Stock Prediction through LSTM: A Holistic Approach

Final Report

James Hu
MIS 464
University of Arizona
United States of America
jameshu@email.arizona.edu

Jingsong Wu
MIS 464
University of Arizona
United States of America
jingsongwu@email.arizona.edu

***Abstract*— Stock price prediction is always one of the most popular and important topics to investors. With the expansion of the information source, more and more information is proved that it has a correlation with the stock price. The current existing approaches are still far away from meeting investor's expectations. The biggest challenge is that the large volumes of information from both company's internal or external sources are very complex. Thus finding the causal relationships between them and a company's stock prices is extremely hard and has been the topic of many academic studies. In this work, we conducted a thorough survey of the current academia landscape on stock movement, stock prediction and our intended methodology: Long Short-Term Memory (LSTM) networks. Through this, we identified a key research gap. While there have been many studies utilizing LSTMs to analyze historical stock prices and news, no study has expanded beyond those two data sources. In light of this, we conclude that the best approach to stock prediction is using a holistic data structure encompassing a wide range of data pertaining to one specific company whilst using LSTM as our prediction model. The resulting framework demonstrated its ability to learn and predict future stock prices and massively outperforms several renowned regression benchmarks. In the end, we propose several promising future directions this study can be taken towards.**

Keywords—stock prediction, long short-term memory, deep learning, transfer learning, regression, neural networks, J48, random forest, multilayer perceptron.

I. INTRODUCTION

The daily fluctuations of the financial market now represents massive amounts of cash flowing between investors and companies. As a result, the good investment decision for all participants in the stock market has become reliant on these daily fluctuations. To address this, many computer algorithms are increasingly used to predict future stock prices. Since the stock market has a total value over 30 trillion USD, an accurate stock prediction algorithm is highly valuable. However, most algorithms underperform when applied to real-life stock movements. This is often because of the volatile nature of stock prices. Many researches have been done on predicting future stock prices using either news, or past stock prices. However, the nature of stock prices movement is affected by multiple factors which make it too complex to be modeled using only two, relatively straightforward sources. To address the inherent chaotic nature of the stock market, we propose the RNN architecture, a neural network architecture used to automatically extract features from time-dependent sequences. Specifically, long short-term memory networks, an evolution of the basic RNN, have shown promise in analyzing long sequences of data, perfect for stock prediction. Thus, we propose a prediction framework that utilizes multiple different datasets, excluding news data, to predict a single stock price over time using a Long Short-term Memory prediction model.

II. LITERATURE REVIEW

For this project, we focused on three areas of the research: stock market movement, stock market prediction and long short term memory. First of all, we want to understand why the stock market moves in such a way. While reading the research papers, we tried to find the specific metrics that can significantly increase the stock price. Many macroeconomics metrics fit in those research. Besides that, each company has some unique metrics to measure its stock performance. We built our model based on these metrics we found. The Long Short-Term Memory tools help us efficiently learn the long-term patterns.

Stock Market Movement

A company's stock price reflects the market's prediction of its future growth and operation performance. Predicting the growth of the stock market has historically been a difficult task due to the chaotic nature of the operating output. Because of this, several different methods have emerged to tackle this challenge. Although each stock in the stock market moves individually, scientists have found that the overall stock market performance might follow a certain pattern. Many literature focus on using the historical stock data to find the pattern. These projects will often employ machine learning (ML) techniques or statistical analysis to draw conclusions from their dataset. Table 1 shows a compilation of related papers that lists their data sources.

Year	Author	Data Source
2018	Das et. al.	US Stock Prices, Twitter
2018	Li et. al.	US Stock Prices, New York Times, WSJ
2017	Chong et. al.	Korean Stock Prices
2017	Singh and Srivastava	US Stock Price
2017	Wang et. al.	Hong Kong Stock Price
2016	Chen et. al.	Shanghai Stock Exchange
2016	Moghaddam et. al.	US Stock Prices

Table 1. Stock Movements and their Data Sources

Stock Prediction

To accurately predict the stock price, it is necessary to find and understand the important metrics that can directly or indirectly affect the stock price. One important thing in stock prediction is that the metrics should be predictable and relevant. Several literature has addressed the many external factors that could heavily impact the stock price such as unemployment rate, oil price, foreign exchange rate, federal interest rate and inflation rate. Majority of the data can be easily accessed from Bloomberg, Yahoo Finance and Bureau, but the relations of each variable to the stock price is very complex. The figure below are the most frequent words from our research literature.

Figure 1. Important Metrics to Stock Price

Although most of these external factors are uncontrollable, these factors could be predicted from the source of the government policies and financial news. Using these sources could help the investors to predict the potential changes of these factors and use these predicted changes in each factor to predict the stock change. The table below shows the recent research on different test objectives on how these objects affect the stock market.

Year	Author	Test Objective	Data Source
2016	Jareño, F., & Negrut, L.	Unemployment Rate	US Stock Prices
2015	Sim, N., & Zhou, H.	Oil Price	US Stock Prices
2014	Zhang, T.	Interest rate	US Stock Prices
2014	Andries et. al.	Interest rate, Exchange rate	US Stock Prices
2013	Chang, K., & Yu, S.	Oil Price	US Stock Prices
2012	Mun, K.	Inflation rate, Exchange rate	US Stock Prices, Japanese Stock Prices
2011	Hamrita, M., & Trifi, A.	Interest rate, Exchange rate	US Stock Prices

Table 2. Stock Prediction and Metrics

Long Short-Term Memory

In order to analyze the data, we need a data mining tool to help us analyze those data efficiently. Also, besides using the metrics we found in the paper mentioned above, another

direction is taking into account the news stream in predicting stock prices. A major part of new stream analysis utilizes sentimental learning, a machine learning technique that categorizes news and other media based on their opinion towards a certain subject, in this case the company or industry in question. Each path has suffered drawbacks, with each side seeing considerable progress which means that we need to analyze the data from a huge database to make our conclusion.

Increasingly, neural networks (NN) and its many forms have become a widespread tool for scientists to analyze and draw conclusions from huge amounts of data. Traditional NN models are not suitable for time-dependent data, leading to the development of the recurrent neural network (RNN) model, which passes information down as it processes a sequential input. As iterates through a time dependent sequence, it will pass information along, allowing them to learn time-dependent patterns.Figure 2 shows the basic architecture of an RNN.

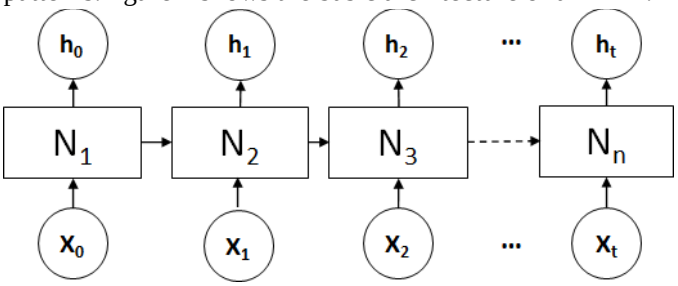

Figure 2. Illustration of the architecture of a basic RNN

However, RNNs are inefficient in learning long-term patterns due to the vanishing gradient problem. This is solved through the long short-term memory (LSTM) models are a type of recurrent neural network with an additional memory unit to retain information as it parses long time-dependent data. Advancements in ML models, especially the introduction of LSTMs, have seen a huge improvement in accuracy for models relying only on stock data. Table 3 shows a compilation of related papers that lists their data sources.

Year	Author	Method	Data Source
2017	Nelson et. al.	LSTM	US Stock Prices
2017	Selvin et. al.	LSTM, RNN, CNN	US Stock Prices
2016	Akita et. al.	LSTM	Korean News and Stock Prices

Table 3. Literature about Long Short -Term Memory

III. RESEARCH GAPS

From our evaluation of related stock market prediction literature, we observe that most papers focus on only two or less data sources: historical stock price and online news media. However, from our previous findings about stock market movement, we know that these two data sources alone cannot accurately model stock market movements. Also, most research papers focused on only one or two variables of the stock market. However, the stock market movement is a complex result from the combinations of multiple factors. Some factors are related to other factors. For example, the unemployment rate could affect the interest rate. This is the most challenging part of our research. Because of this, we believe that in order to accurately predict future stock prices for a specific company, we need to incorporate many sources into our model.

Although some research papers mentioned the impacts of some factors are different from stock to stock due to their different industry, none of them actually went deep about the actual impacts of each factor in each industry. For example, the inflation rate changes might affect more significantly on the stocks in the Real Estate industry due to its high proportion of debt and loans in the balance sheet and less significantly on the stocks in the Retail industry which focused more on the liquidity side. It takes us more time to figure out the impact of each metric in each industry.

With these gaps in mind, we propose a LSTM-based prediction framework that utilizes multiple different datasets, excluding news data, to predict a single stock price over time.

IV. DATA COLLECTION

Before we get started, we need to collect the data for our regression model. Since most of the data we need are also important metrics for the macroeconomy and the money market, it is relatively easier for us to collect those data. Our prediction model includes 5 variables: date, effective federal funds rate, real GDP (in percent change), unemployment rate, and S&P 500 index.

All the data about federal funds rate are collected from the federal reserve, which is the central banking system of the United States of America. The effective federal interest rate is the target interest rate that the banks and other depository institutions borrow and lend their reserves to each other. Theoretically, the stock price should be negatively correlated to the federal interest rate. As the federal interest rate increases, the cost of borrowing increases and the present value of the future dividend payment decreases and will eventually decrease the stock price. Based on our research, we found that the relationship between the stock price and interest rates is significant only at low frequencies or longer horizons (4 to 5 years). Therefore, the interest rate is a good metric to predict the stock price for long term investment. However, when we do a deeper research on the relationship between federal funds rate and the stock price, we found that the impact of interest rate on stock price varies from industry to industry depending on the capital structure of the industry.

The real GDP growth rates are collected from the Bureau of Economic Analysis, which is a government agency that provides official macroeconomic data and industry statistics. Real GDP growth is an important metric to measure the overall economic growth. Normally, if the real GDP growth increased higher than expected, the stock price would increase.

We collect the unemployment rate from the Bureau of Labor Statistics which is a government department focusing on labor economies. Unemployment rate is a key metric of the economic performance of the market. By monitoring the unemployment rate, researchers are able to capture the economic cycle of the market. A low unemployment rate usually reflects strong economic growth with high GDP, CPI and IPI. A high unemployment rate usually reflects a recession of the market with low consumer spending. Meanwhile, the changes in the unemployment rate also affect the interest rate. When the unemployment rate increases, the federal government trends to lower the interest rate and when the unemployment rate is low, the federal government trends to increase the interest rate.

We collected the S&P 500 index from Yahoo Finance. It is an index that measures the top 500 stocks traded in the U.S. stock market. Since most of our data sources are from government agencies, those data are transparent to all visitors and accurately recorded.

V. RESEARCH DESIGN

Our proposed framework consists of two parts, the data preprocessing part and the LSTM model.

Data Preprocessing

The data preprocessing portion of our framework is what allows the entire framework to parse a multitude of different datasets. The portion works in three steps. The first is to aggregate all training data into a singular csv file. While this step can be done by a Python function, we did this step manually. To accomplish this, we linked several CSV files together based on the date and deleted any rows without a stock price datapoint, the datapoint we are trying to predict. Then, we fill in the blanks for the other data, like GDP the federal reserve rate, using previous numbers. Afterwards, we end up with a completed CSV file with a set of numbers corresponding with each stock price datapoint. From here on, everything will be done using a Python script. Step two is to split the aggregate into training and testing data, the former for training our LSTM model and the latter for model evaluation. For this project, we used an 80/20 train/test split. The third is to format our testing and training data for the LSTm model. This involves normalizing the data and converting the data into the correct dimension, both of which is accomplished using the Python numpy library.

LSTM Model

With the data preprocessed, we can finally start using the LSTM model for analytics, which is coded entirely in Python and utilizes the CUDA-enabled Pytorch library heavily. As with its previous mention in the literature review, the LSTM is a type of neural network specially designed to process long sequences of time-dependent data. It does this by automatically extrapolating features from its input sequence, many of which are time-dependent patterns needed for stock prediction. Then, using these extracted features, it will continue attempting to predict future prices over a set amount of iterations, or epochs, each time making an error in its predicted stock price when compared to the actual stock prices, then correcting its algorithm based on these errors through backpropagation. At the end of training, the LSTM should be left with a set of weights that can accurately predict future stock prices given input. This set of weights will be tested on the test data to evaluate how well it can predict.

The following table provides the LSTM's specifications. These specifications were hardcoded due to either the nature of the data or limited hyperparameter tuning.

Attribute Name	Data Source
Train/Test Split	80/20
Training Window	12
Learning Rate	0.0001
Training Epoch Limit	150
Input Layer Size	7
Hidden Layer Size	50
Output Layer Size	1

Table 4. LSTM Specifications

VI. EXPERIMENTAL RESULTS

Throughout this project, we have made several experiments and models to predict the stock price. From our experiments, we gathered three major experimental results. We used the Tableau to help predict the stock price of a company with the price of the company's key product and related materials. The experiment results from that model didn't meet our expectation because it is too company specific. As mentioned above, the company's stock price reacts on some factors differently than other companies because of customers' elasticity of demand on their products and the company's capital structure.

Then, we used the Weka data mining tools to build several models based on the metrics we collected from the company's financial report, the government reports and financial websites. We used a linear regression model, J48 model, random forest model and multiple perceptron model to predict the stock price based on the same dataset. We found that linear regression was the most accurate model with higher correlation coefficient and lower mean absolute error. However, we still think the linear regression model we built is not accurate enough because of its high mean absolute error. Because linear regression can only work on numeric data, we decided to implement the LSTM model to help us analyze the news and add those data into our model.

With the LSTM model, we can analyze the data from news reports in our prediction. Our LSTM model predicted the result with the data from multiple news sources. Based on the results from our LETM model, our prediction on stock price is more accurate than previous models. Besides that, we also found three exciting findings. The first is evidence that our model is capable of learning from the training data. The second is evidence that our model has avoided overtraining. The third is evaluation against other renowned regression benchmarks.

Evidence that Model can Learn

Before we can trust any analytics from the LSTM model, we must first confirm that the LSTM is capable of automatically extracting features from the data. This is actually a fundamental question when using any neural network model: does the model actually learn anything from the data or is it just making random predictions. We can determine this by looking at the model's loss function, which gives us a representation of the difference between the model's predicted prices and the actual prices. If the LSTM is indeed learning something from the model, the loss function's value, or the loss value, the error should decrease over the training epochs. To illustrate this, we plotted the loss value as the LSTM trains over 150 epochs in the following figure.

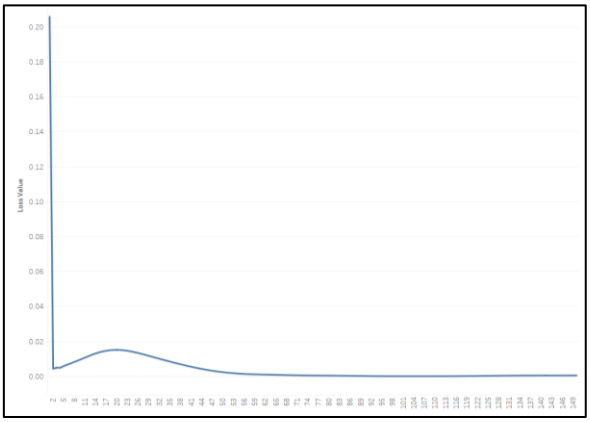

Figure 3. Loss value as each training epoch passes

As clearly shown, in figure 2, as the time passes, the loss value generally decreases over the training epochs, indicating that the LSTM is indeed learning something from the data.

Evidence that Model Avoided Overtraining

However, long tail present in the loss value plot may indicate another potential issue, overtraining. This occurs when a NN is trained so much that it memorizes the training data,

resulting in a network that can't generalize its predictions to data it hasn't seen, which essentially means it's useless. To see if overtraining is present, we also plotted the trained LSTM's predicted stock prices on the test data with the actual test data's stock prices to see if the model can successfully generalize to unseen data. The following figure shows the plot, with blue representing the actual prices and orange representing predicted prices.

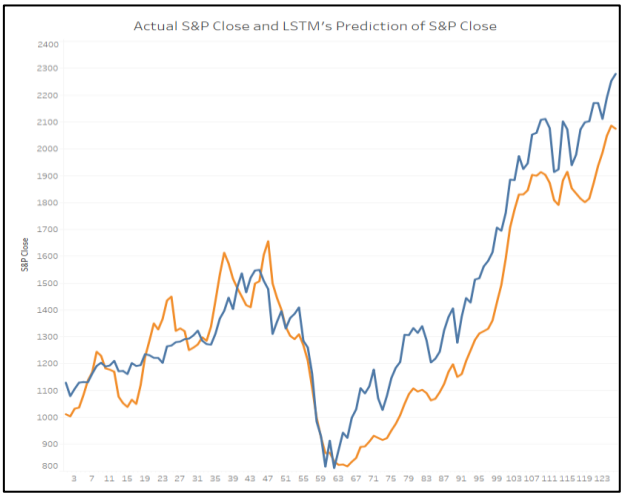

Figure 4. Graph of Actual and Predicted S&P 500 Close Data

From figure 4, we can see that while the result from the model is underperformed after the recession period. The stock price predicted is lower than the actual stock price. Although our result from experiments is not capable of predicting future stock prices with pinpoint accuracy, it is still able to learn the general trends of the data. This means that the data has avoided overfitting and can successfully generalize to unseen data. On the other hand, this model could help investors to avoid the loss from stock investments because it used to underestimate the value of stock and provide lower price. If the investors use this model and receive a lower stock price than they expected, they would sell the stock and avoid the loss during a recession period.

Evidence that Model Avoided Overtraining

For our last experimental result, we wanted to see just how well our model performs when compared to other methods. Because we are working with future value prediction, we chose several regression benchmarks from WEKA to compare with our LSTM framework. The methods we chose are linear regression, gaussian process, random forest, and a multilayer perceptron. The actual architecture of the multilayer perceptron is the same architecture as the best performing multilayer perceptron from our Lab 2 WEKA assignment in order to compare our framework with the best method out there. Table 5 shows the correlation coefficient and the mean absolute error of each regression benchmark and our framework.

Benchmark	Correlation Coefficient	Mean Absolute Error
Linear Regression	0.8654	262.5585
Gaussian Process	0.7618	392.1461
Random Forest	0.8688	392.0959
Multilayer Perceptron	0.6435	708.303
Our LSTM Framework	**0.9638**	**130.892**

Table 5. Benchmark Performance Comparison

From our comparison, we can see that our framework outperforms every benchmark with regards to both statistics. For a better interpretation, the correlation coefficient shows our model being 96.38% of the way to being able to perfectly predict stock prices compared to the 86.54% for linear regression and 86.88% for random forest. This result shows that our LSTM model can help us calculate the correlation between each factor and stock prices. The mean absolute error shows our model's prediction being off by $130.89 on average, which is significantly lower than any other methods we used so far. This surprising result shows that our LSTM model can predict the stock price more accurately given the high dollar value of the S&P 500 index. Our results from the LSTM model show the power behind RNN-based NN algorithms in predicting time-dependent data, especially the LSTM, which is designed for long sequences such as decades of stock price data. Thus, they naturally outperform other regression methods that might struggle to model such chaotic data over long periods of time.

VII. CONCLUSION

In this study, we addressed a surprising gap in the academic literature, that being the lack of using alternative datasets to conduct stock prediction. From our literature review in stock

movement, we know that stock price fluctuation is not just a result of past stock prices and news. Beyond that, there are a complex combination of factors. Those factors might also have correlation between each other. Depending on the elasticity of demand and capital structure, the impact of factors to each company in different industries are different. In the short run, the stock performance is also related to the news. Our previous results from Tableau works, linear regression model, J48 model, random forest model and multiple perceptron models did not meet our expectation on accuracy of the predicting results. None of these models include any categorical data in it and we believe that's the reason why we could not predict the stock price accurately. Thus, we built an LSTM framework that forgoes news data in order to prove the viability of using alternative data sources in predicting stock data. Through our experimental result, we can conclude that our model is capable of automatically extracting time dependent features to then learning and successfully predicting future stock prices based on data seen and unseen. When evaluated against other regression benchmarks, we see our framework vastly outperforming other methodology. By comparison with the real S&P 500 index, we found that our predictions from our LSTM model underestimate the stock price after the last economic recession. We believe that might be because of investor's investing strategies on the stock portfolio have changed from the last recession and the weight of bonds in the portfolio changed. These might be the cause of money supply in the stock market and therefore impact the stock price. More research and experiment is needed to prove this.

From our work, we can see the power of the LSTM when used in stock prediction, which RNN-based models seem to be almost specifically made for. However, the bigger picture is how we conducted our predictions. We currently have less factors listed in our LSTM model than our previous regression models, this might be a potential reason that our predicted stock price is lower than the actual price. Rather than just focusing on the news, which still has its impact on the stock market movement, we chose to explore different avenues of data. With this, we demonstrated the viability of using non-news related data in stock prediction and showed promise in this alternative direction of approaching stock prediction.

VIII. FUTURE WORK

Based on our results from the experiments, we realized that there is more research we need to do in the future to accurately predict the stock price. We are now considering three directions to increase the accuracy of our prediction. Each of the three future directions envisioned revolves around further increasing the accuracy of our prediction.

First, although we have collected a lot of data for our model, it seems we still need to do more research and gather more data. Since the stock price index is influenced by many variables, we need to keep researching and collecting different variables into our model to decrease the standard error of our regression model. One interesting topic we can study is the impact of the global crisis to the stock market. Starting from March, we saw the stock market negatively impacted by the coronavirus. And we believe that the reason why the stock performs in this way is due to the news reports of the disappointing results and increasing cases and deaths. Because of this, we suggest also incorporating news data into our data feed. While we actively avoided news data in this study to prove the viability of using non-news data in stock prediction, the news is still a major player in stock market movement, and incorporating sentimental analysis on news data may further enhance our predictions.

The second direction is to conduct further hyperparameter tuning to optimize our LSTM's performance. Because of our limited time, we were unable to explore every possible parameter combination four our LSTM model. One immediate area to hyperparameter tuning is our training epoch number. From the loss value graph, we can see a super long tail, and while it fortunately did not result in our model overtraining, it still represents a considerable amount of unnecessary time spent training. By cutting down on training time, we can speed up the entire framework, thus being more efficient with our time.

The last direction is exploring alternative RNN-based algorithms. While LSTM was specially developed to deal with long-sequenced data, it is relatively old, there have been even more recent developments. One such development is the Gated Recurrent Unit, an further evolution of the LSTM that can deal with long sequences even more efficiently than the LSTM. Another development is the much more recently proposed Independent RNN, or IndRNN, which seems to outperform both LSTMs and GRUs. However, because we were constrained by our limited time and knowledge of the field, we

were unable to fully explore these exciting new fields, but the mere existence of these possibility hints towards the potential of this project.

ACKNOWLEDGMENT

This work is supported by the University of Arizona under Eller College of Management and Management Information System department.

Reference

[1] Akita, R., Yoshihara, A., Matsubara, T., Uehara, K. (2016). Deep Learning for Stock Prediction Using Numerical and Textual Information. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, 2016, pp. 1-6

[2] Andries, A., Ihnatov, I., & Tiwari, A. (2014). Analyzing time-frequency relationship between interest rate, stock price and exchange rate through continuous wavelet. Economic Modelling, 41, 227-238.

[3] Bengio Y., Simard P, and Frasconi P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, Vol. 5, No. 2 (March 1994), 157–166.

[4] Chang, K., & Yu, S. (2013). Does crude oil price play an important role in explaining stock return behavior? Energy Economics, 39, 159-168.

[5] Chen, W., Hao, Z., Cai, R., Zhang, X., Hu, Y., & Liu, M. (2016). Multiple-cause discovery combined with structure learning for high-dimensional discrete data and application to stock prediction. Soft Computing, 20(11), 4575-4588.

[6] Chong, Eunsuk & Han, Chulwoo & Park, Frank. (2017). Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies. Expert Systems with Applications. 83.

[7] Chung, J., Gulcehre, C., Cho, K., Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. ArXiv:1412.3555

[8] Das, S., Behera, R., Kumar, M., & Rath, S. (2018). Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction. Procedia Computer Science, 132, 956-964.

[9] Hamrita, M., & Trifi, A. (2011). The Relationship between Interest Rate, Exchange Rate and Stock Price: A Wavelet Analysis. International Journal of Economics and Financial Issues, 1(4), 220-228

[10] Jareño, F., & Negrut, L. (2016). US Stock Market And Macroeconomic Factors. Journal of Applied Business Research, 32(1), 325-340.

[11] Li, X., Xie, H., Lau, R., Wong, T., & Wang, F. (2018). Stock Prediction via Sentimental Transfer Learning. IEEE Access, 6, 73110-73118.

[12] Moghaddam, A., Moghaddam, M., Esfandyari, M. (2016). Stock market index prediction using artificial neural network. Journal of Economics, Finance and Administrative Science, Volume 21, Issue 41, December 2016, Pages 89-93

[13] Mun, K. (2012). The joint response of stock and foreign exchange markets to macroeconomic surprises: Using US and Japanese data. Journal Of Banking & Finance, 36(2), 383-394.

[14] Nelson, D., Pereira, A., Oliveira, R., (2017). Stock market's price movement prediction with LSTM neural networks, 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 1419-1426.

[15] Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V., Soman, K. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model, 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, 2017, pp. 1643-1647.

[16] Sim, N., & Zhou, H. (2015). Oil prices, US stock return, and the dependence between their quantiles. Journal of Banking and Finance, 55, 1-8.

[17] Singh, R., & Srivastava, S. (2017). Stock prediction using deep learning. Multimedia Tools and Applications, 76(18), 18569-18584.

[18] Wang, Feng, Zhang, Yongquan, Rao, Qi, Li, Kangshun, & Zhang, Hao. (2017). Exploring mutual information-based sentimental analysis with kernel-based extreme learning machine for stock prediction. Soft Computing, 21(12), 3193-3205.

[19] Zhang, T. (2014). Stock Price, Real Riskless Interest Rate and Learning. IDEAS Working Paper Series from RePEc, IDEAS Working Paper Series from RePEc, 2014