock Exchange. The model used a deep belief network (DBN) and a long short-term memory (LSTM) network. The results reveal that the suggested model delivers significant gains in prediction performance. It's also been discovered that some businesses are more predictable than others, implying that the proposed approach can be utilized to build financial portfolios. These machine learning models and deep learning models need data from which they learn because they are based on supervised learning methods. The mentioned techniques usually include the following basic steps: Data collection and preprocessing stages, and then model training. The model is performed after training, prediction, and evaluation (Jain, Gupta, and Moghe2018). Neural Networks Architecture We have used three different deep learning architectures for this work (RNN, LSTM and CNN). This part describes briefly the architectures of the neural networks used in the experiment.

VII. RECURRENT NEURAL NETWORK

Recurrent Neural Network is an ANN class in which contacts neurons form a driven graph, or in simpler terms have a self-loop in the hidden layers. This allows RNNs to know the current state from the previous state of the hidden neurons. RNNs take the data they have recently learned in a period together with the current input instance. To learn sequential knowledge, they utilize the internal state or memory. This helps them to practice several tasks such as recognition of speech, recognition of handwriting, etc. Data can be moving in any direction and RNNs may use their internal memory to handle arbitrary input series. This might make RNN useful for solving more advanced and challenging tasks as well as for having higher computational complexity than feeding up neural networks. Thus, as shown in Figure 1 (Ozbayoglu, Gudelek, and Sezer2020), we can unroll the network. This unrolled network illustrates how we can supply the RNN with a stream of data.

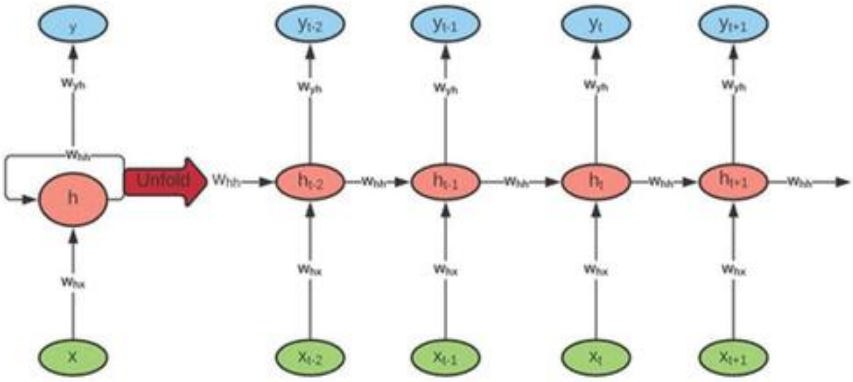


Figure 1. Recurrent neural network cell structure.

VIII. LONG SHORT-TERM MEMORY

LSTM networks are an extension for recurrent neural networks that extend their memory essentially. it's one of the most successful RNN's architectures. Such networks are specifically designed to avoid the issue of long-term dependence but their normal behavior is to retain knowledge for a long-time period back. This is because LSTM's store their information in memory similar to a computer's memory since this network can read, write and remove information from its memory Hiransha et al. (2018). The LSTM model comprises a specific set of memory cells that replace the RNN's hidden layer of neurons and its key is memory cell status. This model extracts information via the gate structure to retain and update the memory cell state. LSTM structure includes three gates: input, forget and output gate. Also, each memory cell has three layers of sigmoid and one layer of tanh. These gates decide whether to let new input into

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

(input gate) or not, remove the information (i.e input) because it is not necessary (forget gate) or allow it to affect the output at the current time stage (output gate). In other words, the gates are used to manage and control the interaction of memory cells among themselves and neighbors. An illustration of the structure of LSTM memory cells with its three gates is shown in Figure 2.

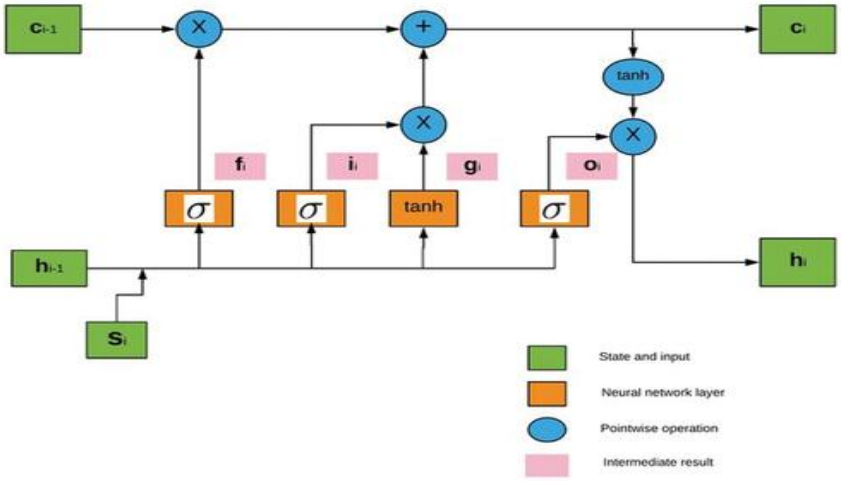


Figure 2. Long short-term memory cell

IX. PROPOSED SYSTEM

In this research we perform both numerical analysis and the textual analysis on the stocks and news dataset to try predicting the future price of the stock. Numerical analysis will be performed by treating the stock trend as a time series and we try to forecast future prices by observing the prices over last x number of days. In textual analysis we perform sentiment analysis of the news articles and learn the influences of news on stock prices. Finally, predictions from these two models will be used as input to a merged model to output final predictions.

9.1 Numerical Analysis

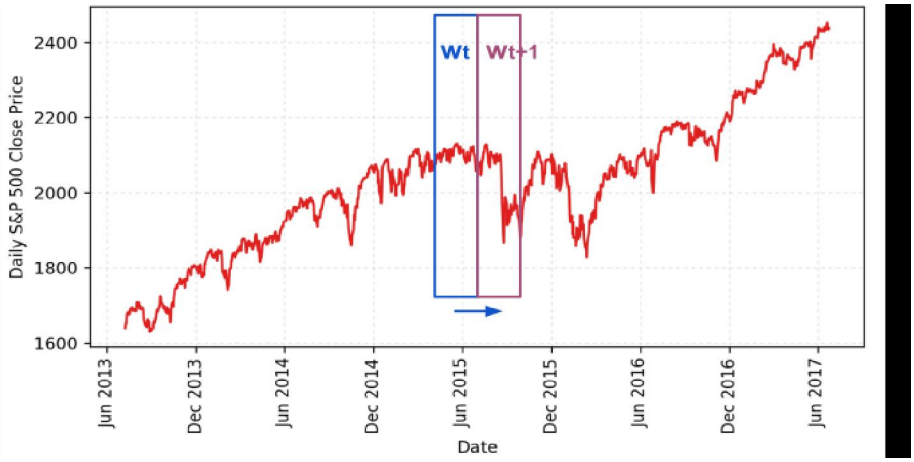


Figure 3. Sliding window learning the trends in stock price

Numerical analysis aims at building a recurrent neural network-based model to predict the stock prices of S&P500 index. RNNs are good at learning and predicting time series data. As stock market data is a time series, RNNs are best suitable for this task. For this purpose, we use a specific type of RNNs called as Long Short-Term Memory. LSTM is specially designed cell which can help the network to memorize long term dependencies. In numerical analysis, we try to learn the patterns and trends of stock prices over the past, and this information will be augmented by the textual information later.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

S&P 500 index data from 3rd January 1950 until 31st December 2017 downloaded from Yahoo! Finance GSPC is used for this purpose. To simplify the problem, we use only the closing prices of the stock index. Stock prices are a time series data of length N. We choose a sliding window w of variable size, which moves step by step from the beginning of the time series. Figure 3 illustrates the sliding window wt which is used as the input to predict wt+1.

While moving the sliding window, we move it to the right by the window size so that there is no overlap between the previous window and the current window. Each input window at each step is passed to a LSTM cell which acts as the hidden layer. This layer would predict values of the next window. While predicting the prices for window Wt+1, we use values from the first sliding window W0 until the current time Wt, where t is the time.

$W0 = (p0,p1,\ldots,pw-1)$

$W1=(pw,pw+1,\ldots,p2w-1)$

$Wt=(ptw,ptw+1,\ldots,p(t+1)w-1)$

The function we are trying to predict is:

$f(W0,W1,\ldots,Wt) \approx Wt+1$

The output window Wt+1:

$Wt+1=(p(t+1)w,p(t+1)w+1,\ldots,p(t+2)w-1)$

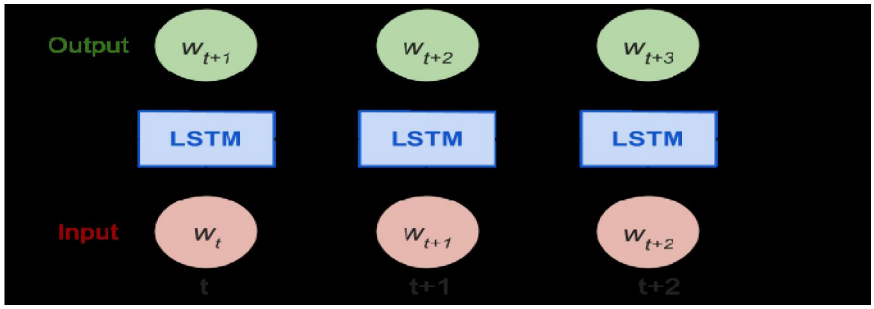


Figure 4. Unrolled version of the network for Numerical Analysis

During the training process, predicted output is calculated by using the randomly assigned weights and compared with the actual value. Error is calculated at the output layer and it is propagated back through the network. This is called back propagation and as this is applied to a timeseries like data as in our case, we call it Back Propagation Through Time (BPTT). During backpropagation, we update weights to minimize the loss in the next step. To update the weights, we calculate the gradient of the weight by multiplying weight's delta and input activations then subtract a ratio of this gradient from the weight. This ratio would influence the quality and speed of the training. This ratio is called Learning Rate. If the learning rate is set high, the network learns fast but the learning will be more accurate when the learning rate is low. Learning process is repeated until the accuracy or loss meets a threshold.

By the design of RNNs, they depend on arbitrarily distant inputs. Due to this, backpropagation to the layers very far away requires heavy computational power and memory. In order to make the learning feasible, the network is unrolled first so that it contains only a fixed number of layers. The unrolled version of the recurrent neural network is illustrated by Figure 4 Unrolled version of the network for Numerical Analysis. The model is then trained on that finite approximation of the network.

During training, we feed inputs of length n at each step and then perform a backward pass. This is done by splitting the sequence of stock prices into non-overlapping small windows. Each window will contain "s" numbers. Each number is one input element and "n" consecutive such elements are grouped to form a training input.

For example, if s=2 and n=3, the training example will be as follows.

Input1 = [[p0, p1], [p2, p3], [p4, p5]], Label1 = [p6, p7]

Input2 = [[p2, p3], [p4, p5], [p6, p7]], Label2 = [p8, p9]

Input3 = [[p4, p5], [p6, p7], [p8, p9]], Label3 = [p10, p11]

Where p is the stock price at a single instance of time.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

9.2 Dropout Parameter:

Deep neural networks with many parameters are powerful systems to learn most of the complex tasks. However, they suffer from the problem of overfitting. The problem is that generally the available data will be limited and in most cases, it will not be sufficient enough to learn complex patterns in the data. As the underlying patterns between inputs and the output can only be mapped using a non-linear approximation function, the hidden layers tend to fit a higher order function to accurately map the outputs. By doing so, they perform best during training but fail miserably during testing. Therefore, we use regularization techniques to avoid overfitting. In general machine learning algorithms, regularization is achieved by using a penalty to the loss function or combining the results from different models and then averaging it. But neural networks are more complex and as they require comparatively higher computational resources, model averaging is not a solution. Therefore, we use a technique called dropout, in which we randomly drop a few units from the network during training, thus preventing co-adaptation between the units.

We use the dropout technique in our model to avoid overfitting. The model will have a configurable number of LSTM layers (n) stacked on top of each other and in each layer, there are a configurable number of LSTM cells (l). Then there is a dropout mask with a configurable percentage (d) of number of cells to be dropped during the dropout operation. We calculate the keep probability (k) from d, which is basically derived from the below equation

$k = 1 – d$

The training process takes place by cycling over the data multiple times. Each full pass over all the data points is called an epoch. In total, training requires m number of epochs over the data. In each epoch, the data is split into b sized groups called mini-batches. In one Back Propagation Through Time learning, we pass one mini batch of data as input to the model for training. After each step, we update the weights at a rate defined by the configurable parameter “l”, which represents learning rate. As explained earlier, higher learning rate lets the network to train faster, but to achieve better accuracy, we need a lower learning rate. In order to find an optimal value, initially we set learning rate to a higher value and at each succeeding epoch it decays by multiplying it with a value defined by “d” (stands for learning rate decay).

9.3 Implementation:

We used tensorflow library to model the network. We first define the placeholders for inputs and targets as follows.

```
inputs = tf.placeholder(tf.float32, [None, n, s])
targets = tf.placeholder(tf.float32, [None, s])
learning_rate = tf.placeholder(tf.float32, None)
```

We create the LSTM cell and wrap it with a dropout wrapper to avoid overfitting of the network to the training dataset.

```
def _create_one_cell():
lstm_cell = tf.contrib.rnn.LSTMCell(l, state_is_tuple=True)
lstm_cell = tf.contrib.rnn.DropoutWrapper(lstm_cell, output_keep_prob=k)
return lstm_cell
```

Number of layers of LSTM cells in the network are configurable. This will be passed as an input to the program. When multiple LSTM cells are used, we try to stack them using the tensorflow’sMultiRNNCell class. This class helps in connecting multiple LSTM cells sequentially into one composite cell.

```
cell = tf.contrib.rnn.MultiRNNCell( [_create_one_cell() for _ in range(l)],
state_is_tuple=True)
```

The Dynamic RNN class present in the TensorFlow library will create the RNN specified by the RNNCell created in previous step.

```
val, _ = tf.nn.dynamic_rnn(cell, inputs, dtype=tf.float32)
```

The last column of the result of above dynamic RNN is then multiplied with weight and bias is added to get the prediction of this step.

```
prediction = tf.matmul(last, weight) + bias
```

Loss is calculated by squaring the difference between the predicted value and the true value, then finding the mean of the result.

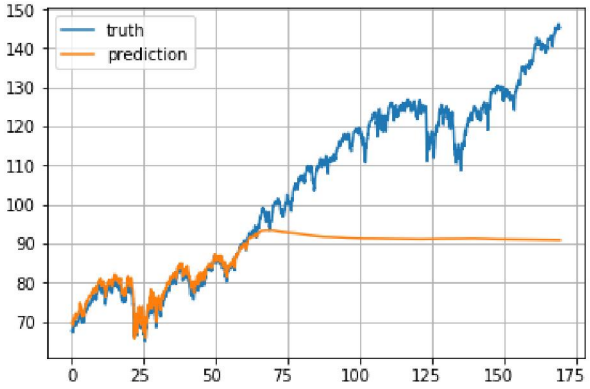
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Where Yi is the vector representing actual values of the stocks and Y^i is the vector representing predicted stock values. We use RMSProp optimization algorithm to perform gradient descent. Learning rate is passed as a parameter to the RMSPropOptimizer and this learning rate will be gradually decayed by the decay parameter in each epoch. After training for the specified number of epoch, the training stops.

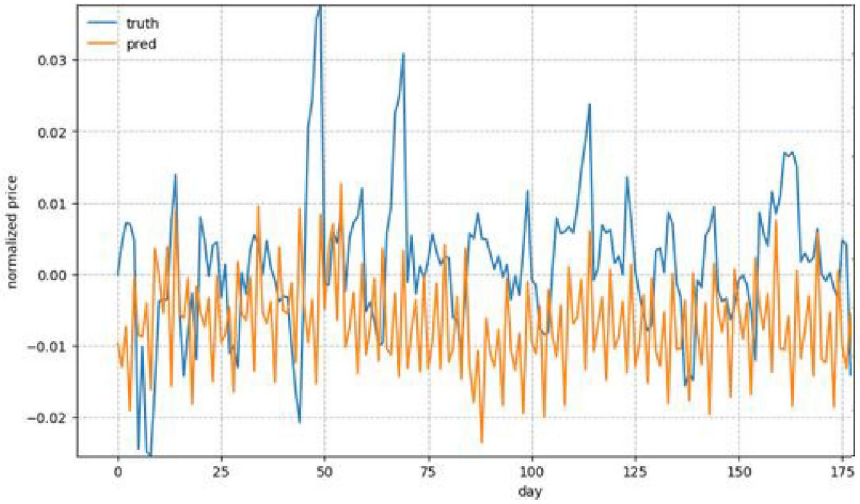
9.4 Normalization:

At each step, latest 10% of the data is used for testing and as the S&P 500 index gradually increases over time, test set would have most of the values larger than the values of training set. Therefore, values which are never seen before are to be predicted by the model. Following figure shows such predictions.



We can see from the above figure that the predictions are worse for new data. Therefore, we normalize the values by the last value in the previous window, thus we will be predicting relative changes in the prices instead of the absolute values. Following is the equation for normalizing the values in a window. W't is the normalized window and t is the time.

$$W'_t = \left(\frac{p_{tw}}{p_{tw-1}}, \frac{p_{tw+1}}{p_{tw-1}}, \ldots, \frac{p_{(t+1)w-1}}{p_{tw-1}}\right)$$



Above figure shows that after normalization, model is able to predict better for new data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

9.5 Textual Analysis

In textual analysis, we analyze the influence of news articles on stock prices. For this purpose, news articles from websites like reddit were downloaded, and sentiment of the headlines was calculated. Then we tried to correlate them with the stock trend to find their influence on the market. Even though sentiment can reveal whether the news is positive or negative, it is hard to know how much it would affect the stock price. Therefore, the problem of predicting the exact stock price was simplified to problem of predicting whether the stock price rises or falls, by transforming the values of close price column to Boolean having values 0 or 1. A value of "1" denotes that the price increased or stayed the same, whereas "0" indicates that the price reduced.

Following are some of the examples of news headlines which influenced the stock prices.

Date	Headline	Influence
2014-03-19	A Surprising Number of Places Have Banned Google Glass In San Francisco	Negative
2014-03-20	Lawsuit Alleges That Google Has Crossed A 'Creepy Line'	Negative
2014-05-02	Google to Acquire Favorite Live Streaming Service Twitch for $1B	Positive
2014-05-21	Google overtakes Apple as Most Valuable Global Brand	Positive
2014-05-22	Google Inc (GOOGL) Plans to Spend $30 Billion On Foreign Acquisitions	Positive
2014-06-19	Heads up: Hangouts is being weird today (other Google services too)	Negative
2014-07-09	Google Co-Founders Talk Artificial Intelligence Just a Matter of Time	Positive

Table 1 Sample news headlines and their influence on stock price

Then this vector was passed to machine learning algorithms to find mapping between sentiment values and output trends.

Output:

Date	Close	Predictions
2018-07-27	268.149994	273.169891
2018-07-30	267.500000	274.561707
2018-07-31	264.100006	276.339691
2018-08-01	265.049988	277.771332
2018-08-02	260.850006	278.965332
...	...	...
2020-03-18	75.500000	119.214340
2020-03-19	72.949997	114.244820
2020-03-20	77.300003	109.748322
2020-03-23	66.199997	106.353691
2020-03-24	68.550003	102.927559

404 rows × 2 columns

TATA MOTORS

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Date	Close	Predictions
2018-07-27	129.750000	134.438599
2018-07-30	132.750000	134.696014
2018-07-31	133.000000	135.391373
2018-08-01	129.600006	136.202576
2018-08-02	128.850006	136.379395
...	...	...
2020-03-18	90.050003	115.103569
2020-03-19	83.550003	110.387978
2020-03-20	90.599998	104.674706
2020-03-23	74.650002	100.369614
2020-03-24	77.150002	94.927986

404 rows × 2 columns

INDIAN HOTEL

GODREJ PROPERTIES

Date	Close	Predictions
2018-07-27	687.099976	716.803650
2018-07-30	714.500000	717.515259
2018-07-31	714.750000	721.163086
2018-08-01	743.250000	726.165283
2018-08-02	725.500000	733.972473
...	...	...
2020-03-18	707.750000	905.930481
2020-03-19	654.950012	877.951904
2020-03-20	699.950012	846.758972
2020-03-23	560.000000	820.281799
2020-03-24	571.099976	786.575684

404 rows × 2 columns

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

X. CONCLUSION AND FUTURE WORKS

In this research, different neural network approaches, namely RNN, LSTM, and CNN, have been applied to the forecasting of stock market price movements. This study discusses the use of neural networks to predict future stock price patterns focused on historical prices. We focused on the importance of choosing the correct input features, along with their preprocessing, for the specific learning models and predicting trend on the basis of data from the past 5 years. For analyzing the efficiency of our models we used four different evaluation metrics. For each model, we measure the error percentage that exists in the training and testing dataset. Then, we compare the obtained results using various sets of features with a specific number of epochs. After we conducted several experiments with different features and epochs, we have found that LSTM is the best model. As we have established through our work that deep learning can stably predict the stock price movement, we think there is more scope to work on making the investment more dynamic and intelligently responsive to the market. We can combine many models for better prediction and efficiency. Besides, the models in this experiment can be updated with other stock indices and they can be optimized using hyper-parameter optimization. In the future, it could be possible to combine multi-agent system and our deep learning methods to enhance predicting the exact price value

REFERENCES

[1] Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. 2014. Stock Price Prediction Using the Arima Model. Paper presented at the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation (UKSim), Cambridge, UK, March 26–28.

[2]Atkins, M. Niranjan, E. Gerding, Financial news predicts stock market volatility better than close price. J. Finance Data Sci. 4(2), 120–137 (2018).

[3]Hernández-Álvarez, Myriam, Edgar A. Torres Hernández, Sang GuunYoo., 2019. Stock Market Data Prediction Using ML Techniques.‘‘ In International Conference on Information Technology & Systems, Springer, Cham, pp

[4]E. F. Fama, The Distribution of the Daily Differences of the Logarithms of Stock Prices, Unpublished Ph.D Dissertation, University of Chicago, 1964.

[5]J. Zupan, Introduction to Artificial Neural Network (ANN) Methods: What They Are and How to Use Them, Acta ChimicaSlov, 41-327, 1994.

[6]Obthong, M., Tantisantiwong, N., Jeamwatthanachai, W. and Wills, G., 2020. A survey on machine learning for stock price prediction: algorithms and techniques

[7] Song, Y. and Lee, J., 2019, December. Design of stock price prediction model with various configurations of input features. In Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing (pp. 1-5)

[8]Sharma, V., Khemnar, R., Kumari, R. and Mohan, B.R., 2019, September. Time series with sentiment analysis for stock price prediction. In 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT) (pp. 178-181).

[9]Tran, D. T., Iosifidis, A., Kanniainen, J., &Gabbouj, M. (2018). Temporal attention-augmented bilinear network for financial time-series data analysis. IEEE transactions on neural networks and learning systems, 30, 1407–1418.

[10]Sharma, A., Tiwari, P., Gupta, A. & Garg, P. (2021). Use of LSTM and ARIMAX Algorithms to Analyze Impact of Sentiment Analysis in Stock Market Prediction. Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020, 377–394.

[11]Sezer, O. B., &Ozbayoglu, A. M. (2018). Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. Applied Soft Computing, 70, 525–538.

[12]Hu, Z., Zhao, Y. and Khushi, M., 2021. A survey of forex and stock price prediction using deep learning. Applied System Innovation, 4(1)

[13]Jain, S., Gupta, R. and Moghe, A.A., 2018, December. Stock price prediction on daily stock data using deep neural networks. In 2018 International Conference on Advanced Computation and Telecommunication (ICACAT)

[14]Pasupulety, U., Anees, A.A., Anmol, S. and Mohan, B.R., 2019, June. Predicting stock prices using ensemble learning and sentiment analysis. In 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) (pp. 215-222)

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

[15]Parray, I.R., Khurana, S.S., Kumar, M. and Altalbe, A.A., 2020. Time-series data analysis of stock price movement using machine learning techniques. Soft Computing, 24(21)