iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6030

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Machine Learning Model for Stock Market Prediction

Prof. Ashwini Kanade[1], Sakshi Singh[2], Shweta Rajoria[3], Pooja Veer[4], Nayan Wandile[5]
[1, 2, 3, 4, 5]*Information Technology, BVCOEW ,Pune , India.*

Abstract: Financial markets are fascinating if you can predict them. Also, the traders acting on financial markets produce a vast amount of information to analyse the consequences of investing according to the current market trends. Stock Market prediction is the technique to determine whether stock value will go up or down as it plays an active role in the financial gain of nation's economic status.

The foundation factor for each investor is to gain maximum profits on their investments. If the company's profits go up, you own some of those profits and if they go down, you lose profits with them. We propose a framework using Long Short Term Memory machine learning algorithm and adaptive stock technical indicators for efficient forecasting by using various parameters obtained from the historical data set considered for a particular company. This algorithm works on historical data retrieved from Yahoo Finance.

For prediction of share price using Long Short Term Memory, there are two modules, one is training session and other is predicting price based on previously trained data. The results will attempt to predict whether a stock price in the future will be higher or lower than it is on a given day to increase transparency among investors in the market. The proposed model also attempts to use sentiment analysis of financial news and opinions fetched from social media platform like Twitter as it acts as another influencing factor in governing stock trends.

Keywords: RNN- Recurrent Neural Network, LSTM-Long Short Term Memory, Machine Learning, Prediction, Sentiment analysis, Stock Market, Technical indicators, Yahoo

I. INTRODUCTION

Stock market is the indispensable part of fast emerging economic countries. The booming volume of data has far surpassed the ability of human beings to analyze them manually. Due to the broad range of financial markets and their desire for investment, so many people, even with minute knowledge, can invest in this domain. Considering the fact of inadequacy of understanding and awareness across the people, stock market prediction techniques plays a pivotal role in bringing more people into market as well as to hold on to the existing investors. Precise prediction of stock prices presents a burdensome task for traders and investors.

A. Stock Market

Stock market, a very uncertain sector of investment, involves a tremendous number of investors, buyers and sellers. A stock generally represents possession on business by a particular individual or a group of people. This market has given investors the chance of earning money and having a substantial life through investing small preliminary amount of money, low risk compared to the risk of opening a new business or the need of profitable career. To make potential money, the rule is simple, you only have to figure out the pattern of stock prices and invest on right time in right place. So, if someone has precise prediction at exact time, he will be earning huge from minimal cash flow available.

B. Literature Survey

The primary focus of our literature survey was to inspect common machine learning algorithms and see if they could be adaptive to our system which works on real time stock price data. These algorithms included Regression, SVM and ARIMA. However, as we were proceeding to our model we strive upon a major drawback of finding long term dependencies between stock prices. A brief search of collective solutions to this problem led us to RNN and LSTM. After deciding to use LSTM to perform stock prediction, we referred a number of papers and concluded our literature survey by considering LSTM as an ideal choice for investigating how fluctuation in one stock price can affect the stock prices over a long period of time. It also helps to determine for how long details about certain past trends in stock price movement needs to be retained in order to predict future trends more accurately in the variation of stock prices.

C. Prediction

The attempt of trying to determine the future value of the stock market is known as Stock Market Prediction. Predicting stock market prices is a complicated task as it includes determination of future value of stock unit traded on an exchange. The successful prediction of a stock's future price could grant significant profit. Due to the significant volume of money involved and number of proceedings that take place every minute, there should be accuracy between the volumes of predictions made. The technical analysis approach recommends generating predictions based on the historical price values of selected stocks. The data set of the stock market prediction model contains details like the closing price, opening price, the data and various other parameters that are needed to predict the target variable which is the price in a given day.

D. Sentiment Analysis

Sentiments are an integral part of stock market and the idea of evaluating these based on several data sources can give observations on how stock markets responds to various kinds of news. Nowadays, social media has become a mirror that reflects people's views and opinions to any specific event or news. Sentiment classification tries to predict sentiment from texts and it has now become the presiding approach used for extracting sentiments and appraisals from online sources. This attempts to divide the language units into three categories: negative, positive and neutral. Any positive or negative sentiment of public related to a particular company can have a diverse effect on its stocks prices. We aim to predict the stock market prices of different organizations by performing sentiment analysis of news as well as social media data such as tweets.

II. METHODOLOGY

Neural network mathematically can be defined as a differential function that maps one kind of variable to another kind of variable. Recurrent neural network is a type of ANN, designed to identify patterns in sequence of data, such as the spoken words, text, handwriting, genomes or numerical time series data originating from stock markets, sensors and government organizations.

A. Recurrent Neural Networks

Recurrent Neural Networks make use of back propagation algorithm, but it is registered for each timestamp which is oftentimes known as Back propagation Through Time (BTT). The issues in RNN are- vanishing gradient and exploding gradient. These issues are point of concern because RNN is trained by back propagation through time and therefore extend into multiple layers. When gradient is returned through multiple time steps, it tends to grow or vanish.

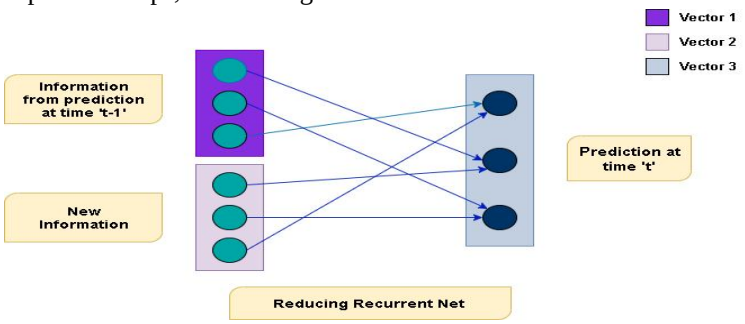


Fig. 1 Recurrent Neural Networks

B. LSTM

Long Short Term Memory Network is a kind of RNN which is capable of learning long-term dependencies. This model is capable enough to decide when to remember and forget and how long to hold onto past details and also make connections between old details with the new input. LSTM has a chain like framework but the repeating modules are different. The key to LSTM is a cell state.

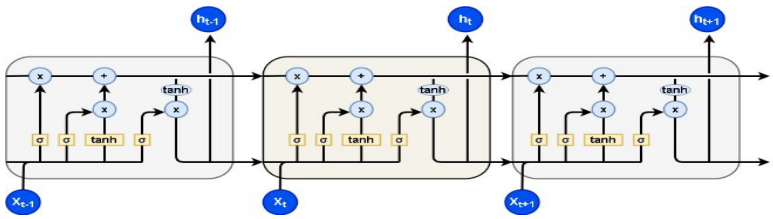


Fig. 2 LSTM

The horizontal line at the top of the diagram is the cell state which runs straight down the entire chain with only some minor linear interaction.

Three gates of LSTM cell:

1) Input gate controls whether the memory cell is updated so it is applied to $\bar{C}$ which is the only vector that can modify the cell state.

$$i_t = \sigma\left(w_i\left[h_{t-1}, x_t\right] + b_i\right)$$

2) Forget gate determines how much of the old state should be unremembered. The state is applied to the output gate to get the hidden vector.

$$f_t = \sigma\left(w_f\left[h_{t-1}, x_t\right] + f\right)$$

3) Output gate- checks if the information of the current cell state is made visible

$$o_t = \sigma\left(w_o\left[h_{t-1}, x_t\right] + b_o\right)$$

C. *Steps*

1) *Step 1:* Identification of information which is of no importance and will be excluded from the cell state and this conclusion is made by a sigmoid function of the forget gate layer.

w_f= Weight

h_{t-1}= Output from previous time stamp

x_t= New Input

b_f= Bias

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

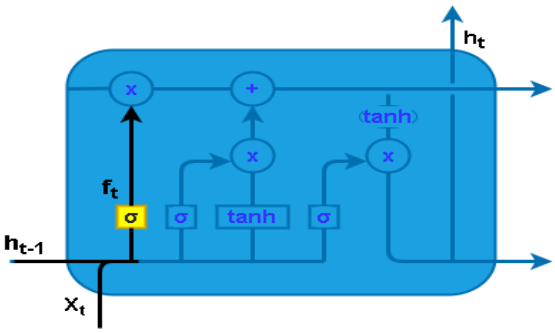


Fig. 3 Identify information using forget gate

2) *Step 2:* Finding what new information should be stored in the cell state. This entire procedure incorporates following steps: a sigmoid function of the "Input Gate layer" which determines the values to be updated next, a tanh function which generates a vector of newest candidate values which could further be concatenated to the state.

$$i_t = \sigma\left(w_i\left[h_{t-1}, x_t\right] + b_i\right)$$

$$\bar{e}_t = tanh\left(w_c\left[h_{t-1}, x_t\right] + b_c\right)$$

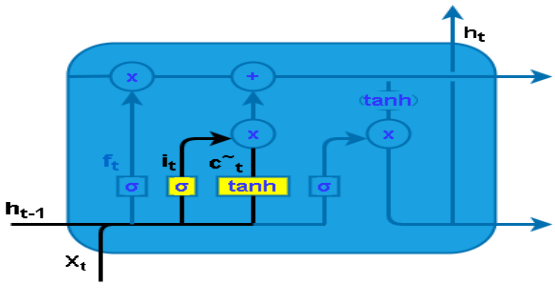


Fig. 4 Decide new information using input gate

In the next step, we will combine these two to update the state.

3) *Step 3:* Update the previous state, C_{t-1} into the new cell state C_t. This is applied by first multiplying the old state C_{t-1} by ft and forgetting the things we decided to neglect earlier and further we add it to $\tilde{C}_t$. The new candidate value obtained by this is scaled by the decided updated value for each state.

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$$

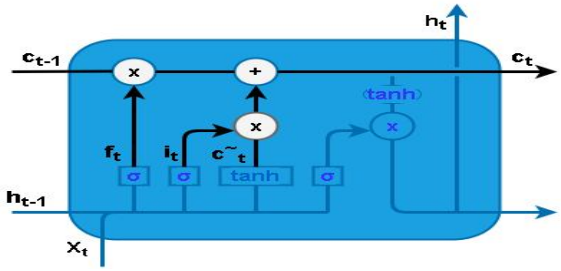


Fig. 5 Update the states

4) *Step 4:* Execute a sigmoid function which identifies which segment of the cell state should be displayed in the output. Then, by putting the cell state through tanh function (push the values to be between -1 and 1) and multiplying it by the output acquired from the sigmoid gate, so only the decided part should get displayed on the output.

$$o_t = \sigma\left(w_o[h_{t-1}, x_t] + b_o\right)$$

$$h_t = o_t * \tanh\left(c_t\right)$$

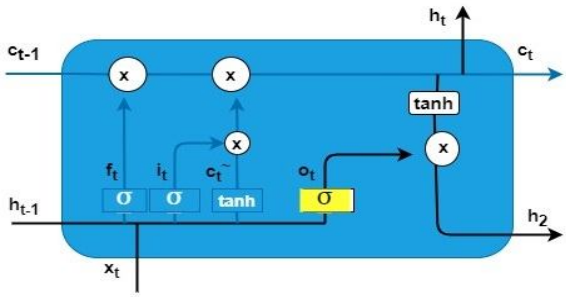


Fig. 6 Identify output

III. PROPOSED SYSTEM

Pandas data reader, which is a sub package that allows one to create a data frame from various internet data sources, is used to fetch the raw data from Yahoo Finance. The start date and the end date is accepted as a parameter based on which the length of the data will depend along with the name of the organization. There are seven columns or attributes that describe the rise and fall in stock prices. Some of these attributes are (1) DATE, which contains all the dates between start date and end date (2) HIGH, which describes the highest value the stock had in a previous year (3) LOW, is quite the contrary to HIGH and resembles the lowest value the stock had in previous year (4) OPEN is the value of the stock at the very beginning of the trading day (5) CLOSE stands for the price at which the stock is valued before the trading day closes (6) VOLUME, tells you how many shares of that particular stock were traded that day (7) ADJ. CLOSE, closing price adjusted for splits and dividend distributions.

Date	High	Low	Open	Close	Volume	Adj Close
2012-01-03	58.928570	58.428570	58.485714	58.747143	75555200.0	50.994907
2012-01-04	59.240002	58.468571	58.571430	59.062859	65005500.0	51.268970
2012-01-05	59.792858	58.952858	59.278572	59.718571	67817400.0	51.838169
2012-01-06	60.392857	59.888573	59.967144	60.342857	79573200.0	52.380054
2012-01-09	61.107143	60.192856	60.785713	60.247143	98506100.0	52.296970

Fig. 7 Dataset

This is a pictorial representation of the data set fetched from Yahoo Finance. The number of record varies as per the parameter passed. These parameters includes the interval between the start date and the end date to be considered for fetching the data set. Also the values of the dataset would differ depending on the company chosen by the user. This dataset is used to train the LSTM model. This dataset is divided into training data and testing data, where training data contains 80% of the dataset and testing data contains 20% of the dataset.

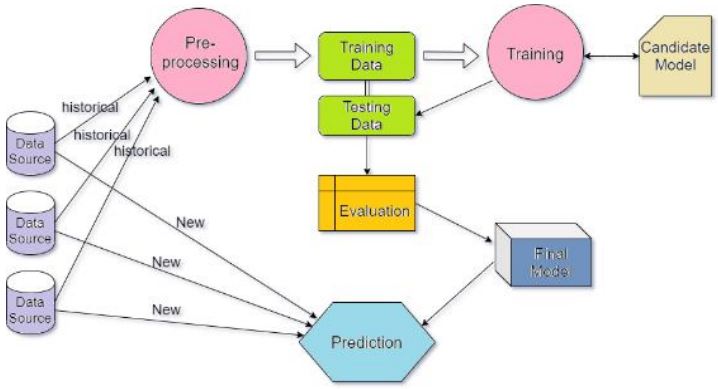


Fig. 8 System Architecture

Now, the first step is to describe our network and construct an instance of the Sequential class. Following this we can create our layers and add them in the sequence in which they should be connected. The LSTM recurrent layer consists of memory units called as LSTM(). A fully connected layer that often goes around LSTM layers is used for outputting a prediction and is called as Dense(). By transforming a 2D dataset to a 3D dataset using the function reshape() in NumPy. Once we describe our network, we must compile it. Compilation process requires a wide range of parameters to be specified, particularly moulded for training our network. Specifically, the optimization algorithm is used to train the network and to evaluate the network that is minimized by the optimization algorithm the loss function is used.

Once the network is compiled, it can be considered as fit, which means it can adapt the weights on a training dataset.

Fitting the network needs two specifications on the training data such as a matrix of input patterns, X, and an array of matching output patterns, y.

The back propagation algorithm network is used for training and optimized according to the optimization algorithm and the loss function specified while compiling the model. Once the network is trained, it can be evaluated for the results. Once the satisfaction with the performance of our fit model is achieved, we can use it to make predictions on new data.

This is as easy as calling the predict() function on the model with an array of new input patterns.

IV. ADDITIONAL FEATURES

A. Twitter Sentiment Analysis

Sentiment Analysis, also known as Opinion Mining, is a supportive device inside natural language processing which permits in recognizing, measuring, and learning subjective data.

Sentiment Analysis is a process of determining ‘computationally’ whether a content is positive, negative or neutral. It also derives the opinion or attitude of an individual. Twitter, as we all know, has been gaining fame nowadays and it is used regularly to convey opinions regarding multiple issues and topics. The workflow looks like:

1) Gather relevant tweets from Twitter: We’ll start by grabbing the tweets we want from Twitter by defining keyword “Yahoo Finance”.
2) Pre-processing (stopword removal): Stopwords are words that aren’t integral to the meaning of a text, and are usually removed as part of a Natural Language Processing workflow. Reorganization of tweets are done into sentences without using stopwords.
3) Apply the right sentiment analysis algorithm: Choosing which sentiment algorithm to use depends on a number of factors: you need to take into account the required level of detail, speed, cost, and accuracy among other things.
4) Analyse the results: The results are analysed.
5) Discuss further improvement and next steps: Gather (a lot) more tweets and explore other algorithms.

```
Positive tweets percentage: 52.77777777777778 %
Negative tweets percentage: 18.055555555555557 %
Neutral tweets percentage: 29.166666666666668 %
```

Fig. 9 Twitter Sentiment Analysis

B. News Sentiment Analysis

Similarly we can extract live news and news analysis can be performed. The workflow looks like:

1) Find some news source: A news source has to be selected and the data is fetched using built-in functions.
2) News extraction: Parse the html and extract the content with BeautifulSoup.
3) Display live news: The live news is displayed to the user.

C. Learning Module

Stock market is not a difficult subject to understand and anyone can learn to trade stocks. So it is important to have a basic understanding of how stock market related activities work. While new investors should educate themselves as to the common mistakes that people make in investing, they should also understand that the market landscape is in a permanent flux. This project also involves a learning module which will help people understand stock trading and predictions in a proper manner.

V. EXPERIMENTAL RESULT

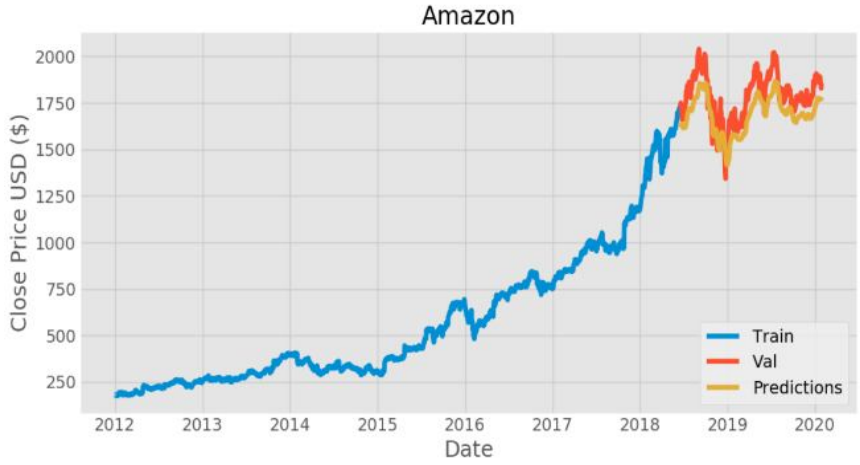


Fig. 10 Prediction for Amazon

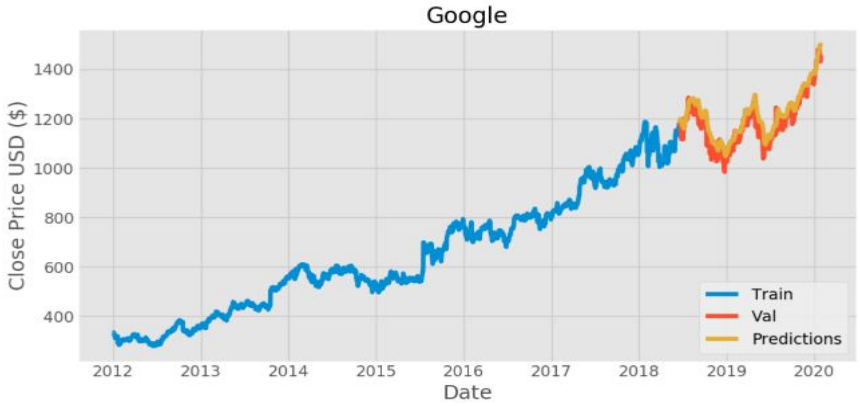


Fig. 11 Prediction for Google

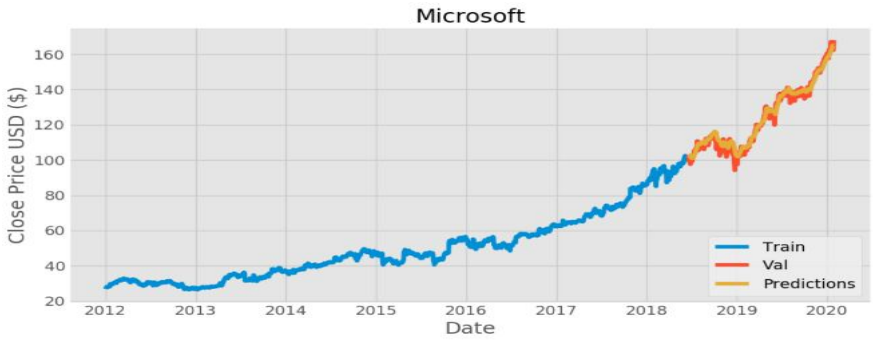


Fig. 12 Prediction for Microsoft

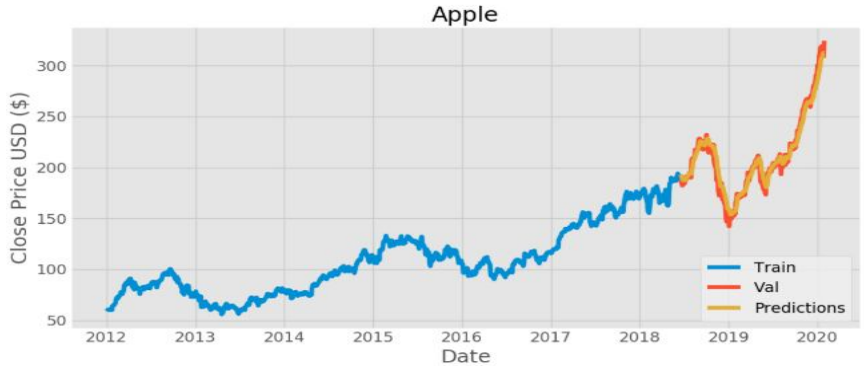


Fig. 13 Prediction for Apple

VI. ADVANTAGES

LSTM is robust in sequence prediction problems because they are capable enough of storing historical information. This is extremely important because for the proposed model, the previous price of a stock is significant in predicting its future price. The other primacy of LSTM is its potential to learn context specific temporal dependence. LSTM contributes in resolving vanishing gradient problem- since the value stored in a memory cell is not iteratively modified, the gradient does not vanish when trained with back propagation. LSTM are not sensitive to time lags between input data points as compared to other RNNs. It works effectively over a wide range of parameters like input gate bias, output gate bias and learning rate.

VII. CONCLUSIONS

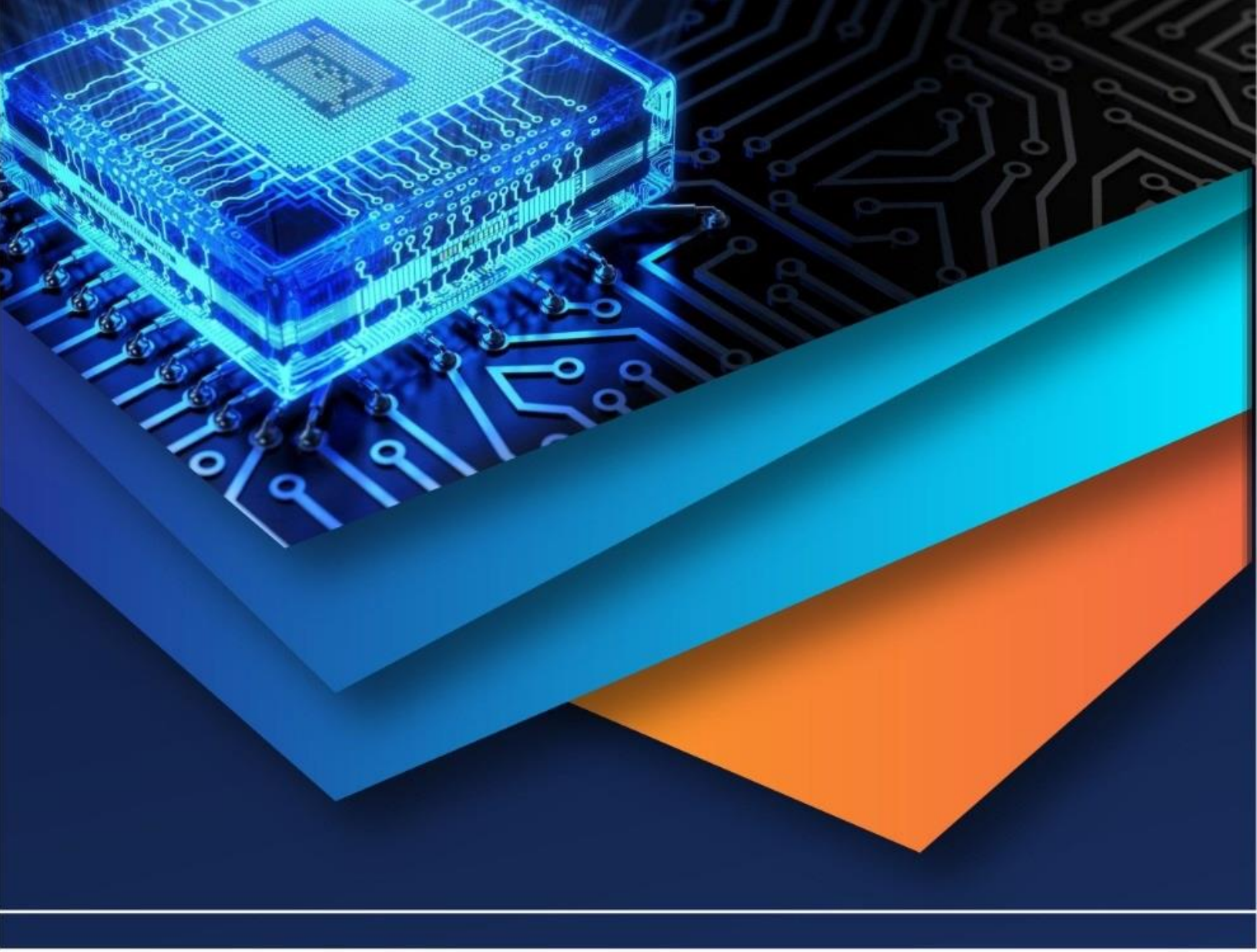
In the paper, we proposed the utilization of the data collected from the financial markets with machine learning algorithms in order to predict the stock price fluctuations. We predicted the stock market movement which takes input as the closing price of stock and news heading. The technique which has been employed in this paper is LSTM which is applied to the Yahoo finance dataset. We also used sentiment analysis to calculate whether the type of article that has a positive or negative influence on the stock and those can be used for further analysis. The acquired results are used to evaluate the prices of stock and the data after calculating is updated and displayed to the user as a graph. This interpretation of results has led to the conclusion that it is possible to predict the stock market movements with more precision using machine learning techniques.

VIII. ACKNOWLEDGMENT

It is a matter of great pleasure to present this paper on " Machine Learning Model For Stock Market Prediction
". I'm grateful to the institute , Bharati Vidyapeeth's College of Engineering for Women for giving this opportunity to develop a web application on this major project .
We would like to express my special thanks of gratitude to my guide Prof. A. V. Kanade who has been an invaluable assistance to this project with her advice and suggestions. We are very thankful to Principle I/C Prof. Dr. S. R. Patil and Prof. Dr. D. A. Godse, Head of Information Technology Department, and other staff members for encouraging us and contributing in stimulating suggestions. They have always been prompt at extending their helping hand and valuable technical known.

REFERENCES

[1] Introduction to Data Mining and Knowledge Discovery (1999), Third Edition ISBN: 1892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854.
[2] Larose, D. T. (2005), “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2, John Wiley & Sons, Inc.
[3] Dunham, M. H. & Sridhar S. (2006), “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition.
[4] Eugene F. Fama “The Behavior of Stock Market Prices”, the Journal of Business, Vol 2, No. 2, pp. 7–26, January 1965.
[5] Robert K. Lai, Chin-Yuan Fan, Wei-Hsiu Huang and Pei-Chann Chang, “Evolving and clustering fuzzy decision tree for financial Time series data forecasting”, An International Journal of Expert Systems with Applications,Vol.36, No.2, pp. 3761-3773, March 2009.
[6] An IEEE paper on “Survey of Stock Market Prediction Using Machine Learning Approach” by Ashish Sharma, Dinesh Bhuriya and Upendra Singh,2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY