Stock Price Prediction using LSTM-ARIMA Hybrid Neural Network Model with Sentiment Analysis of News Headlines

Mahim Dashora
dept. of Computer Science and Engineering
PES University
Bengaluru,India
contact.mahim@gmail.com

Nityam Churamani
dept. of Computer Science and Engineering
PES University
Bengaluru, India
nityam.churamani@gmail.com

Darshil.V Shah
dept. of Computer Science and Engineering
PES University
Bengaluru, India
darshil.vs23@gmail.com

Professor Badri Prasad
dept. of Computer Science and Engineering
PES University
Bengaluru, India
badriprasad@pes.edu.

***Abstract*—Financial markets are extremely volatile, which ends in people losing their money within the exchange. Our project- Stock Price Prediction using LSTM-ARIMA Hybrid Neural Network Model with Sentiment Analysis of News Headlines, is one amongst the many approaches to unravel the matter and predict accurate stock prices. This paper uses a unique method to predict next day's final stock prices using a combination of Long Short Term Memory (LSTM), Auto Regressive Integrated Moving Average (ARIMA) statistical model and Sentiment analysis. This project will mainly provide an insight to traders and investors about future stock prices, thus helping them make the right decisions. They're going to thus be able to minimize the loss of their money and resources.**

1. Introduction

Stock prediction is of extreme importance to banking, investment and stock exchange firms. A huge amount of stock price data is thus used for the prediction of stock prices. Stock values, however, are affected by a lot of factors like news, opening price, closing price, etc.. The prediction of stock prices in this model involves the use of ARIMA and LSTM along with sentimental analysis. The combination of ARIMA and LSTM is achieved by using the concept of weighted average. This way, we are able to combine the two methods along with sentiment of the stock to make the model as apt to the real world as possible.

2. Literature Survey

We referred a number of papers to develop our project. Below, is a small survey about the papers we referred and the implementation techniques present in them.

[1] used LSTM-Model along with sentimental analysis of stock news. This model was very accurate as it also took sentiment of the stock into account. We have to take into account the news of a given timeline alone and going beyond that timeline wouldn't produce results.

[2] uses regression along with LSTM to predict the values of stock. This method doesn't take sentiment of the stock into account and therefore isn't as effective as the ones with sentimental analysis. However, this model is relatively simple to understand and implement.

[8] uses CNN model combined with LSTM to predict stock prices. CNN is used to extract features in data and summarize them. It also helps

with highlighting the most prominent features. LSTM helps with remembering data values. However, this model doesn't take sentiment of the stock into

account.

[6] indicated the use of a hybrid ARIMA - LSTM model. This was an innovative approach and we added to it by taking sentiment of the stock into consideration.

3. METHODOLOGY

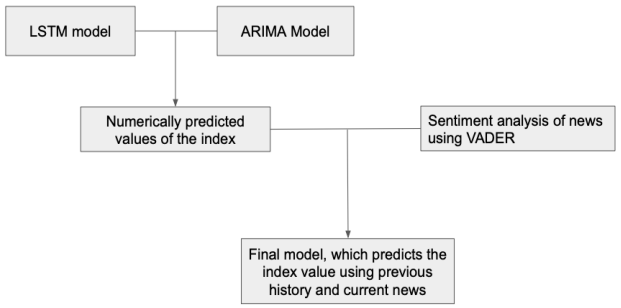


Figure 3.0.1 System Architecture displaying methodology to get final Output

The methodology behind this prediction model involves the use of Long Short-Term Memory[LSTM] and Auto Regressive Integrated Moving Average[ARIMA]. The final prediction then takes the sentiment of the stock into account. Initially, LSTM and ARIMA models are implemented. Data is fed to these models and then training and testing is performed. Once the RMSE[Random Mean Squared Error] of the two models is obtained, a weighted average of those errors is taken and values are predicted. This allows us to combine the above models with minimal error. The final values are then combined with the sentiment score to make the final prediction. Sentiment takes into account the news about the stock. This tends to affect the stock prices. Thus this final combination is used to predict the stock prices.

3.1 Data Preprocessing

Trend varies over time,to eliminate trend we applied transformation to penalize higher values than smaller values , hence we took **log** transformation. News headline dataset and stock price dataset for DJIA index was sourced from [7].The features for stock price data are changed by scaling them between 0 and 1 using MinMax scaling. The dataset was split into 75% Training Set and 25% Test set.

3.2 Long Short-Term Memory (LSTM) Networks

Recurrent Neural Network (RNN) remembers input with the help of their internal memory. Long Short-Term Memory (LSTM) Cells are based on RNN. LSTMs consist of an extra cell state that performs as a memory unit. The decision to update the cell state and the content to be updated is decided by the gates present in the LSTM cell.

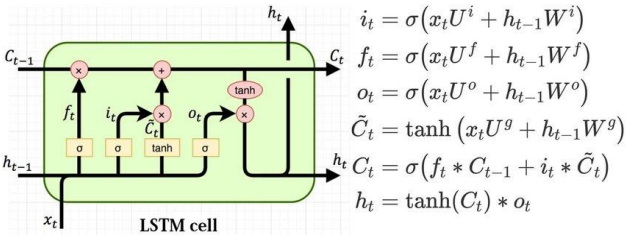


Figure 3.1.1 Long Short-Term Memory [LSTM] cell and equations of its various gates. (taken from https://web.stanford.edu/class/cs379c/archive/2018/class_messages_listing/content/Artificial_Neural_Network_Technology_Tutorials/OlahLSTM-NEURAL-NETWORK-TUTORIAL-15.pdf)

'it', 'ft' and 'ot' represent input, forget and output gate respectively. 'ct' represents the cell state, 'ht' represents hypothesis and 'σ' represents the activation function respectively. The loss function used here is the mean squared error(MSE) with 100 epochs.

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

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

Our LSTM Network has 2 cells and undergoes 100 epochs.

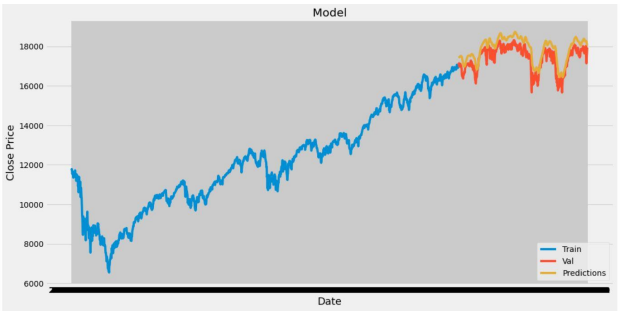


Figure 3.1.1 LSTM model output

3.3 Auto Regressive Integrated Moving Average (ARIMA)

This paper uses a time series model called Auto Regressive Integrated Moving Average (ARIMA), which are applied in cases where data shows non-stationarity with respect to mean . Here, nonstationary time series mean those series where mean and variance are not constant with respect to time period, The system uses ARIMA to predict future price of DJIA (Dow Jones Industrial Average)index.Here, index shows rising trend with respect to time , for the 8 years (2008-2016). Thereby, we difference it to make the series stationary with time, a differenced series calculated by equation.

$$X_{td} = X_t - X_{t-1}$$

ARIMA takes three parameters as input, those are (p,d,q)

- p is the number of auto regressive factors
- d represents times the series is differenced
- q represents number of lagged forecast error

We have used the Dickey Fuller Test to test the stationarity of time series,We also compute the AutoCorrelation Function (A.C.F)and Partial Autocorrelation Function (P.A.C.F)of the time.

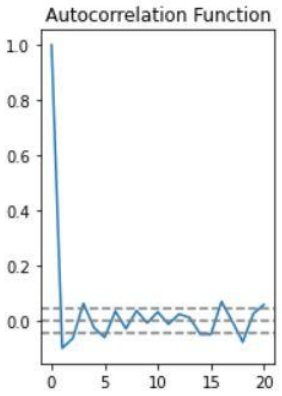


Figure 3.3.1 ACF plot

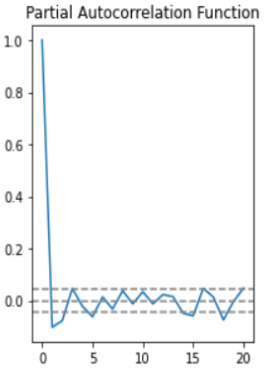


Figure 3.3.2 PACF plot

PACF represents the correlation between current value and lagged value ,but removes indirect effects of prior lags.
ACF represents the correlation between current value and lagged value of variable. We can observe from the following graphs, that p=1 and q=1, from the above ACF and PACF graphs respectively.

This model produces a forecast on 2 years test data given below.

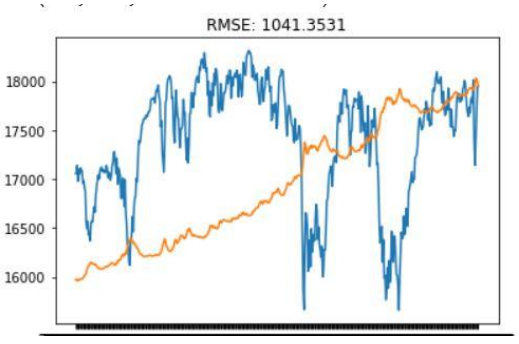


Figure. 3.3.3 ARIMA output

3.4 Sentiment Analysis

The sentiment analysis of the stock headlines is done using the VADER [Valence Aware Dictionary and Sentiment Reasoner] model. It's a part of the NLTK library of python which provides users with the required tools to perform NLP(Natural Language Processing) related tasks. The VADER model generates polarity scores when it processes a text, which gives an indication about the sentiment of the stock. These scores lie between -1 and 1. A score of less than 0 suggests a negative sentiment and a score greater than 0 would suggest a positive sentiment. The VADER model not only predicts the sentiment of the stock but also the extent of the sentiment, i.e, how good or how bad the sentiment is. This method provides a way to numerically represent the sentiment of the stock and thereby improve our prediction.

$$X_f = X_o + \text{compound\_score} * \sigma_\mu$$

$$X_o = \frac{\alpha}{\lambda} * \text{ARIMA\_prediction\_value} + \frac{\beta}{\lambda} * \text{LSTM\_prediction\_value}$$

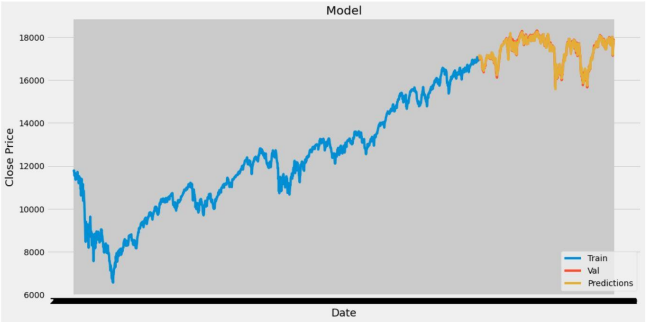
where ,

α=RMSE from the LSTM model

β=RMSE from ARIMA model

λ=(RMSE from LSTM model +RMSE from ARIMA model)

Here ,Xf is the final prediction and Xo represents the predicted values from the ARIMA-LSTM hybrid model ,σ_d is the standard deviation from the last 5 days' simple moving average .
Below , the predicted closing price of DJIA index from dates 14-07-2014 to 30-06-2016.



4. CONCLUSION

The model gives 1.5% error for DJIA index test set. Experimental results based on DJIA index show that LSTM-ARIMA hybrid model with news sentiment can forecast prices with accurate performance as it considers both market sentiment and past price values.

REFERENCES

[1] Xinyi Li1 , Yinchuan Li2,1 , Hongyang Yang1 , Liuqing Yang1 , Xiao-Yang Liu1 , "DP-LSTM: Differential Privacy-inspired LSTM for Stock Prediction Using Financial News " Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] Siyuan Liu, Guangzhong Liao, Yifan Ding, " Stock Transaction Prediction Modeling and Analysis Based on LSTM "

[3] Choi, Hyeong Kyu. "Stock price correlation coefficient prediction with ARIMA-LSTM hybrid model." *arXiv preprint arXiv:1808.01560* (2018).

[4] Bhardwaj, Aditya, Yogendra Narayan, and Maitreyee Dutta. "Sentiment analysis for Indian stock market prediction using Sensex and nifty." *Procedia Computer Science* 70 (2015): 85-91.

[5] Rajput, Vivek, and Sarika Bobde. "Stock market prediction using hybrid approach." *2016 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2016.

[6]Omkar S. Deorukhkar1, Shrutika H. Lokhande2, Vanishree R. Nayak3, Amit A. Chougule4, "Stock Price Prediction using combination of LSTM Neural Networks, ARIMA and Sentiment Analysis ",International Research Journal of Engineering and Technology (IRJET)

[7]Sun, J. (2016, August). Daily News for Stock Market Prediction, Version1.

[8] Sidra Mehtab, Jaydip Sen, Subhasis Dasgupta, "Robust Analysis of Stock Price Time Series

Using CNN and LSTM-Based Deep Learning
Models”