**ECE 2372 FINAL PROJECT PROPOSAL**

**BY MANUEL GARCIA ACOSTA, PAPA FALL, VIVSWAN SHAH**

1. **Project Summary**

Stock markets in capitalist countries allow a wide array of individuals to invest their capital and take part in the economy’s growth. A stock market is often seen as one of the most important indicators of the economic strength of a country, since an increase in the price of stocks can be thought of as an increase in investment in companies of such country.

Because of the importance of the stock market, it is not surprising to notice that stock return predictability is one of the most important concerns for investors. In recent times “investors are exposed to an ever-increasing number of new facts, data and statistics every minute of the day. Assessing the predictability of stock returns requires formulating equity premium forecasts on the basis of large sets of conditioning information, but conventional statistical methods fail in such circumstances”. [1]

Machine learning models have gained popularity in recent times regarding the study of large-scale data sets, and the financial sector has not been impervious to this. Several ML models have been used to study stock returns like Abe and Nakayama [2], Nevalsami [3] and Rossi [1]. These models often use lagged (past) stock returns as features and the future stock return as the output values.

The main objective of our project is to predict daily stock returns of the US stock market and specifically a selection of 50 companies in the S&P (standardized patients) 500 index. We chose the S&P 500 index, which tracks the performance of 500 leading companies in the US economy, since it is one of the most used indices for tracking the performance of the US stock market.

“Directional prediction of stock returns is based on forecasting whether returns are greater than some pre-specified threshold. Previous research mainly focuses on sign prediction, where this threshold is equal to zero (i.e. whether the return is positive or negative)” [3]. For our study we will be using neural networks with data from the years 2014-2019 to train models that predict the next day’s return with the widely used threshold of 0. We do not consider the most recent years because of the large effect of the COVID-19 pandemic on the global economy.

The features we will use are the lagged entries of each stock. We’ll follow Nevasalmi methodology, whom states that lag lengths beyond ten trading days are found to be uninformative [3]. For the performance metric we will be using the accuracy of the classification (up & down) from each of our models.

1. **Project Description**

The stock market is the collection of buyers and sellers of stock or shares, which represent the ownership of a portion of a publicly traded business. By buying stocks or shares a person can invest their capital in the business and can take part in the business’s growth. As the business grows the value of the share of the company also grows, which in turn results in a profit or gain in investment by the buyers. Nevertheless, the opposite of this is also true, if the business loses value, then shares associated with the business also lose value. So, it is extremely important to invest your capital in the right business at the right time, to make the most return on your investments.

The main objective of our project is to predict daily stock returns of the US stock market and specifically a selection of 50 companies in the S&P 500 index. We chose the S&P 500 index, which tracks the performance of 500 leading companies in the US economy since it is one of the most used indices for tracking the performance of the US stock market. For our study, we will be using LSTMs (Long short-term memory) neural networks with sentiment analysis with pre-trained BERT (Bidirectional Encoder Representations from Transformers) using information from the years 2014-2018 to train models that predict the return for 2019. We do not consider the most recent years because of the large effect of the COVID-19 pandemic on the global economy.

There has been a lot of research work done on the predictions of stock with the neural network architecture as shown in Thakkar [4]. Based on this research, we are choosing LSTM for the prediction of stock data as this has been widely shown that LSTMs perform better with time-based data like the stock market data [5]. For sentiment analysis, we are choosing a pre-trained BERT network from Google.

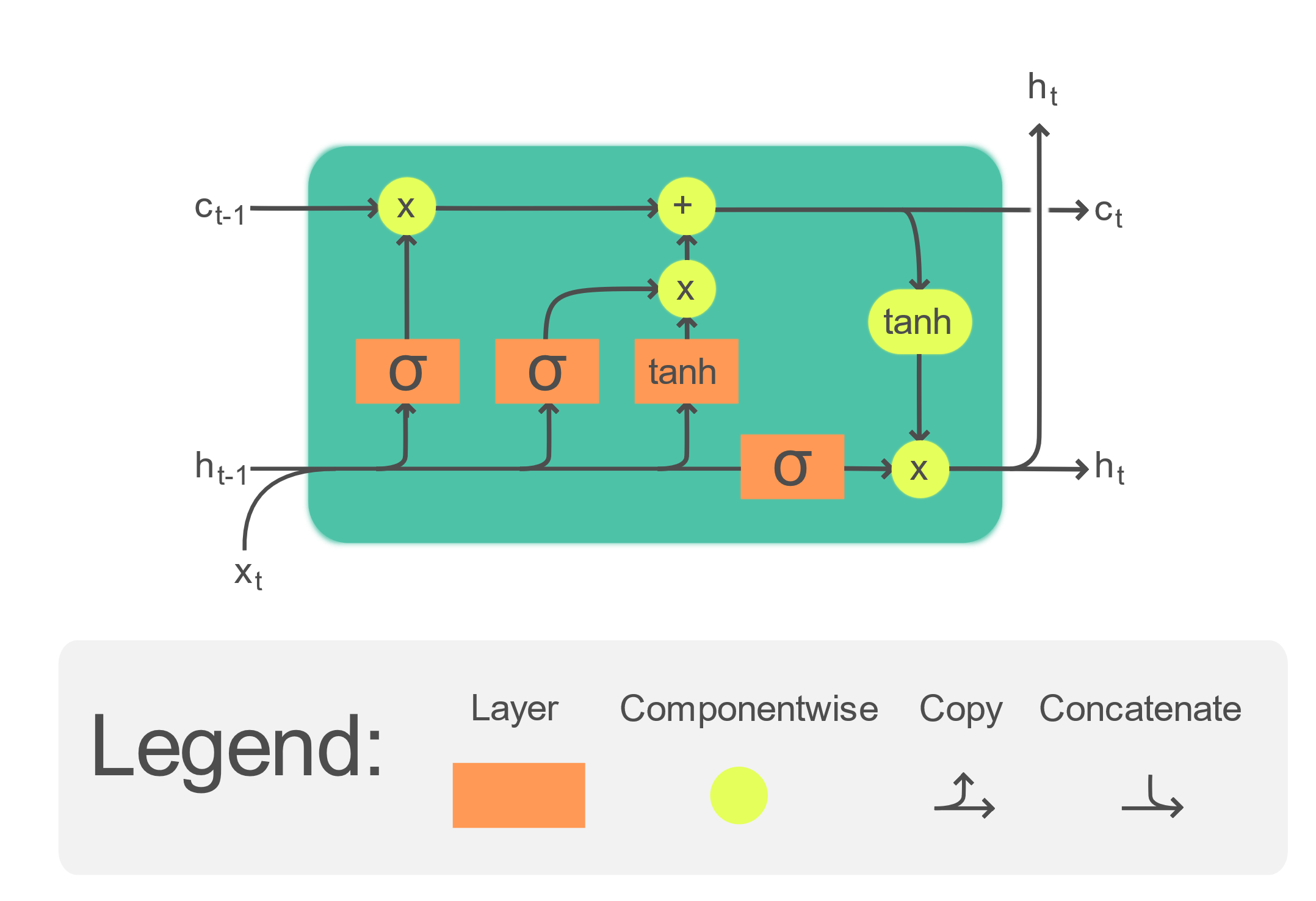
Long Short-Term Memory (LSTM) is designed based on Recurrent Neural Network (RNN) with feedback connections, which give it the ability to learn time-based data [6]. It can handle not only individual data points (such as photos) but also complete data streams (such as speech or video). For example, LSTM is applicable to tasks such as connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems).

Figure 1. Internal mechanics of The Long Short-Term Memory (LSTM). LSTM cell may process input in a sequential manner while maintaining its hidden state through time.

BERT stands for Bidirectional Encoder Representations from Transformers and is a cutting-edge neural network for natural language processing. BERT was created by a group of researchers at Google in 2018 [7]. BERT was trained on English Wikipedia (2,500 million words) and BooksCorpus (800 million words) and obtained the highest accuracies for various NLP (Natural Language Processing) tasks. There are two general BERT versions that have already been trained: The big model is a 24-layer, 1024-hidden, 16-heads, 340M parameter neural network, whereas the small model is a 12-layer, 768-hidden, 12-heads, 110M parameter neural network.

**Training process:**

The training of one epoch of the neural network will be as follows

Note: value of and are set to 1 at the beginning and dates are relative to starting of training dataset (from 2014 to 2018).

1. First, company from 50 companies is chosen
2. Then the date is chosen
3. Then news article related to company and the stock of company (sentiment data) and stock data of the company from date *j* is picked and normalized
4. Then sentiment data is fed through BERT to give a value between -1 (negative sentiment) to 1 (positive sentiment)
5. Then the combined data (stock data and sentiment value) is fed to LSTM to give next day predicted stock data
6. Based on this predicted stock data and real stock data from next day, gradients are computed, and network is updated
7. Repeat from step 2 with , if is in training dataset else continue
8. Repeat from step 1 with , if else training epoch completed

For testing, the same process is repeated except without updating the network in step 6 and giving the network stock data from the previous test results in step 3 (the testing dataset is the year 2019).

We are hoping to this combination of neural networks with data from 2014-2019 will perform better than or at least similarly to the previous research on this in both terms of accuracy and generalizability.

1. **Collaboration Plan**

To complete this project, we will need to fulfill several tasks chronologically. These tasks include data collection, coding, data analysis, and reporting. Our comprehensive plan contains a schedule for each task in order to complete our project by the deadline. We have delegated tasks amongst ourselves based on different experiences and skillsets that we have. Each task has an assigned leader that is the most qualified to coordinate the work with the other group members still helping. The first step of our project will be data collection.

* The task leader of data collection will be Manuel Alejandro. In this task, we will be gathering data from 50 different companies from the S&P 500 index. From this database, we will be able to access the performance of different stocks. The S&P 500 index will allow us the option to pick from 500 American companies leading in the US economy. We will also be using data from the years 2014-2019. It is important to get our data from prior to the year 2020 because in that year stock behaviors were skewed due to the COVID-19 pandemic. This overall task is important because we will not be able to build our algorithm without a dataset. A potential challenge could be us not finding data for a particular company of interest. We have set a deadline for the completion for this task for March 30th. The next task is coding.
* The task leaders for coding will be Vivswan Shah and Papa Fall. Through this section of our project, we will use machine learning models. This is to learn from the data collected to train a model based on the years 2014-2018 and then test the model using the data from 2019. Our plan is to use a neural network which can learn complex functions. Neural networks are also able to perform multiple stages of processing to predict a response. Building this algorithm is important, because we believe it will be able to obtain accurate results despite how complicated the data might be. However, there are challenges involved with this potential learning model. For example, figuring out the correct model, normalizing the data, and making sure there are no biases during training. Our set deadline for completion of this task is April 13th. The next task would be data analysis.
* The task leaders for data analysis will be Vivswan Shah and Manuel Alejandro. Once we obtain results from the neural network model, the next step will be analyzing the results. This is a crucial step, because during analysis it is when we will be able to judge the accuracy of our machine learning model. Once our model can predict outputs for stocks from 2019, we will be able to display an accuracy metric based on a comparison between our predictions and the actual outputs. There are some challenges that our model might face that might affect the analysis. One is making sure that our model accurately represents a time-based analysis. Another challenge is making sure that our model is not overfitting and is generalizing properly. Our set deadline for this task is April 18th. The final task in finishing our project is documenting our findings.
* The task leader for this portion of the project will be Papa Fall. This portion of the project is important because documenting our research results will help simplify the totality of our work. Which is critical since it will help outside readers be able to comprehend the project summary quickly. A potential challenge with reporting our findings would be deciding the most valuable information that needs to be documented. Our deadline for this task is April 23rd and it will be the final portion of this project.

References

[1] A. G. Rossi, “Predicting stock market returns with machine learning,” *Georgetown University*, 2018.

[2] M. Abe and H. Nakayama, “Deep Learning for Forecasting Stock Returns in the Cross-Section,” in *Advances in Knowledge Discovery and Data Mining*, vol. 10937, D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, Eds. Cham: Springer International Publishing, 2018, pp. 273–284. doi: 10.1007/978-3-319-93034-3\_22.

[3] L. Nevasalmi, “Forecasting multinomial stock returns using machine learning methods,” *The Journal of Finance and Data Science*, vol. 6, pp. 86–106, Nov. 2020, doi: 10.1016/j.jfds.2020.09.001.

[4] A. Thakkar and K. Chaudhari, “A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions,” *Expert Systems with Applications*, vol. 177, p. 114800, Sep. 2021, doi: 10.1016/j.eswa.2021.114800.

[5] A. Moghar and M. Hamiche, “Stock Market Prediction Using LSTM Recurrent Neural Network,” *Procedia Computer Science*, vol. 170, pp. 1168–1173, 2020, doi: 10.1016/j.procs.2020.03.049.

[6] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “The Performance of LSTM and BiLSTM in Forecasting Time Series,” in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, Dec. 2019, pp. 3285–3292. doi: 10.1109/BigData47090.2019.9005997.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv:1810.04805 [cs]*, May 2019, Accessed: Mar. 21, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805