

# Applying Recurrent Neural Networks with Long Short-Term Memory in Clustered Stocks

Felipe Affonso, Thiago Magela Rodrigues Dias

Centro Federal de Educação Tecnológica de Minas Gerais(CEFET/MG)  
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

***Abstract.** Predicting the stock market is a widely studied field, either due to the curiosity in finding an explanation for the behavior of financial assets or for financial purposes. Among these studies the best techniques use neural networks as a prediction tool. More specifically, the best networks for this purpose are called recurrent neural networks (RNN), and provide an extra option when dealing with a sequence of values. However, a great part of the studies is intended to predict the result of few stocks, therefore, this work aims to predict the behavior of a large number of stocks. For this, similar stocks were grouped based on their correlation and later the algorithm K-means was applied so that similar groups were clustered. After this process, the Long Short-Term Memory (LSTM) - a type of RNN - was used in order to predict the price of a certain group of assets. Results showed that clustering stocks did not influence the effectiveness of the network and that investors and portfolio managers can use it to simplify their daily tasks.*

*Keywords: Neural Networks, Clustering, Stock Market, Deep Learning.*

## 1. Introduction

Predicting stock market movement is a well-known research field. Some studies argue that it is possible to predict stock market movement [Filho and Júnior (2009); Zhang et al. (2018)]. Based on that, it is natural to use all available resources in our dispose in order to confirm these hypotheses. The latest discoveries on the artificial intelligence (AI) area shows that algorithms are able to learn by themselves, recognize patterns, find the best features for a model and many others [LeCun et al. (2015)]. By using the deep learning (an AI subarea) methods it is possible to understand financial market better than just reading and looking the data by itself.

Deep Neural Networks have been used several times in order to predict the financial market movement. Studies show that this method is getting 48%-54% right predictions [Gerlein et al. (2016)]. Thus, we can certify that this field of research can be improved. Overall, the key advantage of deep learning is the ability for feature abstraction and for detecting highly complex interactions between these features – resulting in state-of-the-art performance across many applications [Krauss et al. (2017)]. Recurrent Neural Network (RNN) is a type of neural network which offers the best outcomes when dealing with sequential data, for example, stock returns.

However, using neural networks in order to predict the stock market is a complex and expensive task. It is necessary to prepare the data, prepare the network, adjust the

parameters, evaluation metrics and others. Therefore, the main purpose of this study is verifying if it is possible to forecast the movement of a group of similar stocks. If so, investors and portfolio managers will be able to conduct their investments in a much simpler and less expansive environment.

In order to predict the movement of a group of stocks, they must be previously clustered together. There are several recent studies proposing different ways of clustering stocks [Bini and Mathew (2015); Jung and Chang (2016); Momeni et al. (2015); Nanda et al. (2010)]. Most of them use more than one technique and focus on portfolio management. In this paper, the K-Means algorithm will be used once it outperforms other methods in terms of clusters compactness [Bini and Mathew (2015); Nanda et al. (2010)].

The remainder of this paper is organized as follows. Section 2 briefly covers some previous studies and works in this area. Section 3 provides an in-depth discussion of the methodology, explaining which techniques were used in each part of the process. In Section 4 the results are presented. Finally, Section 5 presents some discussions about most relevant findings and proposes future works.

## **2. Literature Review**

Långkvist et al. (2014) explains the importance of time-series in data mining and how it can be used as a feature to perform better tasks in AI. By adopting unsupervised training methods, it is possible to use the time-series as an important feature of the model and consequently achieve better results. They also explain that dealing with time-series can be tough, the data is noisy and contains several dimensions. When techniques to simplify the data are applied, important information can be lost. Different algorithms are shown and explained, each one of them have their own characteristics and can be used in a different way. Focusing on the stock market deep learning strategies can solve most of the problems on dealing with such a complicated type of data and provides new approaches that have not been tested yet. The author also explains how time-series are used in different fields as music recognition, videos, physiological data, and others.

Bini and Mathew (2015) present some data mining techniques used for clustering and data prediction. After explaining the importance of each categories of data mining the paper goal is presented. The work focuses in finding out the best companies in the market using clustering techniques and predicting their future stock price using regression techniques. Among the methods used, the K-means and Expectation Maximization algorithms performed better when compared to hierarchical and density based technique. Whereas the prediction algorithm used was the multiple linear regression that is best than simple linear regression.

Kumar and Ravi (2016) describes the past studies and surveys on the text mining field including several applications of these techniques. They also present the advantages and disadvantages of the methods used on these researches. The paper consists in a literature survey on different databases about the text mining subject applied to the financial domain. In order to transform a normal text into a structured format several techniques can be applied. The first step is called pre-process, it plays an important role in text mining tasks. Normally, when the pre-process step was done with excellence the results tend to be superior. Feature selection is the second step to be done, in this part the unnecessary features are removed and the data set retains only the important ones. The authors present several important information about the number of papers and the year of

publication of each one, a distribution of methods and approaches, and also a summary containing the most relevant information about the studies.

Cavalcante et al. (2016) proposes a new literature review on recent approaches designed to solve financial market problems. The main purpose of the study is to survey the machine learning methods applied to the financial context published from 2009 to 2016. Besides summarizing the main studies, the authors also identify the key challenges on this research field. The articles are categorized by their pre-processing, forecasting and text mining methods. Some concepts are detailed through the article, as the difference between technical and fundamental analysis, and also the machine learning methods: clustering, artificial neural networks, support vector machines, and others. The authors also explain the advantages of using these techniques in financial data and how they can be used in order to achieve better results. One of the most important results of this paper is the summary presented, where it is possible to verify features as: main goal, application, which inputs the authors are using, the techniques and if they used some trading system or not. It represents an important research once it presents the key publications and the techniques used in several relevant studies in the past years.

Chong et al. (2017) aims to use deep learning techniques for financial market prediction for 38 stocks from the Korean stock market. The authors compare three deep neural networks methods against a standard regressive model and an artificial neural network. The authors utilize the Normalized Mean Squared Error (NMSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mutual Information (MI) as measures to evaluate the prediction performance the methods utilized. As a result, the deep neural network performs better on the training set, but on the test set the autoregressive model performs better. Both of the methods can be joined in order to combine the advantages of each one and eliminate the disadvantages. Also, it is necessary to use more data aiming high quality outcomes.

The research on the finance and neural networks field is motivated because just a few articles on the area can be found. Krauss et al. (2017) writes that "This point can be illustrated with The Journal of Finance, one of the leading academic journals in that field. A search for "neural networks" only produces 17 references whereas the journal has published about two thousand papers during the last thirty years. An even more limited number of papers uses neural network techniques in their empirical studies." According to the paper, this is a unique study because they use three different state-of-the-art machine learning techniques and compare them with a deep learning algorithm. Also, they follow the financial literature and provide a holistic performance evaluation, making it easy for investors to understand the results achieved. And lastly, they focus on a daily investment horizon instead of months and weeks, as normally on the literature. By using daily values the model can be trained with more data and consequently performs better. The main goal of the paper is to bridge the gap between academic and professional finance, providing a new perspective for both fields equally.

Fischer and Krauss (2018) proposes a new method for financial market predictions using deep learning with long short-term memory (LSTM) networks. This work outperforms Krauss et al. (2017) and achieves returns of 0.46 percent per day. The main goals of the study are: use LSTM networks on the financial area, provide a better understanding about artificial neural networks and summarize the strategies used by the LSTM in choosing between winning and losing stocks.

Nelson et al. (2017) studies the applicability of recurrent neural networks (RNN), in this case the LSTM was used in order to predict stock prices movement for four Brazilian companies. It was possible to observe that this technique outperforms the baselines with few exceptions. It proves that LSTM networks can offer good predictions when compared to other approaches, mainly in terms of accuracy and gains obtained. Even when compared to other finance strategies, like buy and hold, the trained model outperforms traditional methods and offers less risks when observing the maximum losses.

Zhang et al. (2018) created a model called Xuanwu, which utilizes unsupervised pattern recognition methods in order to remove the human factor in the division of training and test samples. Experimental results show that the proposed framework outperforms the best methods in stock movement prediction. On a more detailed explanation, the first fact is that, it is impossible for human to compete with big data algorithms, so, the separation between train set and test set should be made by the algorithm. The training sample differs from other literature models because it works utilizing the probability a known shape be formed in a fixed duration of time when only the first days of analyze were conducted. Xuanwu also utilizes patter recognition algorithms to understand the movement of the stock being analyzed, these patterns are called shapes. The authors claim that this framework can be used by small startups that doesn't trade in the stock market so frequently, the analysts could use the method to narrow down they search for good stocks to invest in

### **3. Materials and methods**

The data set used in this paper was downloaded from Kaggle<sup>1</sup>. It contains the stock code, date, and values of close, high, low and volume for a given date. The first step is defining the period in which the data must be extracted. Later, it must be analyzed in order to obtain the period that contains the biggest number of stocks traded. Data preparation and handling was entirely conducted in Python 3.5 [Python Software Foundation (2018)]. The deep learning LSTM network was developed using Keras [Chollet et al. (2015)].

In order to achieve the desired results, it was necessary to complete a sequence of steps. First, it is necessary to (i) prepare the data for future utilization, it consists in downloading the data set, extract the data to be used in the prediction and evaluating if the extraction was done successfully. Later, the preprocessed data must be (ii) clustered, it guarantees that it will be possible to predict values from a large data set, once it would take a long time to run a neural network algorithm through thousands of stocks. Therefore, it will be necessary to find similarities between stocks, this will be done using K-Means algorithm, which will be applied on a correlation matrix generated by Pearson correlation coefficient metric. Once the data is grouped by its similarities, neural networks algorithms will be implemented to (iii) predict stock market prices. Finally, the results generated in last step will be evaluated.

#### **3.1 Clustering**

Mirkin (1996) defines clustering as a mathematical technique designed for revealing classification structures in the data collected in the real-world phenomena. Which means

---

<sup>1</sup> <https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs>

that clustering is the process of creation of clusters of similar objects [Bini and Mathew (2015)]. K-means clustering aims to partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean [Nanda et al. (2010)]. However, the data must be prepared prior to being used by this algorithm.

The stock return was calculated using Equation 1, where  $R_t$  represents the return for a specific period  $t$ .  $P_t$  is defined as the price of the stock at time  $t$  and  $P_{t-1}$  represents the price of that same stock at the instant just before  $P_t$  analysis [Affonso et. al (2017)]. In this study, the 1-day interval was used.

$$R_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \quad (1)$$

Pearson's correlation coefficient is used to measure how two linear variables are correlated. The result goes from -1 to 1, the signal indicates the direction of the relationship and the value suggests how strong it is. By other side, if the value is zero, there is no linear correlation between the variables [Filho and Júnior (2009)].

$$C_{i,j} = \frac{\langle R_i R_j \rangle - \langle R_i \rangle \langle R_j \rangle}{\sqrt{\langle R_i^2 \rangle - \langle R_i \rangle^2} \sqrt{\langle R_j^2 \rangle - \langle R_j \rangle^2}} \quad (2)$$

The math behind Pearson's correlation coefficient is shown in Equation 2, where  $\langle R_i \rangle$  is the average value of return prices, so the  $\langle R_i R_j \rangle$  can be explained as the average of the sum of the mean log return values of two corresponding stocks. In other words, it is the mean of the mean. The correlation coefficient is measured for all pairs of stocks in both exchanges.

Cluster analysis is an unsupervised method and one of the biggest problems with clustering is identifying the optimum number of clusters [Gan et al. (2007); Liu et al. (2010)]. Therefore, determining the optimum number of clusters for a given data set is extremely important. Some validation method must be used to achieve the appropriate number of clusters. There are several methods that can be applied, but in this paper the "elbow method" will be used. It consists in plotting a graph where the y-axis represents some kind of error metric and the x-axis represents the number of clusters. A marked flattening of the graph suggests that the clusters being combined are very dissimilar, thus the appropriate number of clusters is found at the "elbow" of the graph [Ketchen and Shook (1996)].

After obtaining the correct number of clusters, the K-Means algorithm was applied on all of the correlation matrices and the output was a file containing the stock and its correspondent group.

### 3.2 Neural Networks

Neural network consists of an interconnected group of artificial neurons, while deep learning, a method of machine learning, has developed several layers on the basis of neural networks [Li et al. (2018)]. These networks have been used to achieve state-of-the-art results on a number of benchmark data sets and for solving difficult AI tasks [Långkvist et al. (2014)].

As we are dealing with stock prices, the data set consists of just an array of numbers through time. Recurrent Neural Networks(RNN) are mainly used to predict a data sequence, for example the next word in a text or message. RNNs process an input sequence one element at a time, maintaining in their hidden units a 'state vector' that

implicitly contains information about the history of all the past elements of the sequence [LeCun et al. (2015)]. The RNN problem is that important information is faded away. It uses a back-propagation algorithm in order to replicate its results through layers. Although their main purpose is to learn long-term dependencies, theoretical and empirical evidence shows that it is difficult to learn to store information for very long [Bengio et al. (1994)]. For example, it would be harder for a RNN to predict the next correct word for a 2000 words text than for a 10 words sentence. This situation is called vanishing gradient problem.

LSTM networks are specifically designed to learn long-term dependencies and are capable of overcoming the previously inherent problems of RNNs, i.e., vanishing and exploding gradients [Fischer and Krauss (2018)]. This technique was first proposed by Hochreiter and Schmidhuber (1997) and aimed solving the vanishing gradient problem. It does so by keeping the error flow constant through special units called "gates" which allows for weights adjustments as well as truncation of the gradient when its information is not necessary [Nelson et al. (2017)]. Thus, the vanishing problem is solved and the RNN doesn't lose the important data through time.

#### 4. Results

The results are presented in three stages. First, information about the data is presented. The second step consists in using the data set in order to cluster the stocks by its similarity. Lastly, by using the groups created in the last part, the LSTM is trained and the predictions are made. The results are discussed in Section 5.

After analyzing the download data set, it was possible to observe that the period between 01/01/2008 and 29/03/2018 contains the biggest portion of the data. The result of the extraction is shown in Table 1, both exchanges had their size data reduced drastically.

**Table 1. Representation of the dataset before and after extraction**

Exchange	Number of Files	Number of Extracted Files
NYSE	3795	1225
NASDAQ	3066	804

##### 4.1 Clustering

In order to illustrate the results obtained in the last step, a correlation matrix was plotted and is presented in Figure 1. For each stock, in each stock exchange, the correlation between all of them was measured. It is possible to confirm that the main diagonal equals to zero and the rest of the matrix is between -1 and 1. Therefore, it is possible to use this result to group similar stocks.

The elbow method was used to obtain the best number of clusters. It consists in plotting a graph, shown in Figure 2, where it is possible to check the number of clusters versus the generated error, in this case it was utilized the Within-Cluster Sum of Squares. It is important to emphasize that we cannot just choose the biggest number of clusters, with the minimum error value, because it would cause an over-fit.



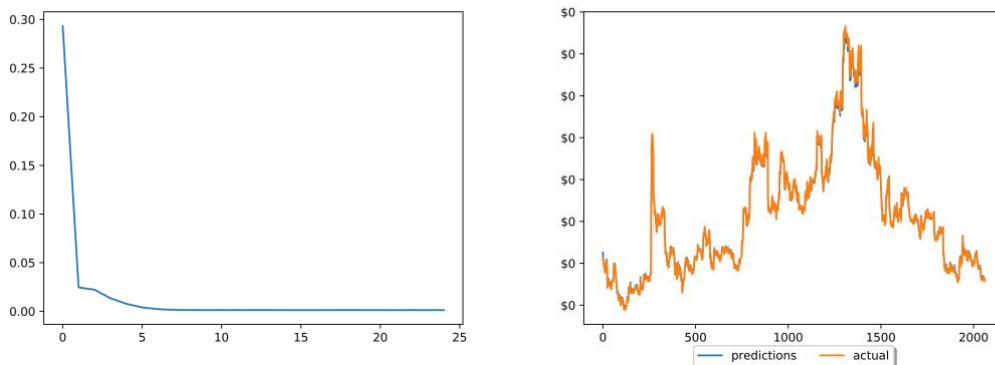
as a metric and the mean squared error(MSE) as the loss function. Some of the predicted values will be shown next.

**Table 3. Results for both Exchanges**

Exchange	Group	MSE Error
Nasdaq	Group_11	0.000241121
Nasdaq	Group_10	0.001105024
Nasdaq	Group_0	0.001893583
Nasdaq	Group_5	0.002498377
Nasdaq	Group_17	0.002949402
Nasdaq	Group_3	0.003049409
Nasdaq	Group_16	0.003102871
Nasdaq	Group_2	0.003516029
Nasdaq	Group_9	0.004704485
Nasdaq	Group_19	0.004859025
Nasdaq	Group_6	0.005061181
Nasdaq	Group_4	0.005986108
Nasdaq	Group_8	0.009971577
Nasdaq	Group_14	0.011123924
Nasdaq	Group_1	0.016487216
Nasdaq	Group_18	0.034465782
Nasdaq	Group_7	0.054812843
Nasdaq	Group_15	0.066097262
Nasdaq	Group_12	0.107150873
Nasdaq	Group_13	0.145795589

Exchange	Group	MSE Error
NYSE	Group_9	0.000353639
NYSE	Group_17	0.000464707
NYSE	Group_15	0.000620887
NYSE	Group_12	0.000789169
NYSE	Group_10	0.002226865
NYSE	Group_19	0.002670706
NYSE	Group_3	0.003475358
NYSE	Group_5	0.003642246
NYSE	Group_4	0.004226033
NYSE	Group_7	0.00573763
NYSE	Group_13	0.005756983
NYSE	Group_14	0.010200528
NYSE	Group_11	0.011839856
NYSE	Group_1	0.012058058
NYSE	Group_18	0.013770473
NYSE	Group_16	0.049443774
NYSE	Group_6	0.055835894
NYSE	Group_8	0.063193569
NYSE	Group_2	0.129570389
NYSE	Group_0	0.165804342

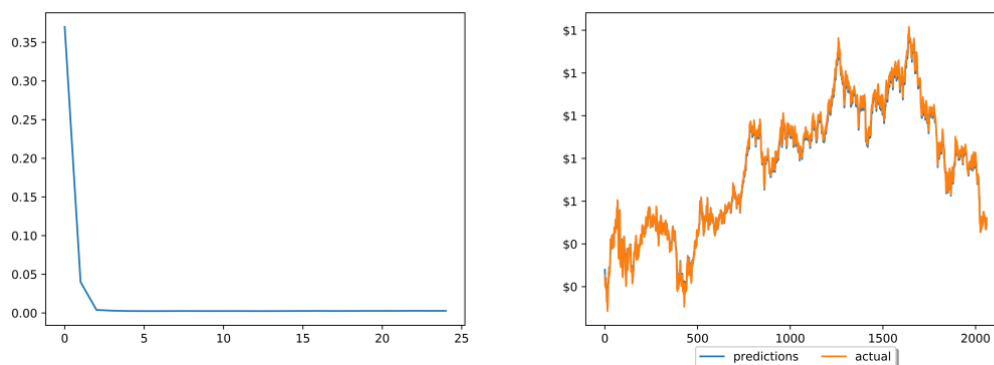
First, we must evaluate the metrics generated. Table 3 shows all the values sorted by their respective error for both exchanges. We can check that the error is small, but also, the accuracy is really small. In Figure 3 (a) it is possible to confirm that the error decreased over the time but - as we can check in Figure 3 (b) - the accuracy of the predictions is not so good. The predicted values (represented by a blue line) are really close to the real value(orange) but not on top of each other. Thus, the predicted values are not the ideal ones, but the network produced potential results.



**Figure 3. Results generated by the network for the Group 9 for the NASDAQ Exchange. a) Error vs. Epochs (Left); b) Prediction made by the LSTM (Right).**



The same behavior can be observed in Figure 4 for the NYSE Exchange. In Figure 4 (a) it is possible to observe how the error decreases, but the predictions are not on top of the real values (Figure 4 (b)). We can also observe that as the neural network tries to predict one value at a time, we can observe that a mistake in the beginning of the process influences the rest of the prediction.



**Figure 4. Results generated by the network for the Group 9 for the NYSE Exchange. a) Error vs. Epochs (Left); b) Prediction made by the LSTM (Right).**

## 5. Discussion and Future Work

We have shown how two different machine learning techniques can be used in pursuit of a single goal. The raw data was organized and important information was extracted. The remaining stocks were clustered by their correlation and groups of stocks were generated using K-Means algorithm. Finally, a LSTM network was applied in the data set with the purpose of predicting the value of a group of similar stocks. The process was done successfully and it proves that stocks can be clustered using K-Means algorithms and that the network prediction capacity did not suffer any impact.

These results can have a huge impact for investors and portfolio managers. By using the proposed techniques, it is possible to buy stocks that have the same behavior. Thus, less time and less computational power will be necessary in order to evaluate a whole group of companies.

Potential directions of future work include analyzing the impact of different similarity measures and how it can influence the process of clustering stocks. Also, when dealing with clustering, it will be necessary to study further options and algorithms that better represent the similarity between stocks. Furthermore, it seems that, even with some error in the beginning of the series, the stock tendency keeps unaltered, so it should be taken in consideration in future work. Additionally, better evaluation metrics should be used in the LSTM network aiming a better accuracy.

## Acknowledgment

This work is funded by Coordination for the Improvement of Higher Education Personnel (CAPES).

## References

- Affonso, F., de Oliveira, F., and Dias, T. M. R. (2017). Uma Análise dos Fatores que Influenciam o Movimento Acionário das Empresas Petrolíferas. In Ibero-Latin American Congress on Computational Methods in Engineering (CILAMCE).
- Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2), 157-166.
- Bini, B. S., and Mathew, T. (2016). Clustering and Regression Techniques for Stock Prediction. *Procedia Technology*, 24, 1248-1255.
- Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., and Oliveira, A. L. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55, 194-211.
- Chollet, F. (2015). Keras. Available at <https://keras.io>.
- Chong, E., Han, C., and Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83, 187-205.
- Filho, D. B. F. and Júnior, J. A. d. S. (2009). Desvendando os mistérios do coeficiente de correlação de pearson (r). Universidade Federal de Pernambuco.
- Fischer, T., and Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654-669.
- Gan, G., Ma, C., and Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (Vol. 20). Siam.
- Gerlein, E. A., McGinnity, M., Belatreche, A., and Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications*, 54, 193-207.
- Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- Jung, S. S., and Chang, W. (2016). Clustering stocks using partial correlation coefficients. *Physica A: Statistical Mechanics and its Applications*, 462, 410-420.
- Ketchen, D. J., and Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6), 441-458.
- Krauss, C., Do, X. A., and Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689-702.
- Kumar, B. S. and Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114, 128-147.
- Långkvist, M., Karlsson, L., and Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42, 11-24.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.

- Li, Y., Jiang, W., Yang, L., and Wu, T. (2018). On neural networks and learning systems for business computing. *Neurocomputing*, 275, 1150-1159.
- Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures, In *ICDM* , 911–916.
- Mirkin, B. G. (1996). *Mathematical classification and clustering*. Dordrecht, The Netherlands: Kluwer Academic Publishing.
- Momeni, M., Mohseni, M., and Soofi, M. (2015). Clustering Stock Market Companies via K-Means Algorithm. *Kuwait Chapter of the Arabian Journal of Business and Management Review*, 4(5), 1.
- Nanda, S. R., Mahanty, B., and Tiwari, M. K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12), 8793-8798.
- Nelson, D. M., Pereira, A. C., and de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. In *Neural Networks (IJCNN)*, 2017 International Joint Conference on (pp. 1419-1426). IEEE.
- Python Software Foundation (2018). Python 3.5.5 documentation. Available at <https://docs.python.org/3.5/>.
- Qiu, M., Song, Y., and Akagi, F. (2016). Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. *Chaos, Solitons & Fractals*, 85, 1-7.
- Zhang, J., Cui, S., Xu, Y., Li, Q., and Li, T. (2018). A novel data-driven stock price trend prediction system. *Expert Systems with Applications*, 97, 60-69.