Prediction of Stock Market Returns using LSTM Model and Traditional Statistical Model

Seth Gyamerah
Department of Computer Science
C.K. Tedam University of Technology and Applied Sciences
Navrongo, Ghana

Dennis Redeemer Korda
Department of ICT
Bolgatanga Technical University
Bolgatanga, Ghana

ABSTRACT
Despite the growing interest in time series data specifically, stock market predictions in the financial world as well as its development stages of most related studies, this article aim to provide a good structure and suitable model for predicting the trend movement of stock market returns. This is done through a classification wavelet LSTM network model and the result compare to a baseline model. The results show that there are high returns of S&P500 stock market prices with a 15minutes interval range as compared to the wavelet-logistic (W-LR) regression model. It is therefore obvious that the enhanced deep neural network (W-LSTM, LSTM model) performs better in stock market prediction as compared to the traditional statistical models (LR W-LR).

Keywords
LSTM model, Logistic regression model, Stock market (S&P500) and classification problem, Wavelet transform function.

1. INTRODUCTION
Recently there has been a lot of research publications on financial time series especially in the area of returns on stock markets. The interest is that if the returns on stocks are successfully predicted investors may be better guided. The profit in investment and trading by the individual, institution in stock market returns to a large extent depends on predictability. If any program is developed and can continuously predict the returns of the dynamic stock markets, it would offer enormous support in decision making by stock investors. Moreover, the forecasted returns of the stock market will help the market analyst in making correct investment decisions. Another concern for research in this field is that it has many theoretical and experimental challenges. The most concern is the Efficient Market Hypothesis (EMH: EUGENE FAMAS 1970). The Hypothesis explained that stock prices are not the precise information about the market and its constituents and that the opportunity of earning more profit ceases to exist. So, it is assumed that no system is expected to outperform the market predictability. Hence modelling any market under the Expected market Hypothesis, the assumption is that it is only possible on the speculative, stochastic components but not on the changes in value. There are a lot of arguments on the volatility of the Efficient Market Hypothesis and random walk theory. The advancement in computation and finance intelligence and economist have formulated opposite hypothesis called inefficient market hypothesis which states financial markets are not always affected, the markets are not always in a random walk.To handle these complicated components and make an accurate prediction, a lot of scholars choose to use machine learning and deep learning to create a model which has been applied in financial time series forecasting and published in computer science, economics and finance journals. In this paper, we propose a classification problem on the S&P500 stocks market prices, where we were more concerned with the returns of the closing prices. Our classification problem was built on the LSTM model and we compare the model prediction to the traditional statistics model (Logistics regression). The rest of the paper is structured as follow: literature review, the methodology and development, experiment, experimental result, discussion and conclusion.

2. LITERATURE REVIEW
Over the years many traditional statistical models and machine learning algorithms have been employed to predict the trend movement of stocks market data [1][2]. The most popular ones used are Autoregressive integrated moving average, ARIMA [3] is one of the most popular techniques used in financial time series analysis. The improved modern techniques in the forecasting of financial time-series data, Deep LSTM, Shallow LSTM, 1-CNN and machine learning models were used to predict stock market data [4]. Attention LSTM model has been used in prediction of financial time series data [5] Wind power short-term prediction based on LSTM and discrete wavelet transform functions by Liu Y, Goun, L, Hou, C et al [6]. Leveraging Financial News for Stock Trend Prediction with Attention-Based Recurrent Neural Network [7]. Application of support vector machines in financial time series forecasting [8]. Using support vector machine with a hybrid feature selection method to stock trend prediction [9]. Forecasting stock market using wavelet transform and recurrent neural networks, an integrated system based on artificial bee colony algorithm.[10]. Yatong and Takahiro in [11] implemented deep learning models as a strategy for trading. Manuel R.Vergas et al in [12] use the LSTM model and Technical indicators to forecast the S&P500 index. In [13] a hybrid model based on the LSTM network and several GARCH models is used to forecast the stock prices volatility. Other notable researchers, applied interesting models for forecasting stock returns in the cross-section [14][15][16][17][18].

3. METHODOLOGY AND DEVELOPMENT

3.1 Propose classification problem
The proposed classification problem consists of several steps that will be detailed in this section. First, we assume the volatility of the prices at time step t can move either high or low. In such instance we focus on the returns of the market

price by taking logarithm returns of the prices, we define our returns as:

$R_t = \frac{P(t+h)-P(t)}{P(t)}$, Where $P(t)$ is the closing price at time t and it shows the changes of the price between time t and $(t+h)$ relative to the reference price $P(t)$

$R_t = \frac{m(t)-P(t)}{P(t)}$, where $m(t)$ is moving averages from $(t+1)$ to $(t+h)$ period and where

$$m(t) = {}^{1}/_{h} * \sum_{i=1}^{h} P(t+i)$$

$$R_t = \log\left(\frac{{}^{1}/_{h} * \sum_{i=1}^{h} P(t+i) - P(t)}{P(t)}\right)$$

We partition it by 1/3 and 2/3 quartiles and since the distribution of R_t is approximately symmetric and stationery which is denoted by $P_t^{(high)}$, $P_t^{(low)}$. We will then have classification problem as: $X_t \epsilon (high, low)$

3.2 Class Label

We defined our class label as one-hot vector where the true class (High returns) is set to 1 otherwise (Low returns) is set to 0 and θ as our threshold. We use a threshold to control the imbalance of the class labels.

(S&P500 θ = 0.23)

$$y_t = \begin{cases} 1 & if\ , R_t > R_{t+T} - \theta \\ 0 & if \quad R_t < R_{t+T} + \theta \end{cases}$$

In such instance our simplified classification problem is

$$X_t \epsilon (y_t)$$

3.3 Background of LSTM Network Model

The LSTM network unit above figure (1) has vectors $h_{(t)}$ and $c_{(t)}$ which represents short term and long-term memory states. Each of the cells has three gates thus: for the gate $f_{(t)}$, Input gate $i_{(t)}$ and output gate $o_{(t)}$.The logits' function which is the activation function output is in the range of [0][1]. The network gates use element by element to control how data should be maintained. The mathematical computations are as follows:

$$i_{(t)} = \sigma(W_{xi}^T . X_{(t)} + W_{hi}^T . h_{(t-1)} + b_i)$$

$$f_{(t)} = \sigma(W_{xf}^T . X_{(t)} + W_{hf}^T . h_{(t-1)} + b_f)$$

$$o_{(t)} = \sigma(W_{xi}^T . X_{(t)} + W_{ho}^T . h_{(t-1)} + b_o)$$

$$g_{(t)} = tanh(W_{xy}^T . X_{(t)} + W_{hg}^T . h_{(t-1)} + b_g)$$

$$C_{(t)} = f_{(t)}(\times) . C_{(t-1)} + i_{(t)}(\times) g_{(t)}$$

$y_{(t)} = h_{(t)} = o_{(t)}(\times)\tanh(C_{(t)})$, where σ is given in the function as $\sigma_{(x)} = \frac{1}{1+\ell^{-x}}$ and tanh represent the function as: tanh(x) = $\frac{\sin hx}{\cos hx} = \frac{\ell^x - \ell^{-x}}{\ell^x + \ell^{-x}}$

The weight matrix **W** and the bias **b** will be generated by the backpropagation algorithm. Where f_t i_t and O_t are the activation of the input, forget and output gates at time step **t**, whichcontrol how many input and previous state will be and how many cell states will be added in the hidden activation of the network. The protected activation cell state at (time-step) time-step is represented by $C_{(t)}$ and $h_{(t)}$ the activation that will be given to the othercomponents of the LSTM model.

3.4 LSTM Model Construction

In the model, we use Xto represent the features of stock at time t. we treat a time series as {X} (t = 1, 2, ..., T) and returns as an input instance of our LSTM network which will possibly predict the high or low of the next period.We make our problem more feasible by the already proposed classification model built at the initial stage which will depend on the LSTM network.

The above gives details of the LSTM layer model, in the dropout layer, each cell has a defined probability that stops working during the training process to avoid the model being over-fitted. Predicting the neurons will undertake the work to maximize the role, the dense layer has no activation function and the output score is the probability that predicts instance belongs to a pre-defined category and the real score is converted to the corresponding probability by the Softmax layer. We defined the Softmax layer as:

$$P_i = \frac{\ell^{i}}{\sum_j \ell^i}$$

where P_i and i represents the probability of the predicted score that belongs to a particular class.

3.5 Baseline approach

In comparing the LSTM model to a baseline model thus statistical model, we chose the logistic regression model which is a traditional model and compare it to the LSTM model. The logistic regression which is a probabilistic review of the linear regression model has a logit function that finds the probability P such that

$$P\left(y = \frac{1}{x}\right) = p$$

where X is the feature input vector of our features and Y is the target class. We assume that the probability P depends on X such that is P=βX andβ is an input feature co-efficient. Here we use the Sicket-learn library in our final implementation of the logistic regression model. We use the same input OHLC and returns.

3.6 Dataset and Processing

We evaluate the performance of the proposed classification problem on S&P500 ranging from Jan 1, 2010, to 31 Dec 2015. There was a total of 1250 trading days for every 15 minutes interval. We chose open prices, high price, low price, close price and volume price and drop the volume prices in the progress of the work.

Due to the difficulty and volatility of the stock market and its various trading restrictions, the stocks prices are very noisy. Non-stationarity of time series data with overlapping noise and signals, make an accurate prediction of trend very complex to achieve. A wavelet transform function is introduced in the preprocessing stage to denoise the S&P500 dataset. The wavelet mathematical function use to transform the dataset is:

$$X_{\omega}(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right) dt$$

The co-efficient with more deviation is dropped and inversely transform a new set of co-efficient to achieve a new set of denoise S&P500 datasets. This would help to achieve an advanced convergence of the deep neural network model and also help to remove negative hyperparameters of the dimension of the selected features on the model. And we kept the initial dataset without denoising to sever as a comparison of metrics score of wavelet function on the dataset and without a wavelet function on a dataset.

S&P500 dataset:

```
2]: # drop some features that do not have information.
    data = pd.read_csv('15_minutes.csv')
    #data = data.dropna()
    data = data.drop(['Time'],axis=1)
    data.head()
```

2]:

	Time	Open	High	Low	Close	
0	2015-12-29 00:00	1.09746	1.09783	1.09741	1.09772	4.8668
1	2015-12-29 00:15	1.09772	1.09800	1.09770	1.09790	4.4592
2	2015-12-29 00:30	1.09790	1.09805	1.09782	1.09792	1.2107
3	2015-12-29 00:45	1.09792	1.09825	1.09775	1.09808	1.1169

4. EXPERIMENT

We use wavelet transform and without wavelet transform on the S&P500 dataset on the proposed classification LSTM model and the baseline approach. In the S&P500 we divided wavelet transform and without wavelet transform dataset into training data, validation and test data. We use 80% training dataset, 10% validation and 10% testing data and created a Numpy array function and set a look-back for the previous time steps to be used to predict next time predictions. In the wavelet transform plus LSTM model (WLSTM network) on the S&P500 stock market, we set the look-back to a period of 60minutes and forecast 100minutes into the future to detect high or low returns.

We optimize the model using the Adam optimizer and softmax function activation in all the layers. We keep changing the number of iteration epochs and the selected number of mini-batches for the stocks and the dropout layers base on the selected mini-batches.

And for the baseline, we use the same input variables of the previous stocks and the same look-back function in terms of period. We implemented this using the Sicket-learn library and evaluated the scores of the proposed classification on the wavelet LSTM model and the Wavelet logistic regression model using these metrics: accuracy, recall, precision and F1-measure. These metrics were calculated based on predictions correctly made for positive classes (true positive-TP), negative classes (true negative -TN), those made for inaccurately for both classes (false positive- FP, false negative-FN) and confusion matrix to show clearly the high returns predictions and low returns predictions.

A simplified figure on confusion matrix for classification problems

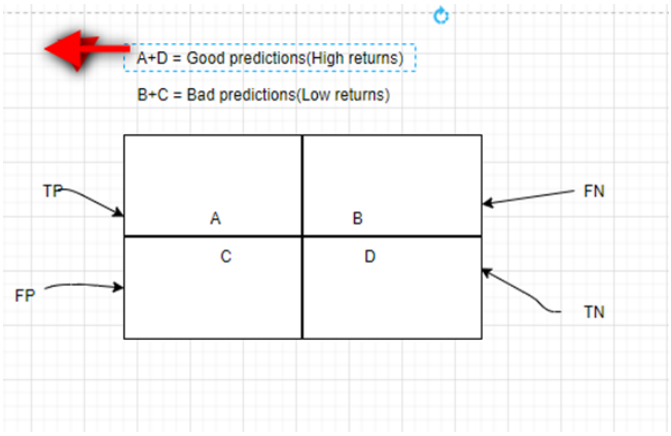

Fig.2: confusion matrix

$$A = \frac{t_p + t_n}{t_p + f_p + t_n + f_n}$$

$$P = \frac{t_p}{t_p + f_p}$$

$$R = \frac{t_p}{t_p + f_n}$$

$$F1 = 2 * \frac{P * R}{P + R}$$

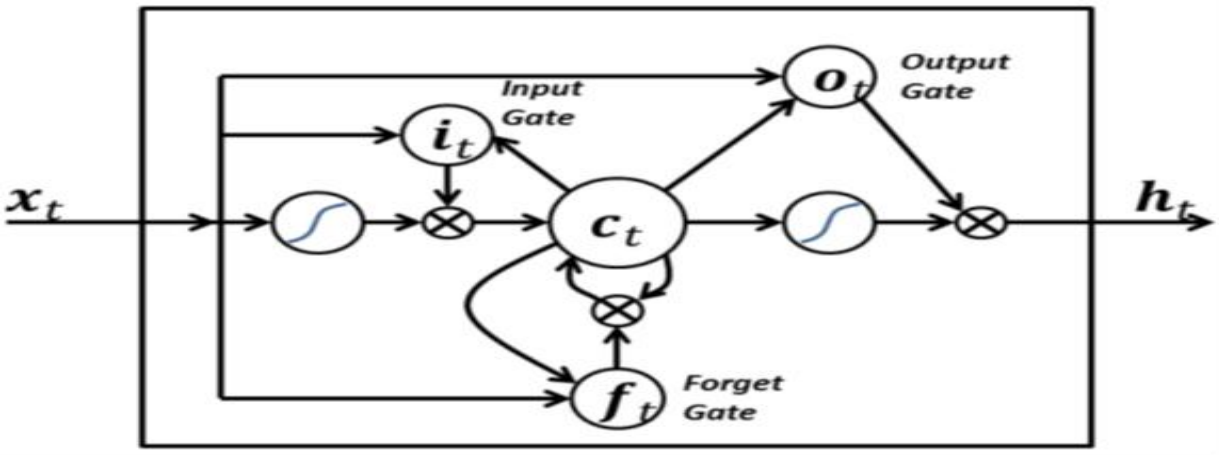

Fig 1: LSTM Network Model

5. EXPERIMENTAL RESULT AND DISCUSSION

We experimented with the wavelet transform dataset and without the wavelet transform the dataset, for the wavelet dataset, we choose the co-efficient with low deviations. We process the S&P500 dataset on a proposed classification problem on the LSTM network model and logistic regression model. We will then have wavelet transform - LSTM network model (W-LSTM) wavelet transforms - logistic regression model (W-LR).

The LSTM network model with wavelet transform function and logistic regression model with wavelet transform function gives good positive returns in terms of the metrics; recall, precision, accuracy and f1-score as compared to the LSTM network model and logistic regression without a wavelet transform on the dataset.

The table below shows the predictions of the S&P500 15minutes dataset.

Table 1. Wavelet transform function with LSTM (WLSTM) model and LSTM model

Metrics	W-LSTM model (%)	LSTM model (%)
Accuracy	0.68	0.58
Precision	0.60	0.56
Recall	0.82	0.68
F1-score	0.69	0.61

Table 2. Wavelet transform function with logistic regression (W- Logistic regression) and Logistic regression model

Metrics	Wavelet transform-Logistic regression model (W-LR) (%)	Logistic regression model LR (%)
Accuracy	0.57	0.55
Precision	0.50	0.50
Recall	0.60	0.55
F1-score	0.54	0.52

6. CONCLUSION

In this paper, it is shown that the proposed classification problem helps to determine the movement of financial stocks market data in terms of high or low returns as compared to most review articles. The stock market is affected by several factors which make it very arguable in terms of its trend movement. The noisy nature of the dataset is difficult to handle, we then denoise the dataset by using the wavelet transform function. We then have an enhanced model of the wavelet transform function plus the LSTM network model (W-LSTM) and compare it to the LSTM network model. In our chosen baseline, we enhance the logistic regression model (W-LR) and compare it to the logistic regression model.

From the evaluation of our W-LSTM model, it shows clearly higher return (Profit) in terms of true positive and true negative with an accuracy of 0.68% as compared to the 0.58% accuracy score of the LSTM network model. In our chosen W-LR (wavelet transform and logistic regression) as a baseline, we achieve a higher returns accuracy of 0.57% as compared to LR accuracy of 0.55% This shows clearly how the wavelet transform function help to reduce the noise in the dataset and enhance the accuracy of the predictive models. In short, the deep neural network model performs better on stock market price predictions as compared to the traditional statistical models. The chosen enhanced LSTM network model (W-LSTM) predicted higher returns on the prices as compared to the enhanced logistic regression model (W-LR). The result is more promising in accessing the trend of stock market returns as compared to the result review in most published articles. In future, we will look at an econometrics model (Elliott waves principles) with long term short memory (E-LSTM) as a hybrid approach to detecting the trend movement of stock market returns and comparing it to this result.

7. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers and editors whose helpful comments and suggestions improved the quality of the paper.

8. REFERENCES

[1] Gyamerah, S., & Awuah, A. Trend forecasting in financial time series with indicator system. J. Appl. Stat, 0-14.

[2] Korda, D. R., Agoha, S. A., & Haruna, D. A. (2017). Stock trend forecasting using regression analysis. [Unpublished manuscript]. Department of Computer Science, University of Cape Coast.

[3] Vijag Kotu; Deshgande, in data science edition (2019) using ARIMA models for financial time series predictions, Science Direct.

[4] Kelvin Li and Glen Yu (2000), Deep Learning for stock price forecasting. Journal of machine learning.

[5] Jiang Q, Tang C, Chen C, et al. Stock Price Forecast Based on LSTM Neural Network [2018] computing and application journal

[6] Liu Y, Guon, LHou,C Hau, H;Liu , Jun, Y, Zheng, M. wind power short-term prediction based on LSTM and discrete wavelet transform (2012) journal Machine learning 1108.

[7] Liu Huicheng. Leveraging Financial News for Stock Trend Prediction with Attention-Based Recurrent Neural Network (2018) Journal of science and informatics.

[8] Tay F.E.H. Cao L application of support vectors machine in financial time series forecasting (2007) journal of university of electronic science and technology, 29(4)-309-337

[9] Lee, M.C. using support vector machine with a hybrid feature selection method to stock trend prediction. Expert system, Appl (2009), 36,10896-10904

[10] H.Suchi, T. J Hsiao, H.F., Yeh W.C. Forecasting stocks market using wavelet transform and recurrent neural networks, an integrated system based on artificial bee colony algorithm 2011) Appl. Software computing 11,2510-2525

[11] Erban Beyaz, Firat Tekiner, Xiao-jun Zeng, john A.Keane(2018) comparing technical and fundamental indicators in stock price forecasting. International conference on High performance computing and communication, IEE 16th conference on smart city.

[12] Rajashree, Dash and Pradipta Kishone Dash (March,2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. The Journal of finance and Data science 2(2016) 42-57.

[13] Francisco J.Ruiz, Alberta Jama, German Sanchez, Jose A. Sanabria and Nuna Agell(2011) An interval technical indicators for financial time series forecasting .Journal of statistical software, vol.

[14] Weihong Huang and Yu Zhang (June 31,2014) Asymmetry Index of stock price fluctuations. Journal of Global Economics

[15] Basit Janvir, khan Moman Jarel etc (August 2017) Evolving Technical Trading strategies using Genetic Algorithms. A case study of Pakistan stock exchange. Conference Paper: IDEAL 2017

[16] Fiu Feng, Xiangnan He, Xiang Wnag etc (March, 2018) Temporal Relational Banking for stock prediction. Journal of HCM transactions information system 37(2) 1-30

[17] Tobias Schadler (Dec,25,2018) Measuring irrationality in financial markets. A chieves Business Research Vol.6 No.12

[18] Gary R Weckmen, Siriam Lakshminarayanan (May2004) Identify technical indicators for stock market prediction with Neural Network. Conferences: proceedings of the IIE Annual Conference.