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

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

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
Bangalore, India
nityam.churamani@gmail.com

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

2022 International Conference on Futuristic Technologies (INCOFT) | 978-1-6654-5046-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INCOFT55651.2022.10094422

***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.**
Due to the noise and volatility of the stock market, timely market prediction is typically regarded as one of the most difficult challenges. We suggest a deep learning-based stock market prediction model that takes investors' emotional tendencies into account to overcome these issues.
This paper uses a unique method to predict next day's final stock prices using a combination of LSTM, 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 are going to thus be able to minimize the loss of their money and resources.

***Index Terms*—LSTM, CNN, ARIMA , Sentiment Analysis**

I. 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, are affected by a lot of factors like news, opening price, closing price, etc.

However, the process by which stock prices are formed is very intricate. Stock prices will vary as a result of the interaction of numerous variables and the unique behaviour of each one, including technological advancements, investor behaviour, political, economic, and market variables. As a result, stock prices fluctuate frequently, which gives room for speculative actions and raises the market's risk. This form of risk not only has the potential to cause financial losses for investors, but it also has the potential to have negative repercussions on the way businesses and nations conduct their economies.
Autoregressive model (AR), moving average model (MA), autoregressive and moving average model (ARMA), and autoregressive integrate moving average model (ARIMA) are common time-series analysis techniques. All of these methods primarily concentrate on the time series itself, neglecting other affecting elements such the background data.

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.

II. Literature Survey

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 would not produce results.
Was able to effectively produce results for short range periods, but not on long range scales like 1 year or 2 years. This was a big disadvantage in LSTM. This model has increased the correlation coefficient by 5.74% and the accuracy of the rise and fall classification by 11.01%.

[2] Uses regression along with LSTM to predict the values of stock. This method does not take sentiment of the stock into account and therefore is not as effective as the ones with sentimental analysis. However, this model is relatively simple to understand and implement. This model cannot be used to predict certain sharp falls and peaks triggered by variations in stock prices. Thus, it does not provide an accurate results for this problem statement as the nature of stock prices is very volatile.

[3] Uses CNN 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

978-1-6654-5046-1/22/$31.00 ©2022 IEEE

features. LSTM helps with remembering data values.
To build their predictive models, historical stock price data of a company listed in the National Stock Exchange (NSE) of India during the period December 31, 2012 to January 9, 2015 was used
However, this model does not take sentiment of the stock into account. Previous one week's stock values are taken as input and forecasting of next week's open values are made. Only one attribute, that is, the open value of the stock is fed to the CNN

[6] Indicated the use of a multimodal ARIMA - LSTM model. The numerical prediction was working in the desired manner here, but the model had a low accuracy as not many optimisations were done on it. And again the problem of prediction of prices on erratic days prevailed.
This was an innovative approach and we added to it by taking sentiment of the stock into consideration. This approach enabled the use of sentiment of a stock obtained via news headlines and thus enabled us to combine the two approaches and come up with a new approach.

By using traditional time-series analysis techniques, the stock market has been predicted numerous times. In order to predict stock prices, Tang [8] evaluated the use of the ARMA-GARCH model expanded by the AR-GARCH model. The author of [9] improved the forecasting of stock market volatility by building an autoregressive dynamic Bayesian network (AR-DBN) based on dynamic Bayesian network (DBN) [10] and inferring the market index. The authors in [11] integrated the traditional ARMA model with SVMs to undertake stock market prediction, maximising the benefits of both models and resulting in a model with more explanatory power.

III. Methodology

The methodology behind this prediction model involves the use of LSTM and 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.

A. Data Preprocessing

Trend varies over time,to eliminate trend we applied transformation to penalize higher values than smaller values

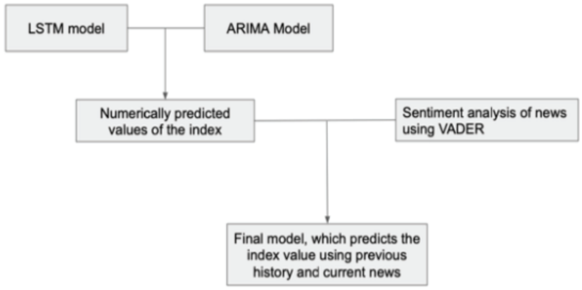
LSTM model
ARIMA Model
Numerically predicted values of the index
Sentiment analysis of news using VADER
Final model, which predicts the index value using previous history and current news

Fig. 1. Methodology

, 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.

On the news dataset, which was a text dataset, we pre processed it by tokenizing it, and lemmatizing the text and avoiding punctuation's for clean and smooth processing.

B. LSTM Networks

Recurrent Neural Network (RNN) remembers input with the help of their internal 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.

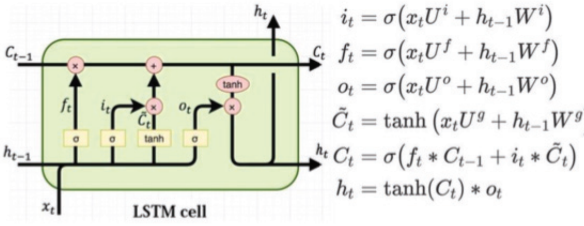
h_t
C_{t-1}
C_t
h_{t-1}
h_t
x_t
f_t
i_t
$\tilde{C}_t$
o_t
σ
tanh
LSTM cell
$i_t = \sigma(x_t U^i + h_{t-1} W^i)$
$f_t = \sigma(x_t U^f + h_{t-1} W^f)$
$o_t = \sigma(x_t U^o + h_{t-1} W^o)$
$\tilde{C}_t = \tanh(x_t U^g + h_{t-1} W^g)$
$C_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t)$
$h_t = \tanh(C_t) * o_t$

Fig. 2. Working of a LSTM cell

i_t, f_t and o_t represent input, forget and output gate respectively. ct represents the cell state, h_t represents hypothesis and ‘’ represents the activation function respectively. The loss function used here is the mean squared error(MSE) with 100 epochs.

$$\text{MSE=} \begin{cases} \frac{1}{2}(y-\hat{y})^2 & if\ |(y-\hat{y})| < \delta \\ \delta((y-\hat{y}) - \frac{1}{2}\delta) & otherwise \end{cases}$$

Our LSTM Network has 2 cells and undergoes 100 epochs.

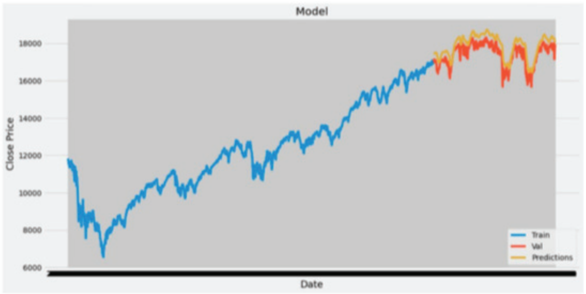

Fig. 3. Accuracy in closing prices using the LSTM Model

C. ARIMA

This paper uses a time series model called 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)

- The number of auto-regressive factors is given by p.
- d represents the number of times the series is differenced.
- q reflects the 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.

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.

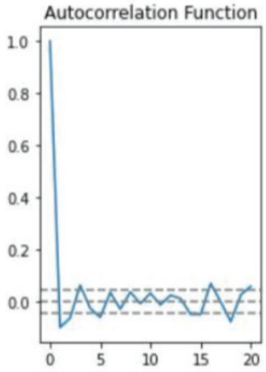

Fig. 4. Auto Correlation Function (A.C.F) Plot

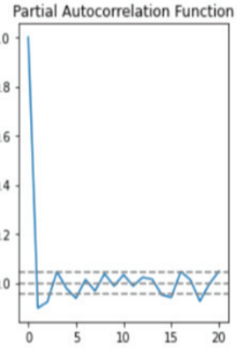

Fig. 5. Partial Auto Correlation Function (P.A.C.F) Plot

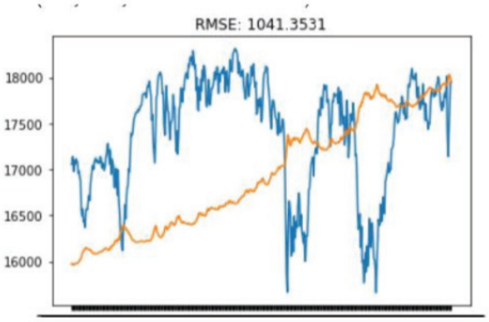

Fig. 6. Root Mean Square Error (RMSE) obtained using the ARIMA model

D. 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.

$\mathrm{X_f} = \mathrm{X_o}$ + compound_score*$\sigma\mu$

$\mathrm{X_o}=\frac{\alpha}{\lambda}$*ARIMA_prediction_value+$\frac{\beta}{\lambda}$*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, $\mathrm{X_f}$ is the final prediction and $\mathrm{X_o}$ 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.

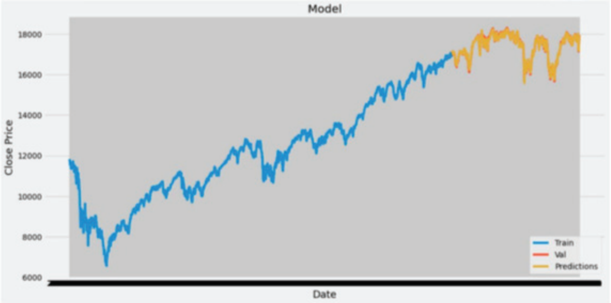

Fig. 7. Final results obtained using LSTM+ARIMA and sentiment analysis

IV. Conclusion

This multimodal approach of using LSTM and ARIMA was useful in predicting the stock prices on normal days without high fluctuations, however it wasn't able to predict stock prices on erratic days where the markets would fluctuate because of external factors. 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.

This novel approach combined two-thought processes, that is, the hybrid LSTM and ARIMA along with sentiment analysis of stocks via news headlines. This approach provides a real time prediction technique for stock prices, taking multiple factors into account and not just the open or close values of stocks.

V. Acknowledgement

Expressing profound gratitude to Associate professor Badri Prasad VR sir, for providing us with the opportunity to work on this project and imparting his wisdom in making us aware of new technologies we can use and helping us design the overall model of the project. In the end we would like to thank the Department of Computer Science and Engineering at PES University for conducting mind stimulating research based courses and allowing the students to explore the depths of subjects. acknowledgement

References

[1] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 205-208, doi: 10.1109/BigDataService.2019.00035.

[2] Hum Nath Bhandari, Binod Rimal, Nawa Raj Pokhrel, Ramchandra Rimal, Keshab R. Dahal, Rajendra K.C. Khatri, Predicting stock market index using LSTM, Machine Learning with Applications, Volume 9, 2022, 100320, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2022.100320.

[3] S. Mehtab, J. Sen and S. Dasgupta, "Robust Analysis of Stock Price Time Series Using CNN and LSTM-Based Deep Learning Models," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1481-1486, doi: 10.1109/ICECA49313.2020.9297652.

[4] 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)

[5] Bhardwaj, A. , Narayan, Y. , Vanraj, Pawan, Dutta, M. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. Procedia Computer Science, 70 . doi: 10.1016/j.procs.2015.10.043

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

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

[8] Tang H, Chiu KC, Lei X (2003) In: Proceedings of 3rd international workshop on computational intelligence in economics and finance (CIEF'2003), North Carolina, USA, September 26–30, pp 1112–1119

[9] Duan T (2016) Auto regressive dynamic Bayesian network and its application in stock market inference. In: IFIP international conference on artificial intelligence applications and innovations. Springer, Berlin

[10] Hinton G, Deng L, Yu D et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97

[11] Zhang DZD, Song HSH, Chen PCP (2008) Stock market forecasting model based on a hybrid ARMA and support vector machines. In: International conference on management science and engineering. IEEE