LSTM Versus Transformers: A Practical Comparison and Combination of Deep Learning Models for Trading Financial Instruments

Ruiru Daniel[1] [a], Nicolas Jouandeau[2] [b] and Dickson Odhiambo[1] [c]

[1] *Strathmore University, Nairobi, Kenya*

[2] *Université Paris 8: Saint-Denis, France*

{druiru, dowuor}@strathmore.edu, n@ai.univ-paris8.fr

Keywords: Deep Learning, Recurrent Neural Networks, Long Short Term Memory (LSTM), Transformers.

Abstract: Predicting stock prices is a difficult but important task of the financial market. Often two main methods are used to predict these prices; fundamental and technical analysis. These methods are not without their limitations which has led to the use of machine learning by analysts and investors as they try to gain an edge in the market. In this paper, comparisons and combinations are made between Long Short Term Memory (LSTM) and the Transformer model in predicting five financial instruments; Gold, EURUSD, GBPUSD, S&P500 Index and CF Industries. This work starts with base models of LSTM, Bidirectional LSTM and Transformers. From the initial experiments, LSTM and Bidirectional LSTM have consistent results but with more trainable parameters. The Transformer model then has few trainable parameters but has inconsistent results. To try and gain an edge from their respective advantages, these models are combined. LSTM and Bidirectional LSTM are combined with the Transformer model in different variations and trained on the same financial instruments. The best models are then trained on the larger datasets of the S&P500 index and CF Industries (1990-2024) and their results are used to make a simple trading agent whose profit and loss margin (P/L) is compared to the 2024 Q1 returns of the S&P500 index. From the experiments LSTM+Transformer, Transformers and Bidirectional LSTM had the best predictions, having accuracies of 95%, 93% and 91% respectively. Ultimately, the first model was the best performer and was used to develop a basic trading agent achieving competitive returns in its best test runs (1.2 to 7.68%).

1 INTRODUCTION

Predicting stock prices is a difficult task owing to the ever changing and unpredictable nature of the financial market. Today, this goal of forecasting financial markets has garnered a lot of attention from both academic and industrial practitioners who hope to provide provable answers for price outcomes. Many methodologies have been suggested, often falling within two major categories, technical and fundamental analysis (Lui and Mole, 1998; Kehinde et al., 2023). The former defines techniques that evaluate investments based on price patterns/trends while the latter focuses on the intrinsic value of assets as well as the factors that influence price outcomes (Krishnapriya and James, 2023). While somewhat effective, these methods still face serious difficulties in providing solid answers or even forecast of financial prices as exemplified by significant fluctuations such as the 2008 financial crisis which unfolded despite the existing technical and fundamental analyses. As such, a new phenomenon of using machine learning algorithms to predict financial markets has been gaining popularity owing to the success of these techniques in other domains. Machine learning methods use data to provide predictions which academics and market traders/regulators can use to forecast prices. Frameworks like Long Short Term Memory Neural Networks (LSTMs) and other Recurrent Neural Networks (RNNs) are of particular interest in the financial domains because of the demonstrated superiority in dealing with time series problems (Fischer and Krauss, 2018a). LSTM further stands out as it has the ability to exploit and retain sequential data patterns over long periods of time. Recently, this attribute of LSTM was enhanced through the use of Transformers which shine in handling long dependencies of input elements on top of enabling parallel processing

[a] https://orcid.org/0000-0003-3197-8821
[b] https://orcid.org/0000-0001-6902-4324
[c] https://orcid.org/0000-0002-0968-5742

(Vaswani et al., 2017). Transformers have thus become more effective in dealing with tasks that use sequential data which explain their popularity in Natural Language Processing domain. A similar challenge is exhibited by the financial markets as they pose a task in the sequential decision making problem domain which this paper tries to solve. This paper compares LSTM, Bi-directional LSTM and Transformers through their performance in predicting a variety of financial instruments. Moreover, these two algorithms are combined to try and yield a superior prediction model.

The rest of the paper is organized as follows. Section 2 outlines the Related Works, Section 3 briefly discusses the Methodology, section 4 then reviews the Experiments which is then followed by the Results and Discussions in Section 5. Finally, Section 6 concludes the study.

2 RELATED WORKS

Deep learning algorithms have shown great ability to learn and approximate functions for non-linear problems when provided with a lot of data. LSTM and Transformers can learn a lot of information from sequential data which explains their recent application in complex time series modeling problems such as stock/financial price prediction (Lin, 2023). The application of these algorithms in the financial domain are discussed below.

2.1 LSTM and Bidirectional LSTM

The LSTM model is actually a type of recurrent neural network (RNN) that was developed for modeling sequential data with less memory dependencies than typical RNN. The time dependence in LSTM is split into a hidden state for short term and cell state for the long term. It uses gate units to manage the flow of information which addresses the issues of gradient vanishing and explosion found in conventional RNNs (Hochreiter and Schmidhuber, 1997). The LSTM cell shown in Figure 1 outlines a base architecture that includes input gates, forget gates and output gates. These gates selectively discard or allow information flow through the network which helps LSTM handle long sequences of information and have better memory effect for any recurring patterns. The input, forget and output gates are described the following equations:

$$\mathrm{i}_t = \sigma(\omega_i[h_{t-1}, x_t] + b_i) \tag{1}$$

$$\mathrm{f}_t = \sigma(\omega_f[h_{t-1}, x_t] + b_f) \tag{2}$$

$$\mathrm{o}_t = \sigma(\omega_o[h_{t-1}, x_t] + b_o) \tag{3}$$

Where: *i_t is the input gate, f_t is the forget gate, o_t is the output gate, σ is the sigmoid function, w_x are the weights of the respective gates, h_{t-1} are the outputs of the previous lstm block, x_t is the input of the current timestamp and b_x are the biases for the respective gates.*

The mentioned attribute of LSTM make it ideal for financial price modeling as it facilitates the exploration of long term dependencies found in market prices. Many scholars and industrial practitioners

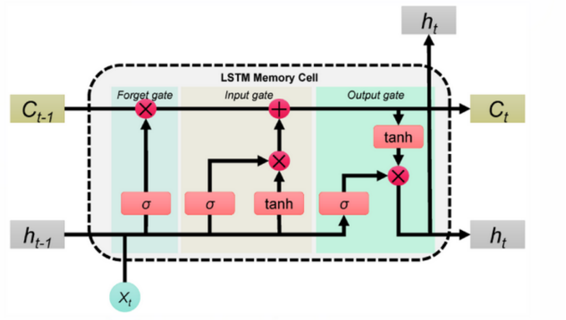

Figure 1: LSTM Cell

have applied this basic structure in the financial market. Roondiwala et al., (2017); Cao et al., (2019); Bao et al., (2017); Fischer and Krauss (2018b) all used LSTM as a deep learning model for predicting the financial market. In the Bao et al., (2017) case, they combined the base LSTM structure with stacked auto encoders which yielded better results by getting 63.026 percent returns in mainland China when compared zith other models that achieved rates below 40 percent. In the case of Siami-Namini et al., (2019), bidirectional LSTM (BiLSTM) were used to increase the prediction performance by learning long term dependencies across time steps of sequence of data (time series) in both directions. This variation of LSTM is useful if a problem requires an RNN to acquire information about an entire time series at various iterations such as that found in the financial market (Roondiwala et al., 2017; Cao et al., 2019; Bao et al., 2017; Fischer and Krauss, 2018b). A bidirectional LSTM can help gain a market's context in both direction which is key in sentiment analysis, a key component of predicting market outcomes.

2.2 Transformers

On its part, Transformers are deep learning models that also handle sequential data but unlike RNNs they use a self-attention mechanism across their encoder-decoder pair. At a high level, transformers

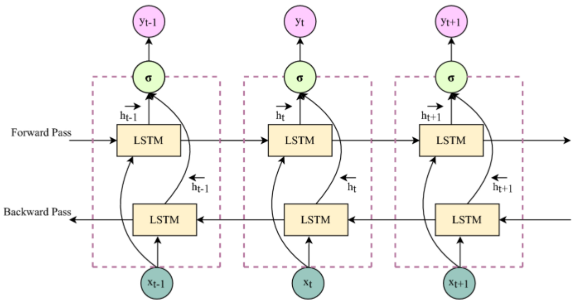

Figure 2: Bidirectional LSTM

can even be seen as a sequence to sequence model (encoder-decoder structure) with self-attention mechanism. The encoder is used to encode the input sequence while the decoder produces the output sequence (Vaswani et al., 2017; Lin, 2023). The self-attention mechanism found in this structure then helps to effectively pass information between the encoder-decoder pair in both directions (Xiao and Zhu, 2023). Moreover, self-attention also increases the encoding of the input sequence by the encoder. Figure 3 illustrates a transformer architecture combined with LSTM. The said attention mechanism is facilitated by three vectors from the encoder's input vectors. These three vectors are known as Query (Q), Key (K) and Value (V) which are simply abstractions used to calculate transformers attention (Cristina, 2023; Chen et al., 2024). In summary, these vectors are combined with a softmax function to give the attention mechanism as shown in equation 4. The role of the softmax function is to convert raw attention scores into probability distributions.

$$Attention(Q,K,V) = softmax(QK^T/\sqrt{d_k})V \quad (4)$$

Where: QK^T is the transpose of multiplying queries and keys in matrices Q and K. d_k is a scaling factor and V is the values vector.

The multi head attention in transformers borrows from this single attention mechanism by providing multiple heads of attention that run in parallel.

$$Multihead(Q,K,V) = Concat(head_1, \cdots, head_h)W^O \quad (5)$$

Where: $head_i$ =Attention(QWQ, KWK, VW ')

This attention techniques make transformers have great efficiency and versatility, attributes that are ideal for sequential decision making tasks. In particular, their ability to handle long sequence of data and parallel process all the available data makes them superior in capturing long range dependencies like those in long term financial results/assessments.

3 METHODOLOGY

3.1 The Problem and Methodology

Comparing and combining LSTM/BiLSTM with Transformers to predict financial market prices was the main goal of this paper. This goal arose from the difficulty of forecasting market prices where numerous analytical methods have been used. While a variety of machine learning models have been developed in this space, minimal work has been done to combine RNNs like LSTM with Transformers.

As such, this paper uses LSTM and Transformers together with their variants as reviewed in the literature review. There was also an initial focus on the broadest samples of these deep learning models, with modifications made to both the architecture and parameters. The predictions made were done in four stages. First, data collection, where a variety of datasets were obtained to evaluate the proposed models' performance. Second, data pre-processing where the collected data was enhanced to meet the requirements of the defined models. Third, LSTM, Transformer and LSTM + Transformer architecture and parameter optimization to achieve the best results possible. Finally, predictions were made based on trade simulations. Figure 3 summarizes this general roadmap used in the experiments of this paper.

To develop the final models used to predict financial instruments, defining the problem also involved modifying the two base models LSTM and Transformer. To start with, a conventional unidirectional LSTM and Bidirectional LSTM were used. These base models were then trained on the financial instruments with their results recorded based on the evaluation metrics identified below (MAPE, Run-time and number of parameters). Thereafter a basic transformer model was implemented. This transformer model was then trained on the financial instruments identified for this study and the results were recorded based on the evaluation metrics.

3.2 Combining LSTM and Transformers

Finally, the (best performing) LSTM model (based on variation of layers) and Transformer model were combined. In this case, the LSTM layer was added just before the attention block of the Transformer as illustrated in Figure 3 and Algorithm 1. This resultant LSTM + Transformer model was then trained on the financial instruments and the results were recorded based on the same evaluation metrics used in the previous models training. This combination was the final

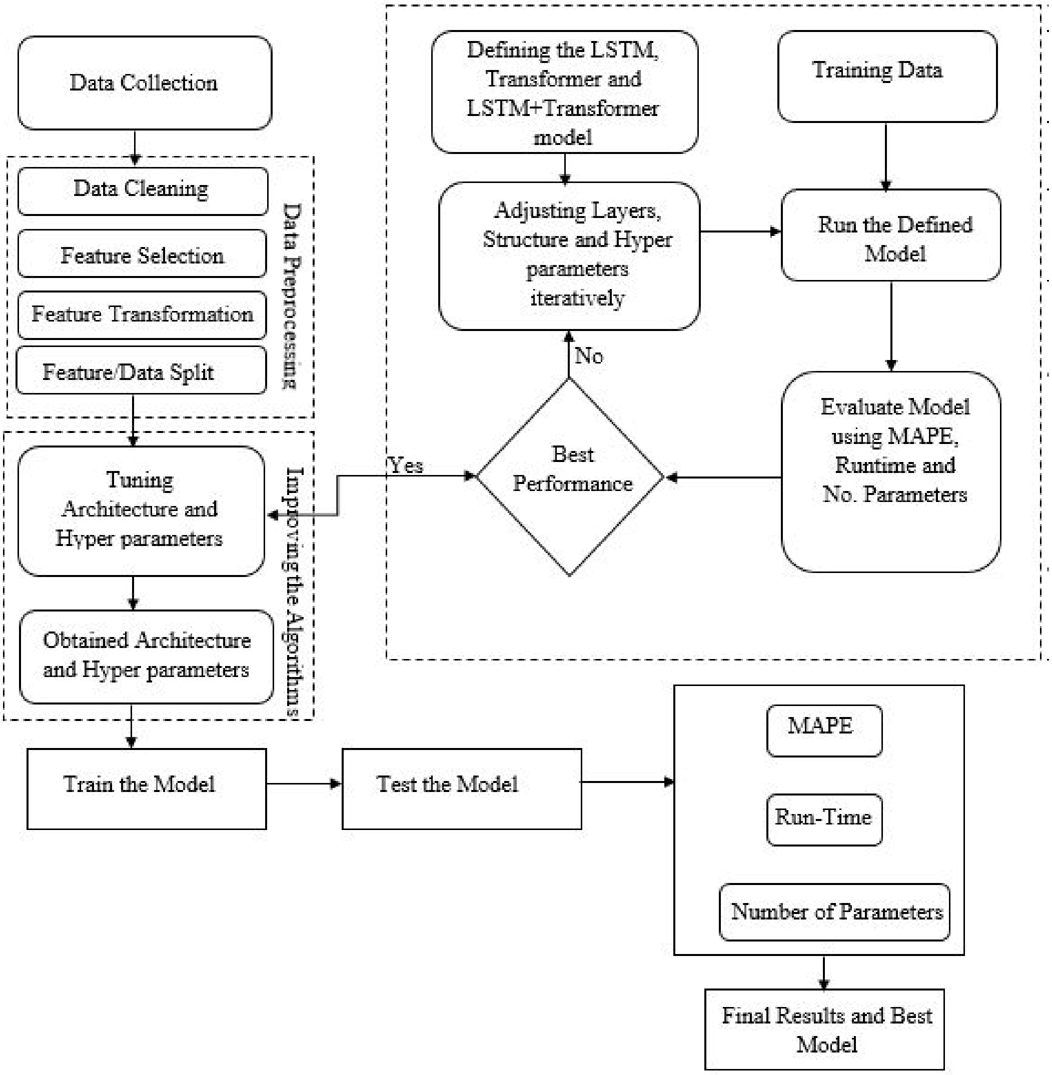

Figure 3: The General Road-Map

step of the experiments conducted.

3.2.1 Combining BiLSTM and Transformers

To cover the baseline model, Bidirectional LSTM was also introduced. A similar structure was followed with the combination of Bidirectional LSTM with Transformer where the BiLSTM layer(s) was placed before the attention block of the Transformer model.

3.3 Evaluation Metrics

Many evaluation metrics exist to assess the accuracy of predicted results each with their own benefits. For the experiments conducted, one conventional metrics was used Means Absolute Percentage Error (MAPE) owing to its ability to define the accuracy of forecasting methods. MAPE outlines the average of the absolute percentage errors of each entrant in a dataset which then calculates how accurate a forecasted result is compared to an actual result (Jierula et al., 2021). The choice of MAPE was also driven by its ability to effectively analyze large sets of data such as those found with the datasets used in this paper's experiments. Reviewing MAPE is also a straightforward process, with a 10 percent MAPE indicating a 10 percent deviation between the predicted value and the actual value. Equation 6 summarizes MAPE.

$$\mathrm{MAPE}(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} \frac{|y_i - \hat{y}_i|}{|y_i|} \tag{6}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value and N is the number of fitted points.

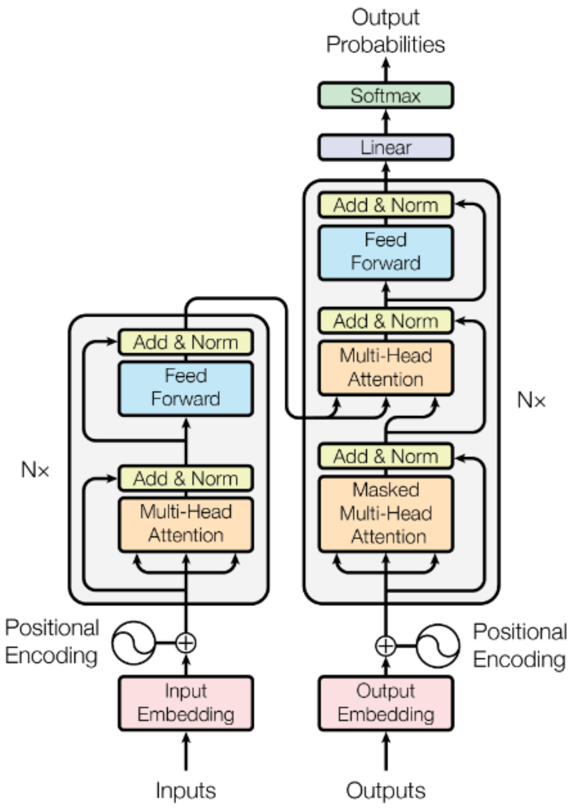

Figure 4: Transformer Architecture - LSTM/BiLSTM to be Added After Both Positional Encoding

Algorithm 1: LSTM/BiLSTM and Transformer

1: **Step 1:** Input
2: **Input:** Sequence $x_1, x_2, \ldots, x_T$
3: **Positional Encoding:** Encoding
4: **LSTM/BiLSTM:**
5: *Process data to output hidden states* $h_1, h_2, \ldots, h_T$
6: **Step 2:** Transformer
7: *Self-attention mechanism*
8: *Multi-head attention*
9: **LSTM/BiLSTM Location**
10: **Decoder**
11: *Transformer decoder generates output sequences*
12: **Output:** Fully connected layers
13: **Loss Function:** Cross-entropy or other appropriate loss function
14: **Step 3:** Backpropagation
15: Compute gradients and update weights
16: **Step 4:** Repeat
17: Iterate over epochs to train the model

In addition to MAPE, Run Time and Number of Parameters used per model were used to assess the final results. Generally, the prediction accuracy of any model was not solely judged by the deviation from the true value but also on the time it took to achieve its predicted values (Run Time). Also, the number of parameters used was a factor as they define the resources used to run any given model. As such, in an ideal case, a high accuracy of the predicted results would be achieved with minimal run time and minimum number of parameters. As such, while trying to achieve high performance/accuracy, the developed model was expected to have minimal training parameters and run-time.

4 EXPERIMENTS

4.1 Data Selection and Preprocessing

The experiments done were based on large dataset for all the five instruments. For the first 2 experiments data from the 1st January 2013 to 2024 was used. From experiment 3 to 5, only S&P500 and CF data was used. The data used here was also larger, as it included the period 1990 to 2024. In all the dataset, the data collected was of the OHLC format (Open-High-Low-Close). To get more stable analysis and results, the daily time frame was used. The data was collected from Yahoo Finance using the “yfinance” API. The S&P500 and CF dataset was also manually obtained from Stooq (stooq.com), an online resource that provides historical data for indices, stocks, bonds, forex and other form of financial data.

4.2 Experiment 1

To get a benchmark for the experiments, the three main models; LSTM, Bidirectional LSTM and Transformer were trained on the five financial instruments. For the LSTM model, a base architecture with one layer of 128 units was used, ultimately generating 73265 trainable parameters. A similar number of layers and units was used for the Bidirectional LSTM (i.e. 1 layer with 128 units) and the total number of trainable parameters ended being 146225. Finally, the transformer model was based on a single transformer block which was then replicated to have the needed number of blocks. For a start, 4 transformer blocks were used resulting in 17205 trainable parameters. These three models were used to conduct the first round of experiments with rankings being done based on MAPE, Run-time and the number of parameters used to achieve their results. The results of this first round (Experiment 1) are highlighted in the Results and Discussion section.

4.3 Experiment 2

In an attempt to combine LSTM and Transformer to yield a superior model, several variations of LSTM + Transformers were developed and trained on the five financial instruments. First, there was the base LSTM with 1 layer of 128 units plus the basic four block transformer model (here named LSTM128+TX). Subsequently, LSTM with one layer of 64 units, 32 units and three layers of 128, 64 and 32 units were combined with the transformer model (there naming following a similar pattern as LSTM128+TX). Similarly, the Bidirectional LSTM was combined with the transformer model with a familiar naming pattern: BiLSTM128+TX, BiLSTM64+TX, BiLSTM32+TX and BiLSTMAll+TX. These combinations and the number of trainable parameters used are summarized below in Table 1. These combinations of LSTM/BiLSTM and Transformer were then ranked based on their MAPE, Run-time and number of parameters. The results of these tests (Experiment 2) are highlighted in the Results and Discussion section.

Table 1: Combined Models and Number of Parameters

Model	No. Parameters
LSTMAll + TX	833394
LSTM128 + TX	1523346
LSTM64 + TX	651474
LSTM32 + TX	301554
BiLSTMAll + TX	2083794
BiLSTM128 + TX	3430930
BiLSTM64 + TX	1392274
BiLSTM32 + TX	618706

4.4 Experiment 3

For the third experiment, the three best models, (either base model or combination models) were then used to predict the S&P500 index and CF industries. Prior to experiment 3, a quick summary of the best performing models, based on experiment 1 and 2 was done to help identify the 3 best models.

4.5 Experiment 4 and 5

In this final round of experiments, the two financial instruments (S&P500 and CF Industries) were predicted with the three best models from experiment 3. Also, this final round of experiment saw an in-depth review of the prediction of the CF industries prediction by the best model. Thereafter, this prediction/forecast was compared with the returns of investing in the S&P500 index in the last month of the dataset used. In all, this final test was to see whether the best model from these experiments would challenge the returns of the S&P500 index in 2024.

5 RESULTS AND DISCUSSIONS

Table 2: Sample of Experiment 1 Results - Gold

Experiment 1: Gold			
Model	**MAPE**	**Run-Time**	**No. Para**
LSTM128	0.0143	84.27s	73265
BiLSTM128	0.0134	51.50s	146225
TX	0.0134	150.24s	17205

The Tables 2 and 3 highlight a sample of the results of the first and second experiment where all models were tested with the five financial instruments. From experiment 1, one results that stood out was the lack of consistency in the results of the Transformer model. Over a run of multiple training iterations including some of the same parameters and financial instruments, the Transformer model failed to have similar or comparable results. Moreover, the first experiment also broke a notion that was held at the start of the tests that the fewer number of parameters in the Transformer model would results in a shorter run-time. Surprisingly, the Transformer model regularly took the longest time to run the training but yielded competitive results. On their part, the LSTM and Bidirectional LSTM were more consistent with their results, including have comparable outcomes in multiple iterations of same financial instruments.

Table 3: Sample of Experiment 2 Results - Gold

Experiment 2: Gold			
Model	**MAPE**	**Run-Tim**	**No.Para**
LSTMAll+TX	0.0140	351.37s	833394
LSTM128+TX	0.0155	333.65s	1523346
LSTM64+TX	0.0183	213.63s	651474
LSTM32+TX	0.0168	151.58s	301554
BiLSTMAll+TX	0.0147	604.48s	2083794
BiLSTM128+TX	0.0142	518.57s	3430930
BiLSTM64+TX	0.0199	280.53s	1392274
BiLSTM32+TX	0.0186	221.33s	618706

On experiment 2, the combination of LSTM/BiLSTM with Transformer model produced models that had consistent results. It was much easier to replicate a prediction with them which validated their use. Of note was the Run-time of the BiLSTMAll+TX and BiLSTM128+TX which was often the longest in any of the categories tested. This

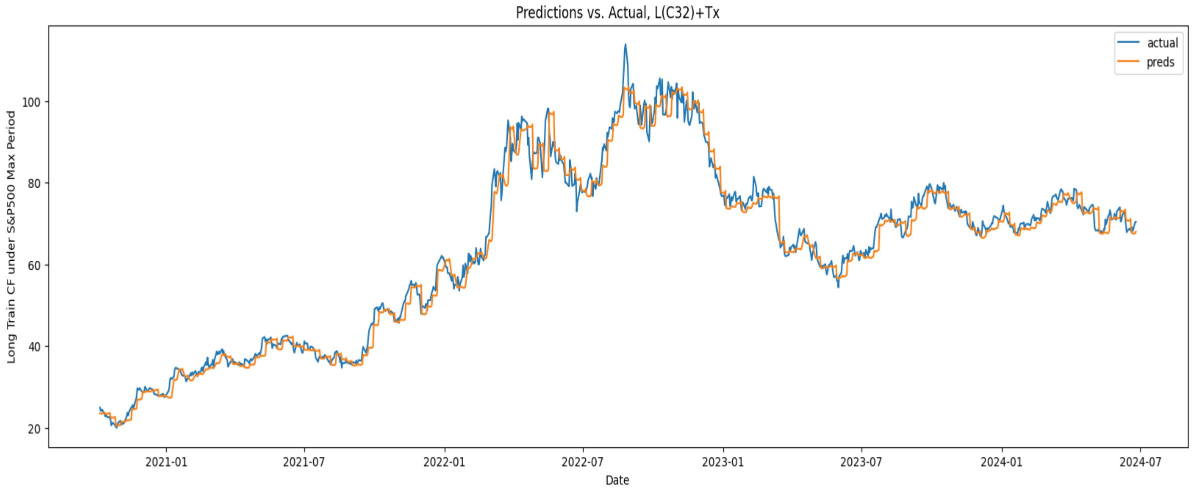

Figure 5: Experiment 4 Prediction

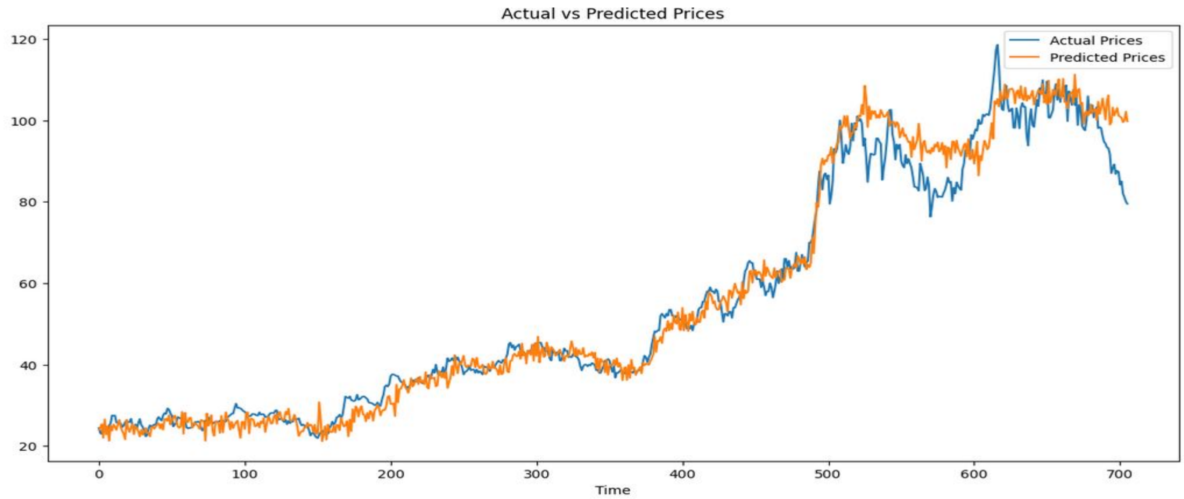

Figure 6: Visualizing CF Industries Trading

outcome is easily attributed to the total number of trainable parameters used; 2083794 and 3430930 respectively.

The ranking done in experiment three saw, 3 models emerge as favorite based on their leading performance in MAPE, Run-time and number of training parameters (fewer being better). The three models were LSTM with 1 layer of 32 units combined with Transformer (LSTM32 + TX), the base Transformer model (TX) and the base Bidirectional LSTM (with 1 layer of 128 units). These three were then used in experiment 4 and 5 where tables 4 and 5 summarizes the overall results.

From these results it is clear that the combined model of LSTM and Transformer outperforms all the other models in predicting CF Industries. This observation further led to the development of a basic trading agent that used the LSTM32+TX model. While its performance changed rapidly, the best, positive results (profitable P/L) ranged between 1.2 percent to 7.68 percent return on initial investment which is not bad compared to the S&P500 index 2024(Q1) returns of about 15 percent (Curvo, 2024); (Speights, 2024).

Table 4: Sample of Experiment 4 Results - S&P500 Index

Experiment 4: S&P500 Index			
Under 75 epochs			
Model	**MAPE**	**Run-Tim**	**No.Para**
LSTM32+TX	0.0149	949.04955s	301554
TX	0.0207	553.503s	17205
Bi-LSTM	0.01509	350.8049s	146225
Under 300 epcohs			
LSTM32+Tx	0.01964	986.7037s	301554
TX	0.02699	547.1845s	17205
Bi-LSTM	0.01334	543.042s	146225

Table 5: Experiment 5 - Predicting CF Industries

Model	Accuracy
LSTM32+Tx	95.9154%
TX	93.2079%
Bi-LSTM	91.6369%

Table 6: Experiment 5 - Trading CF Industries with LSTM32+TX Based Agent

Model-Agent	%Change
LSTM32+Tx	1.2-7.68

Therefore, with good risk management, a potential trader could get good returns by combining fundamental and technical analysis with an LSTM + Transformer model.

6 CONCLUSIONS

In this paper, LSTM, Bidirectional LSTM and Transformer models were compared in predicting five financial instruments. From an initial forecast, three best performing models were used to further predict S&P500 index and CF Industries based on large datasets. The best model was then tested as a trading agent model which yielded good results. This agent however had erratic results which is solid ground for future works, as one tries to stabilise the trading outcomes. In future works, one can try to stabilize the trading results by either adding more trading heuristic, modifying the base model or changing the specification of the trading environment.

REFERENCES

Bao, W., Yue, J., and Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7):e0180944.

Cao, J., Li, Z., and Li, J. (2019). Financial time series forecasting model based on ceemdan and lstm. *Physica A: Statistical mechanics and its applications*, 519:127–139.

Chen, Y., Pu, H., and Qu, Y. (2024). An analysis of attention mechanisms and its variance in transformer. *Applied and Computational Engineering*, 47(1):164–176.

Cristina, S. (2023). The Transformer Attention Mechanism - MachineLearningMastery.com. *MachineLearningMastery*.

Curvo (2024). S&P 500: historical performance from 1992 to 2024. [Online; accessed 9. Jul. 2024].

Fischer, T. and Krauss, C. (2018a). Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, 270(2):654–669.

Fischer, T. and Krauss, C. (2018b). Deep learning with long short-term memory networks for financial market predictions. *Eur. J. Oper. Res.*, 270(2):654–669.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jierula, A., Wang, S., Oh, T.-M., and Wang, P. (2021). Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *Applied Sciences*, 11(5):2314.

Kehinde, T., Chan, F. T., and Chung, S. H. (2023). Scientometric review and analysis of recent approaches to stock market forecasting: Two decades survey. *Expert Systems with Applications*, 213:119299.

Krishnapriya, C. and James, A. (2023). A survey on stock market prediction techniques. In *2023 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pages 1–6. IEEE.

Lin, Z. (2023). Comparative study of lstm and transformer for a-share stock price prediction. In *2023 2nd International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID 2023)*, pages 72–82. Atlantis Press.

Lui, Y.-H. and Mole, D. (1998). The use of fundamental and technical analyses by foreign exchange dealers: Hong kong evidence. *Journal of International Money and Finance*, 17(3):535–545.

Roondiwala, M., Patel, H., Varma, S., et al. (2017). Predicting stock prices using lstm. *International Journal of Science and Research (IJSR)*, 6(4):1754–1756.

Speights, K. (2024). The S&P 500 Jumped Nearly 15% in the First Half of 2024. Here's What History Says It Could Do in the Second Half of the Year. [Online; accessed 9. Jul. 2024].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Xiao, T. and Zhu, J. (2023). Introduction to Transformers: an NLP Perspective. *arXiv*.