Stock Market Prediction Using Deep Learning

Dhanush V Uttarkar[1], Sahana G S[2], Samrudh S Shetty[3], Spoorthi V[4], Shilpa B L[5]
[1234]8th SEM Student, [5]Assistant professor
[12345]Dept. of CSE,
VVIET, MYSURU-570028
KARNATAKA, INDIA

Abstract- Stock market prediction has always been an important priority since its first coming and several people developed several different methods to predict this with a certain amount of accuracy. With the coming of new technological advancements in the fields of machine learning, AI, Deep Learning and NLP there have been improvements in the percentage of accuracy the stock prices can be determined with. The proposed system performs stock market analysis and prediction using LSTM model. The LSTM model used in the proposed system is a special type of recurrent neural network. The proposed system consists of three modules which are linked and made to work together using the LSTM model. This system will use old stock market data as one module which predicts the stock prices using the LSTM model, twitter feed data to performs sentiment analysis which we then use to predict the stock price correlating with stock prices using again the LSTM model this forms the second module. The system will also predict the stock using commodity prices of necessary commodities such as oil and this forms the third module. The LSTM model was chosen for its efficiency in predicting the outcome. For the sentiment analysis part, the system uses the torchtext library from pytroch. The commodity prices, old stock price prediction, stock price prediction using sentiment analysis are correlated using the LSTM model in the pytorch library.

Keywords- LSTM, RNN, Deep Learning, Sentiment analysis, Stock market prediction

I. INTRODUCTION

Stock market is the financial market which acts as a platform for buying selling and exchange of shares. Prediction of stock prices is important since the majority of the capital belonging to the company exists in the form of stocks and predicting their up rise and downfall could affect the company in a major way.

The approaches to predict stock can be broadly classified into two major categories as Fundamental analysis and Technical analysis.

- Fundamental analysis focuses mainly on the credibility and reliability of the company by using the traditional mathematical techniques like P/E ratio.
- Technical analysis is the method of predicting stock price by using technological approaches such as algorithms.

Technical analysis is the currently used approach to predict stock price which includes the usage of machine learning, deep learning techniques, Artificial Intelligence and Natural Language Processing.

Sentiment analysis is the mining of text which identifies and selectively extracts subject information from a certain source. The twitter feed data is used in the proposed system. A positive or negative sentiment related to a particular company can have a ripple effect on its stock prices

Summarization is an important part in different NLP applications like Information Retrieval, Quality Analysis. Summaries can be broadly classified into two categories. One is the Extraction of data where in contents from text like words and the sentences are reused. One more is Abstract that includes regenerating the extracted contents for specific purpose.

The movements in the world's equity markets are controlled by a multiple factors, ranging from large institutional block trades and program trading to earnings and economic reports. However, price of commodities is the one important factor that is overlooked. In fact, frequent change in commodity prices can have a huge impact on the earnings of companies andthe markets. The change in the price of commodities effect the change in stock both directly and indirectly.

This paper is organized as follows: Section II surveys the related work. Section III describes the proposed model which consists of sentiment analysis, historical data and commodity prices. Section IV tells about the experimental analysis. Finally, Section V gives the conclusion of the work.

II. RELATED WORK

In Paper[1] NLP techniques are used to extract subjective expressions. There are 2 tools used for twitter mood effect to predict stock price the first being GPOM (google profile of mood states) and opinion finder. The behavioural graph was calculated based on the effect of social media and on online human behaviour which used the back propagation neural network for analysis and this behaviour graph was further used for sentiment analysis. It also uses decision tree algorithm and support vector machines. The topic sentiment latent Dirichlet allocation model was used to predict stock price movement.

The main model is built around regression learning and is used along the random forest and support vector machine algorithms to form the prediction model.

According to Paper[2] the wavelet transforms are used in some systems, stacked autoencoders, and LSTM. They also describe how this method is newer and better than use of ANN's or SVM's. Conventionally the deep learning approach of choice is that of Convolutional Neural Network (CNN's), but not much was known about other approaches. The author thus uses Stacked Autoencoders not knowing the efficacy of it

initially but also to investigate whether stacked autoencoders perform well for tasks such as these.

The primary model for prediction is the stacked autoencoder, the wavelet transform is used to reduce noise whereas the LSTM is a nonconventional RNN that increases the accuracy of the system.

In certain systems, as in paper [3] the DJIA values and publicly available twitter data is considered.

The DJIA values are fed to the pre-processor to get the processed values.

Meanwhile twitter data is used in sentiment analysis using a methodology which consists of following steps:

1.word list generation –which uses POMS (profile of mood states) technique and Sentiword and standard Thesaurus to get the list of words

2.Twitter filtering

3.Data Score computation

4.Score mapping

And then the data is cross validated using Granger causality analysis.

In model learning and prediction, the learning and studying of actual correlation is done using linear regression, logistic regression, SVM and SOFNN in which SOFNN is best among other algorithms which give 75.56% accuracy[3].

To measure accuracy, k fold sequential cross validation(k-SCV) technique is used.

In Portfolio management, the predicted values are used to make intelligent buy /sell decisions based on the stock values of previous days and the standard deviations.

Jiahong Liet. al.[4]says that, the modern methods are used although the prediction was done by Efficient Market Hypothesis in the past days. It is also mentioned in it that, if it is our goal to study how investor sentiment influences the stock market, we need early assessments of the public mood that are both reliable and scalable at a time-scale and resolution appropriate for practical stock market prediction and sentiment analysis is used here. Here the methodology goes in 3 phases first phase classifies the input data into 3 categories positive, negative and neutral; in second phase there is investors sentiment index constructed to measure the daily mood of the stock market; in the third phase the LSTM deployment is done and the prediction is made. In the final phase the deep neural network model is created that consists of the LSTM, a merge layer, a ReLU linear layer and a SoftMax layer. The prediction accuracy for this particular model was 87.86% on 90% of the training data[4].

In Paper[5] the Data mining is used in conjunction with Machine Learning and Artificial intelligence in some systems. The major concentration is the social media feedback. Social media acts as the mirror of people's thoughts and opinion and hence it plays a major role on the stocks of a particular company. The twitter API is used here and first the tweets are collected and the sentiment analysis is done on that data, the stocks are also analysed by past stock data and suitable machine learning algorithm is used to justify the correlation between the tweet and the stock values. The methodology here includes several modules like Data Collection where in the collection of social media data takes place, Feature extraction module where the classifier is built and trained for sentiment analysis; there are two classifiers used namely Naïve Bayes and the Support Vector Machine. Next is the Training module here the data generated in the previous steps is used as the training dataset to train the sentiment analysis model, finally it is the prediction module which predicts the stock values.

Paper[6] says, unlike traditional RNNs[8]which are not effective for long term dependencies LSTMs can be used for stock price prediction. The attention mechanism allows the model to learn algorithm between different modalities. This model is built using Google's tensorflow.

The model takes dataset from TWSE and calculate the technical indicators. The data is sent to deep learning model which consists of attention based LSTM[9]. These LSTMs have cell states which process the data sequentially and remembers and forgets the information. The information in cell state can be manipulated by input, output and forget gates. It takes the sequence of stock data which include price and technical indicators as input,and produces a multiclass output, where each class represents increase or decrease.

Then this output is sent to trading model. If the predicted class belongs to increasing class, strategy is to buy the stock; if it belongs to decreasing class, strategy is to sell.

III. PROPOSED MODEL

Theproposed system is an integrated model which combines 3 different modules, namely the sentimental analysis[7], commodity prices and the historical data. The architecture of the proposed system is as shown in the Fig1. Each module is explained in detail below.

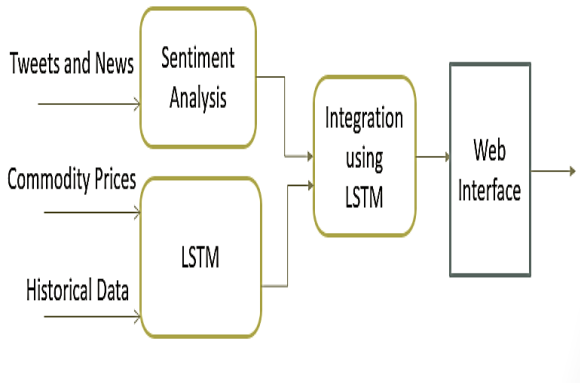

Fig.1: The architecture of the proposed system

I.Sentiment Analysis:

- Obtaining of raw data from twitter API from twitter feeds
- Preparing of data using word embeddings
- The data is fed to the LSTM model
- The LSTM consists of the multi-layered and bidirectional RNNs
- Model training

- Results will be a real number; close to 1 is positive and close to 0 is negative

The word embeddings convert the hot vectors(Sparse vectors) into smaller vectors, the converted vectors will be the real numbers(Dense vector).

RNN:

A. *Bidirectional RNN*:

The bidirectional RNN looks as shown in Fig.2

- It is the recursive neural network that processes the data both forward and backward
- Connects two oppositely directed hidden layers to the same output
- The hidden state tensors returned by the backward and forward RNNs are stacked on top of each other
- Last hidden state of Forward RNN –
  Last hidden state of Backward RNN –

The sentiment prediction is calculated by

$$\hat{y} = f(h_T^{\rightarrow}, h_T^{\leftarrow})$$

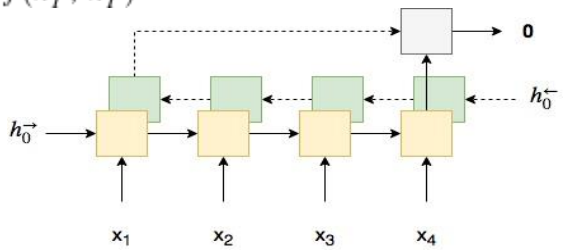

Fig.2: Bidirectional RNN

B. *Multi layered RNN*:

The bidirectional RNN looks as shown in Fig.3

- Additional RNNs are added on top of initial RNN, each RNNs makes up a layer
- The hidden state output by the first RNN at time step 't' will be the input to RNN above it at time step 't'
- Prediction is made from the final hidden state of the final layer

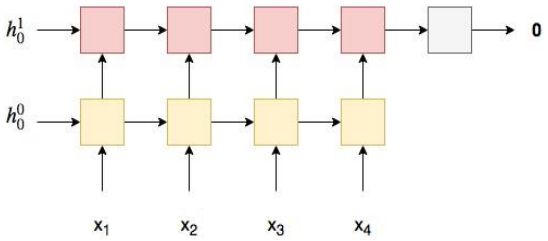

Fig.3: Multi-layered RNN

LSTM:

LSTM[10] is a special type of Recurrent Neural Network, which is capable of learning long term dependencies. It is the most powerful model for time series data, since there can be lag of unknown duration between important events in a time series. Unlike traditional RNNs, LSTM solves the vanishing gradient problem.

LSTM consists of cell, which is the memory part of LSTM and regulators commonly called as gates. Generally, there are three gates in LSTM. They are: input gate, output gate and forget gate. The amount of information that flows into the cell is regulated by the input gate. The retention of the values that should remain in the cell is controlled by the forget gate.

There are two activation functions being used, namely tanh function and the sigmoid function.

tanh function regulates the value that is flowing through the LSTM network and gives values between -1 and +1 whereas the sigmoid function gives the value between 0 and 1.

LSTM improves the longterm dependency of traditional RNN and effectively improves the accuracy and stability of the prediction.

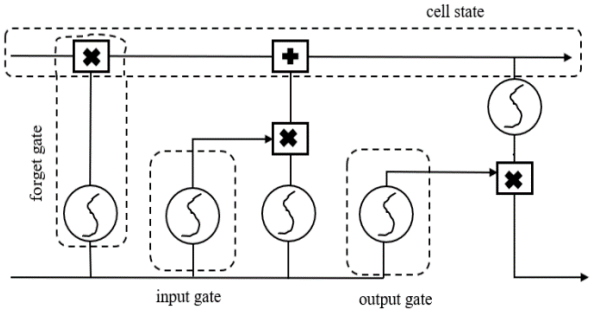

Fig.4: LSTM

II.Historical data and Commodity prices:

Historical data is an important factor which is used to forecast the future performance of the company. It is the collected data aboutthe past events which includes the company's opening price, closing price of a stock in a day. These data are fed into the LSTM[10] model for trend analysis. LSTM takes this time series data to get the future prediction.

The commodity market is the virtual or the physical market to buy, sell or trade products. The rise and fall in the price of these products will have impact on the stock prices. LSTM is used to work on these data and to predict the stock price.

IV. EXPERIMENTAL ANALYSIS

I Dataset

date	symbol	open	close	low	high
########	WLTW	123.43	125.84	122.31	126.25
########	WLTW	125.24	119.98	119.94	125.54
########	WLTW	116.38	114.95	114.93	119.74
########	WLTW	115.48	116.62	113.5	117.44
########	WLTW	117.01	114.97	114.09	117.33
########	WLTW	115.51	115.55	114.5	116.06
########	WLTW	116.46	112.85	112.59	117.07
########	WLTW	113.51	114.38	110.05	115.03
########	WLTW	113.33	112.53	111.92	114.88
########	WLTW	113.66	110.38	109.87	115.87
########	WLTW	109.06	109.3	108.32	111.6

Fig.5: Dataset

The data set was taken from Kaggle[12], the attributes considered were the symbols, open, close. Symbol is particular for a company, open is the stock price of the particular company at the beginning of the day and the close is the stock price of the particular company by the end of the day.

The accuracy of the system can be calculated manually by the equation

$$Accuracy=\frac{□□+□□}{□□+□□+□□+□□} \quad [11]$$

V. CONCLUSION

The stock market prediction is a tedious, complicated and difficult job to carry out. The price of the stock fluctuates due to many affecting factors. In this paper the model uses deep learning techniques to predict the stock prices. The model is fed with the data like the historical data, commodity prices and the twitter data, these data will be processed and the model predicts the stock price; unlike other models, this model integrates three different modules to predict stock prices, the three modules include the historical data, commodity prices and the twitter data, it's a hybrid system that uses the LSTM for prediction purposes. The result of this model states that the deep learning techniques are helpful to predict stock prices with certain amount of accuracy.

VI. REFERENCES

[1]. Salam Al-Augby , Noor Al-musawi and Alaa Abdul Hussein Mezher;"Stock market prediction using sentimental analysis based on social network", 2018.

[2]. Wei Bao, Yulai Rao;"A deep learning framework for financial time series using stacked autoencoders and long-short term memory", 2017.

[3]. Venkata SasankPagolu , Kamal Nayan Reddy , Ganapati Panda , Babita Majhi Anshul Mittal Stanford University, "Stock Prediction Using Twitter Sentiment Analysis",2017.

[4]. Jiahong Li, Hui Bu, Junjie Wu;"Sentiment-Aware Stock Market Prediction: A Deep Learning Method",2017.

[5]. TejasMankar, Tushar Hotchandani, Manish Madhwani, AkshayChidrawar, Lifna C;"Stock Market Prediction based on Social Sentiments using Machine Learning", 2018.

[6]. Li-Chen Cheng, Yu-Hsiang Huang, Mu-En Wu; "Applied attention-based LSTM neural networks in stock prediction"

[7]. AdyanMarendraRamadhani, Hong Soon Goo; "Twitter Sentiment Analysis using Deep Learning Methods", 2017

[8]. Zachary C. Lipton, John Berkowitz;"A Critical Review of Recurrent Neural Networks for Sequence Learning", 2015.

[9]. Mike Schuster and Kuldip K. Paliwal; "Bidirectional Recurrent Neural Networks", 1997.

[10]. Jae Young Choi and Bumshik Lee; "Combining LSTM Network Ensemble via Adaptive Weighting for Improved Time Series Forecasting", 2018.

[11]. Jordan, M. I.; "Serial order: a parallel distributed processing approach. Advances in psychology.", 1997.

[12]. www.kaggle.datasets

Dhanush V Uttarkar,
He is pursuing final year of Engineering in CSE at VVIET Mysuru. His area of interest lies in Machine Learning, Database.

Sahana G S,
She is pursuing final year of Engineering in CSE at VVIET Mysuru. Her area of interest lies in Machine Learning, Database

Samrudh S Shetty,
He is pursuing final year of Engineering in CSE at VVIET Mysuru. His area of interest lies in Machine Learning, Database

Spoorthi V,
He is pursuing final year of Engineering in CSE at VVIET Mysuru. Her area of interest lies in Machine Learning, Database. She has served has Google Developer Student Club Lead for the year 2018-19.

Mrs. Shilpa B L,
Pursed Bachelor of Engineering from Visvesvaraya Technological University, India in 2009 and Master of Technology from Visvesvaraya Technological University, India in year 2011. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and Engineering, VVIET, Mysuru, Affiliated to Visvesvaraya Technological University, India. She is a life member of the IAENG since 2013, ICSES since 2017. Her main research work focuses on Predictive Analytics, Big Data analytics, Natural Language Processing, Data mining and Machine Learning. She has 7 years of teaching experience and 2 years of Industry Experience.