Stock Market Prediction Using Sentimental Analysis in a LSTM Model.

Maithreyan K , Mounika V,

Msc Software Systems

Coimbatore Institute of Technology

suryamaithreya@gmail.com
mounikavenkatesan1998@gmail.com

1 Abstract

Machine learning and deep learning are growing at a rapid speed and it is transforming almost every industry. Stock market is also one of them ,Predicting Stock price is considered to be a great challenge ,but with the power of machine learning applied on various data points, the stock price can be predicted. The price of stock depend on various factors that what makes the prediction challenging. We consider the people sentiment about the company from twitter and other social media. We also take input about the various news about the company. Previous price of the stock is also taken as feature .Stock market price prediction is a time series problem so we are using a recurrent neural network a LSTM model. Price of a stock depends a lot upon its price of stock on previous timestamp, so using a LSTM model increase the accuracy. The model combines both people sentiment and stock market trend and provides valuable insight about the stock price. Thus helping us to make decision to buy and sell the right stocks at the right time.

Keywords: Stock price prediction, Machine learning, Sentimental Analysis, LSTM.

2 Introduction

Stock market prediction has been an active area of research for many years. Early prediction of stock market was based on random walk theory and Efficient Market Hypothesis[1].These models were not that good. The stock market was difficult to predict because its value depended upon various factors. Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making.This show public sentiment about a company can determine a lot about a companies stock price.This papers uses to main factors to predict the stock prices.

- Public sentiment on the company
- Stock price history

Sentimental analysis is performed on publicly available Twitter data to find the public mood and they can be classified into different categories - Calm, Happy, Alert and Kind. Although each so-called tweet, i.e. an individual user post, is limited to only 140 characters, the aggregate of millions of tweets submitted to Twitter at any given time may provide an accurate representation of public mood and sentiment.We use twitter api tweepy to collects tweets about a particular company.

Stock data and prices are a form of time series data. A time series data can be defined as a chronological sequence of observations for a selected variable in our case the stock price.In time series data the value to be predict depends a lot upon the value of the variablein previous time stamp. Analysis of time series data helps in identifying patterns, trends and periods or cycles existing in the data.

To basic classes of algorthims are used for forecasting time series data.

- Linear models
- Non-linear model

The different linear models are AR,ARMA,ARIMA and its variations[2].

The non-linear model are ARCH,GARCH ,Deep learning algorithms [3].

Deep neural networks can be considered as non-linear function approximators which are capable of mapping non-linear functions . Recurrent neural networks have a cycle which feeds activations from the previous time step back in as an input and influences the activations of the current time step. Therefore the activations create an internal state. This in theory can store temporal information for a dynamic indefinite number of time steps in contrast to the fixed number of time steps of feed forward networks. LSTMs are a specific type of recurrent neural network which overcomes some of the problems of recurrent networks[4].Deep learning has the capability of find the hidden patterns and underlying dynamic in the data.

3. DATASET

In this project, we used two main datasets-

•NIFTY 50 companies values from June 2016 to June 2018. The data was obtained using Yahoo! Finance and includes the open, close, high and low values for a given day.

• Publicly available Twitter data about the companies. The data includes the timestamp. Since we perform our prediction and analysis on a daily basis, we split the tweets by days using the timestamp information. The no of tweets about a particular company in a day also calculated.

Data Preprocessing

The data obtained from the above resources are being preprocessed for reliable data analysis.

4. SENTIMENT ANALYSIS

Sentimental analysis is a key part of the model,the sentiment module determines the output that is predicted. A lot of research work is been going on the field of sentimental analysis, and the accuracy has been improved.The tweet about the company is gathered using the twitter api,and the data is sent to be classified. Multiclass classification can be on the text In this project, we use four mood classes namely Calm, Happy, Alert, and Kind. We tried several standard tools like OpinionFinder, SentiWordnet [5] etc.

5 Recurrent neural network

In recurrent neural network the hidden state of the previous timestamp is also feed into the next timestamp. It helps in learning a sequential data well. Each of the computing unit in an RNN has a time varying real valued activation and modifiable weight.

Problems in rnn

1. Exploding gradient problem.
2. Vanishing gradient problem.

One can avoid exploding gradient problem by

- Clipping gradients by threshold
- Truncated BBTT
- RMSprop to adjust learning rate

To avoid Vanishing gradient problem

- LSTM Network

6 Long Short Term Memory

LSTM hidden layers are made up of cells with sigmoidal input, output and forget gates. This allows the network to learn when to forget, take input and output. The LSTM cell has an internal state which is updated based on the previous activations of the layer and inputs through connections to the previous layer and self connections. LSTM is a special kind of RNN, introduced in 1997 by Hochreiter and Schmidhuber [6] . In the case of LSTM architecture, the usual hidden layers are replaced with LSTM cells. The cells are composed of various gates that can control the input flow. An LSTM cell consists of input gate, cell state, forget gate, and output gate. It also consists of sigmoid layer, tanh layer and point wise multiplication operation.The various gates and their functions are as follows

- Input gate : Input gate consists of the input.
- Cell State : Runs through the entire network and has the ability to add or remove information with the help of gates.
- Forget gate layer: Decides the fraction of the information to be allowed.
- Output gate : It consists of the output generated by the LSTM.
- Sigmoid layer generates numbers between zero and one, describing how much of each component should be let through.
- Tanh layer generates a new vector, which will be added to the state.

$$h^t = f(h^{t-1}, x^t; \theta) \qquad (1)$$

The cell state is updated based on the outputs form thegates. Mathematically we can represent it using the following equations.

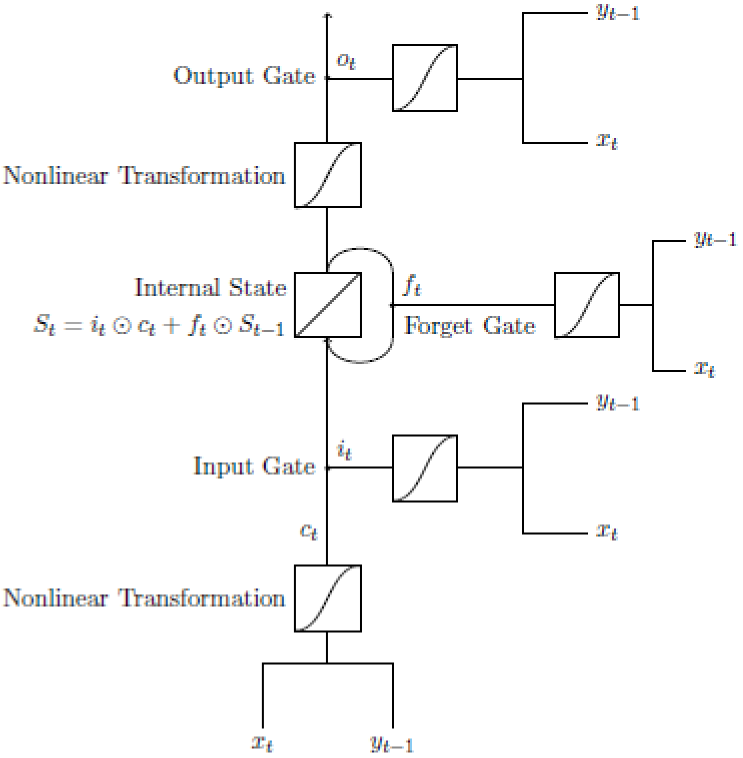
$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{2}$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{3}$$

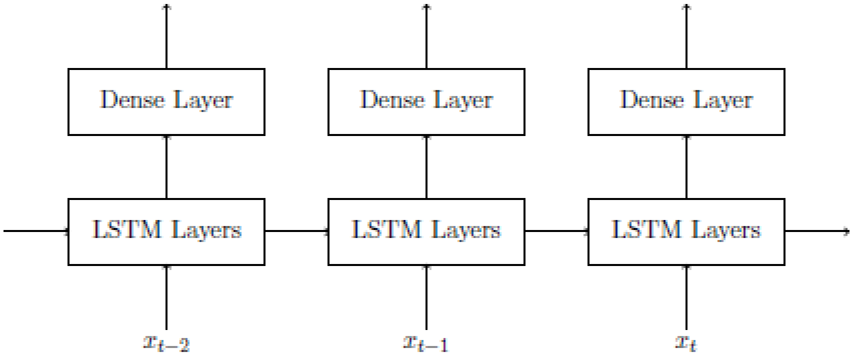
$$c_t = \tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{4}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

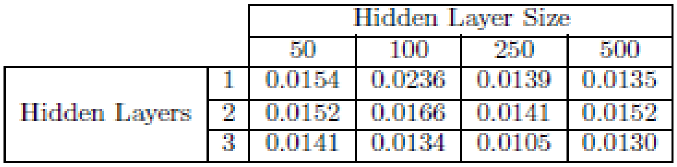


Network structure



RESULT

By varying the number of hidden layer and hidden layer size in LSTM



		Hidden Layer Size			
		50	100	250	500
Hidden Layers	1	0.0154	0.0236	0.0139	0.0135
	2	0.0152	0.0166	0.0141	0.0152
	3	0.0141	0.0134	0.0105	0.0130

Table 1: Returns RMSE Of Specified Networks

Conclusion

In our paper we are using the public sentiment score of a company, and combining the data with the stock market data in a LSTM model. The combination of LSTM model and sentimental analysis helps in increasing the accuracy.In our model we are only using twitter data for analysis, we can also acquire data from various other social media to get better insight of the stock price.

REFERENCE

[1] - Fama, E. F. (1965) The Journal of Business 38, 34–105.

[2] - G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time series analysis: forecasting and control. John Wiley & Sons, 2015.

[3] - G. Batres-Estrada, “Deep learning for multivariate financial time series”. Technical Report, Stockholm, May 2015.

[4] - Felix Gers (2001). Long Short-Term Memory in Recurrent Neural Net- Works

[5] - A. E. Stefano Baccianella and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource forsentiment analysis and opinion mining. In LREC.

[6] - S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neuralcomputation, vol. 9, no. 8, pp. 1735–1780, 1997