ISSN 1870-4069

# Prediction of Stock Indices Using Artificial Intelligence

Francisco Javier Hermosillo Alonso, Alejandro Padilla Díaz,
Julio Cesar Ponce Gallegos, Miguel Ángel Meza de Luna

[1] Universidad Autónoma de Aguascalientes, Aguascalientes,
Mexico

paco.ki114@gmail.com,{julio.ponce, alejandro.padilla,
miguel.meza}@edu.uaa.mx

**Abstract.** This paper will address an approach to hybrid market analysis. It aims to unite the two main types of analysis in the study of market behavior. The fundamental and technical study of the behavior of stock indices is generated a possible description of the behavior that these may have in the future. The computational tools to achieve this purpose will be recurrent neural networks and multiple linear regression in the case of technical analysis, while for the fundamental one, sentiment analysis will be used on Twitter to identify later the correlation between the polarity of tweets and the actual behavior of the stock index. It is shown that the union of fundamental and technical analysis is a whole way of generating with more precision scenarios in the short and medium-term of the course that a stock index can take over time.

**Keywords:** Stock Indices, Neural Networks, Multiple Linear Regression, Prediction.

## 1 Introduction

Today there are a wide variety of applications that allow anyone to easily access the stock market to invest, so it is important to have tools that allow you to predict the behavior of stock indices.

This work shows an implementation that allows predicting some stock indices and proposes a prediction method using hybrid techniques of fundamental and technical analysis, using statistical tools and artificial intelligence that allow a valuable analysis of which are the most recommended variables for predict an index and thus anticipate future behavior and thus make a good investment. A model assembly algorithm will be used to concatenate the results of each development scheme and thus obtain better results.

The hybrid approach to market analysis has been controversial because, as Herta Arias indicates in [1], regression methods have shown the great predictive capacity of fundamental analysis variables, while those of technical analysis would work to explain the short term.

Herta proposes a hybrid model approach, because if it is possible to combine both analysis systems in theory, they would provide a better model for predictive capacity [1].

Agudelo Aguirre in [2] explains that technical analysis is not concerned with whether the price associated with an action is correct or even with the reasons that led an action to take a certain price, as it aims to focus on the price trend and the possible changes associated with that change. Aguirre explains that in this analysis it is said that there are a series of elements of psychological behavior that are not enough to explain the real behavior of the market.

While in the case of fundamental analysis, it is responsible for finding and exploring the various variables that directly impact the future results of the different stock indices. This analysis sets the main objective of knowing the true value of a company, under which it is possible to assume the future behavior of its value [2].

Although it seems that both approaches are mutually exclusive, it is intended to demonstrate that by using both approaches in the prediction of future values, results can be obtained that more accurately explain behavioral trends in the analysis of a given market.

This will be achieved by using data mining techniques, neural networks, and various statistical tools such as multiple linear regression to achieve a robust model that achieves this hybrid approach in the analysis of stock indices.

## 2 Theoretical Framework

This document aims to demonstrate how through technical analysis it is possible to make a prediction about the values that the price of a share can take over time. This through linear correlation techniques and neural networks.

To do this it is necessary to understand some key concepts that will be used throughout this research work, this will be divided into two blocks.

### 2.1 Financial Concepts

For Mireles Vázquez 2012 [3] The stock exchange is a private organization that provides facilities to brokers or intermediaries to various clients, regarding possible negotiations for the purchase and sale of securities.

These values range from shares of companies or societies of different kinds, public and private bonds, as well as a wide range of investment tools. All decisions made in the stock market are based on prices that are known in real time, where the transaction systems are regulated by legal and security systems.

Martínez Mendoza in [4] describes the securities market as a series of tools and mechanisms that allow the issuance and distribution of securities. He mentions that as in any market there is a supply component and another that corresponds to demand.

In financial terms, a share is defined in [5] as the share capital of a company or corporation that represents what it owns, in general and under certain exceptions, a

share can be transferable, it grants both economic and political rights to the owner of the action.

In [1] it is explained that NASDAQ lists the 100 stocks of the most important companies in the industry sector including hardware and software companies, telecommunications, retail / wholesale, and biotechnology companies listed on the Stock Exchange New York (NYSE). Both American and international companies can be included in the index, this index is the one that will be taken as a reference to select 10 of the best companies.

For [6] and [7] the following elements are essential when acquiring shares in a certain company:

1. Sales that the company has.
2. Utilities.
3. Flow of assets (Cash).
4. Value of the company in the market.
5. Current assets (That is, constant flow of assets).
6. Short-term and long-term liabilities (the behavior of the company must be monitored at different terms)
7. Social and accounting capital
8. Utility of the company

## 2.2 Model Concepts

Currently there are different types of models used in the prediction of values, as well as for the analysis of the foreign exchange market[9].

### 2.2.1 Linear Correlation

For Guillen in [8] Linear correlation measures the degree of precision in a prediction that depends on a relationship between two variables, which is known as the degree of correlation or even by association of two variables. The linear correlation corresponds to a mathematical model that studies the dependence between two quantitative variables (In it there are both independent and dependent variables).

Correlation is thus basically a measure of a normalized nature of the linear covariance between two variables. This model or index can vary in the range of -1 to +1, where a type of perfect correlation can be indicated at the end of the range, either positive or negative. While the value of $r = 0$ indicates that there is no type of linear correlation between the two variables analyzed. Whereas if we find a negative value of r we would be facing a relationship that goes in the opposite direction [10].

### 2.2.2 Neural Networks

Computational system made up of many elements that are considered simple, of interconnected process elements that oversee processing information through its dynamic state product of external inputs [11].

LSTMs are a special type of recurring networks. The main characteristic of recurring networks is that information can persist by introducing loops in the network diagram,

so they can basically remember previous states and use this information to decide what will be next. This feature makes them very suitable for managing time series [12].

Currently there are various works that try to predict the behavior of the stock market using neural networks, as can be seen in [13, 14].

# 3 Metodology

The research methodology will be developed in different stages that will seek in the first instance to look for the generalities of the stock market on supervised learning methods, data collection, generation of databases, analysis of variables, definition, and creation of systems of prediction, tests, obtaining reliability of data obtained and finally the comparison of results obtained in order to generate good prediction scenarios.

The development of the research is divided into 3 development blocks that will contain different phases with which it is intended to achieve a specific objective.

## 3.1 Development of Technical Analysis

Stage 1: Construction of databases based on historical information on actions, as well as the selection of the most relevant variables to make the prediction.

Stage 2: Development of the linear regression and recurrent neural network prediction models.

Stage 3: Evidence and reliability of data obtained

In the case of the linear regression method, the PANDAS library will be used, so that the system can read the data from a CSV file for the databases. Later the data will be converted into a single dimension vector. Once you have the data you are going to work with, you will divide the data set into the training set and the test set. To finally use the sklearn library, as follows:

```
sklearn.linear_model.LinearRegression(*, fit_intercept=True,
normalize=False, copy_X=True, n_jobs=None, positive=False)
```

This will allow to fit the data to a linear model with coefficients w = (w1,…, wp) to minimize the residual sum of squares between the objectives of the data set and the objectives stated by the linear approximation. To finally find the error and check the accuracy of the model.

Finally, in the case of the neural network system, PANDAS will be used in Python for the same case as in the linear regression model, to later use a normalization function to adjust the scale of the model with which the inputs can be chosen. and outputs of the neural network.

After the above, the input and output of the network are restricted according to the operation of LSTM and thus initialize the network by adding the input layer and the LSTM layer and thus generate the training set with which the neural network can work recurrent.

The linear regression algorithm is based on the following scheme:

1. Read the data from a CSV file to get the database.

2. Convert the data into a one-dimensional vector so that it can be processed by the model.
3. Divide the data set into the training set and the test set.
4. Use the sklearn library to build the model and thus minimize the residual sum of squares between the objectives of the data set and the objectives stated by the linear approximation.
5. Fit the model.
6. Find the error and check the accuracy of the model.
7. Graph the prediction by comparing the actual value against the predicted value.

The input parameter used for the linear regression models is the "test_size" that will indicate the percentage of use of the data taken from the database to be used for the training of the model.

The LSTM algorithm obeys the following:

1. Read the data from a CSV file to build the database.
2. Import the training set
3. Obtain only the price of the shares open for the entry of RNN.
4. Use normalization as a function of scale.
5. Get the inputs and outputs
6. Restrict entry and exit based on LSTM performance.
7. Build the RNN, in addition to initializing the RNN
8. Add the input layer and the LSTM layer
9. Add the output layer
10. Compile the RNN and all the layers together.
11. Adapt the RNN to the training set.
12. Make predictions and visualize results
13. Find the error and check the accuracy of the model.
14. Graph the prediction by comparing the actual value against the predicted value.

While the input parameter used for the LSTM model will be the "Epochs" that indicate the number of times the model is executed. In each cycle (epoch) all the training data goes through the neural network so that it learns about each one of them.

### 3.2 Development of Fundamental Analysis

Stage 1: Selection of keywords from which you can obtain number of tweets about the company you are looking to analyze.

Stage 2: The polarity of a tweet will be classified given its intention and it will be determined whether the opinion in the tweet is positive, negative, or neutral.

Stage 3: Subsequently, the machine learning algorithm will be used to predict sentiment and find the correlation between sentiment and stock prices.

This algorithm will make use of Naïve Bayes and vector support machine techniques to classify tweets and then measure the correlation between their polarity and their market value. Ranking will be used to generate a prediction. In this case a series of tweets will be classified that will be labeled as "positive", "negative" or "neutral".

**Table 1.** Results of the implemented models.

| Model | Company | Best result precision | Actual value vs Prediction value | Graph |
|---|---|---|---|---|
| Linear correlation algorithm 1 | Facebook | 93.62% | Actual Value 266.649994 Prediction Value 159.69893247 | Fig. 1 |
| Linear correlation algorithm 2 | Berkshire Hathaway | 99.35% | Actual Value 350460.00 Prediction Value 350822.210991 | Fig. 2 |
| Neural Networks Algorithm 1 | Procter & Gamble | 99.87% | Actual Value 133.089996 Prediction Value 137.25214 | Fig. 3 |
| Neural Networks Algorithm 2 | Johnson & Johnson | 99.72% | Actual Value 160.5 Prediction Value 146.1570281982422 | Fig. 4 |

For the classification of these, two of the most common text classifiers were used: Naïve Bayes Bernoulli and vector support machines. Using this ranking you will find the correlation of the statistics of the tweets about stocks and the prices of the stocks for a given day.

The following must be followed for the fundamental method algorithm:

1. Obtain Twitter data using keywords for a specific company.
2. Collect the tweets of the company's shares and then save them in a csv file that separates the tweet time and the text in different columns.
3. Clean up the tweet by removing stopwords and create a set of functions for the training data.
4. Classify all training data so that through this classification the sentiment of the tweets is analyzed.
5. Calculate the total number of positive, negative, and neutral tweets.
6. Generate a dataset for stock prediction that contains a set of functions such as tweet sentiment statistics and a set of targets such as upward or downward direction of stocks.

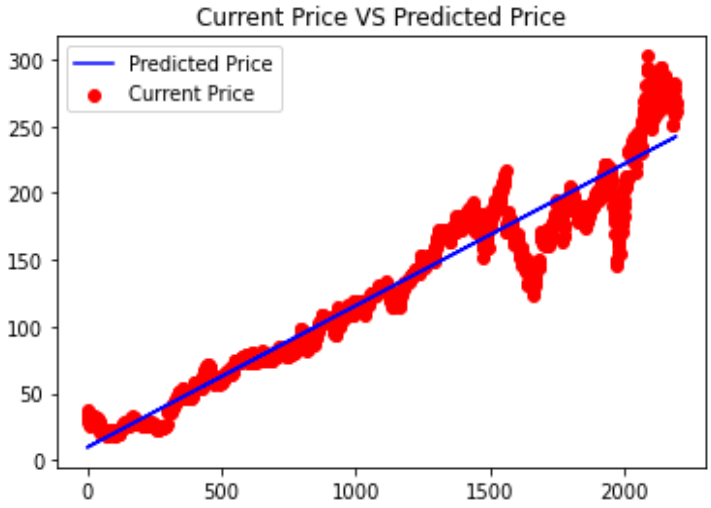


**Fig. 1**. Facebook.


**Fig. 2.** Berkshire Hathaway

7. Classify the stock prediction data set to predict the next direction of stocks.

### 3.3 Hybrid Prediction Development

Stage 1: Using a model assembly method, a process will be created by which the models derived from fundamental and technical analysis are combined in order to generate a prediction method given the historical and opinion data of a particular market.

Stage 2: Comparison of the 3 blocks developed in what is going to verify the efficiency of the hybrid approach exposed in [1] by Herta Arias.

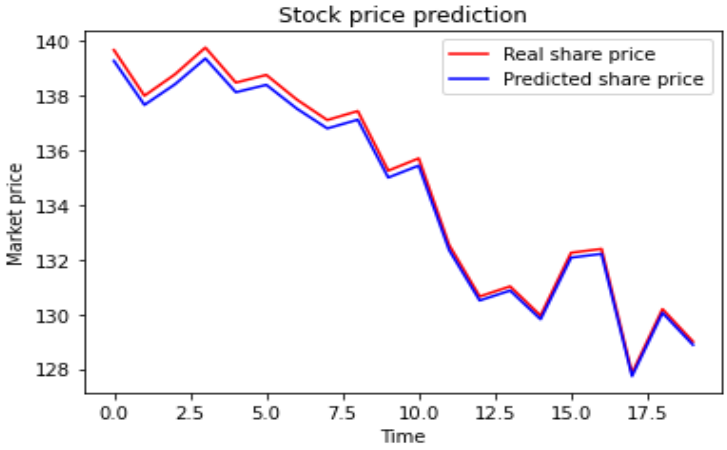


**Fig. 3.** Procter & Gamble

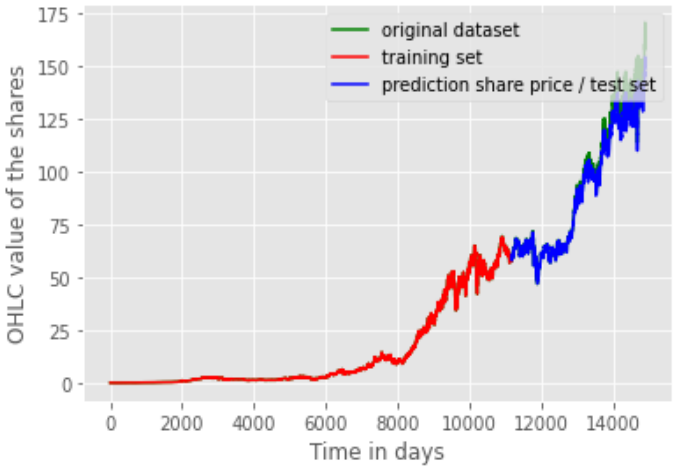


**Fig. 4.** Johnson & Johnson

The assembly model will be built from the concatenation of the fundamental analysis model with the technical one in which the outputs of both models will be exposed to a correlation measure in which it will be described if there is any relationship between the real value of the stock on a given day with the value that was extracted from the correlation measure.

To later make use of an LSTM network that will be trained with the correlation of both models and the real value of the stock, to finally predict the value or behavior that it may take in the future.

The following must be followed for the fundamental method algorithm:

1. Get data from Twitter and Yahoo Finance to feed fundamental and technical models.
2. Initialize fundamental and technical models.

3. Train both models.
4. Get outputs from the models.
5. Obtain the multiple linear correlation between the output of both models and the real price of the stock.
6. Use normalization as a function of scale and get the inputs and outputs.
7. Build the RNN, in addition to initializing the RNN.
8. Add the input layer and the LSTM layer and add the output layer.
10. Compile the RNN and all the layers together.
11. Adapt the RNN to the training set.
12. Make predictions and visualize results.

## 4 Prediction with Technical Analysis

Implemented 2 algorithms of which are using a linear correlation and neural networks, described in sections 3.1 and 3.2 respectively, to analyze the efficiency tested with historical data since they began to trade on the stock exchange, of some of the most important companies that are taken into account in the NASDAQ index, as shown in table 1.

## 5 Conclusions

The results of this research make it clear that the fundamental analysis methods are very complex when it comes to being translated into a supervised learning algorithm that can generate a good scenario for the prediction of stock indices, this due to the great variability of variables at the same time. that are exposed.

Well, it is not enough to measure the opinion or popularity of a company with social groups to know the behavior that it will have in any period, but this does not mean that it does not contribute anything to the prediction of the behavior of a security.

Well, if the fundamental analysis algorithm will be complemented with the help of other variables that are not only those of the popularity of a company, such as its financial situation, its level of confidence, its capitalization and it is also mixed with a correct approach technical analysis could lead to a more robust hybrid prediction method that can accurately predict behaviors over multiple time frames.

As the hybrid algorithm was evident in this first approach, it left somewhat regular results, but it is for the aforementioned that it is necessary to adjust the fundamental analysis model so that both methods are cohesive with the assembly algorithm and result in a much analysis. more complete information on the real situation of the company to generate higher quality predictions.

The hybrid approach is still far from providing accurate predictions of market behavior, but its potential development will change the way markets are understood. Well, the more algorithms and variables that are provided to machine learning models, the higher the quality of the generated model and with it the prediction of any financial medium.

## References


1. Arias Huerta, S., González Berríos, M., Fuentes Castro, H.: Predictibilidad de los análisis técnico y fundamental en mercados latinoamericanos evidencia empírica y aplicación práctica (2015)
2. Aguirre, A. A. A.: Aplicación de una estrategia activa de inversión en renta variable en el mercado de acciones colombiano. NOVUM, revista de Ciencias Sociales Apliacadas, 2(8), 140–178 (2018)
3. Mireles Vázquez, I.: Bolsa de Valores "¿Cómo? ¿Por qué? Y ¿Para qué?", No 21, UAM (2012) http://tiempoeconomico.azc.uam.mx/wp-content/uploads/2017/07/21te4.pdf
4. Martínez Mendoza, M. A.: Cómo invertir en la Bolsa Mexicana de Valores. Doctoral dissertation, Universidad Autónoma de Nuevo León (2000) http://eprints.uanl.mx/6460/1/1080111914.PDF
5. Díaz, A.: El mercado bursátil en el sistema financiero. México: McGraw-Hill (2005)
6. Gitman, L. J., Zutter, Ch. J.: Principios de administración financiera (12.ª ed.). México: Pearson Educación (2012)
7. Herrera, C. E.: Mercados financieros. México: SICCO-Gasca (2003)
8. Badii, M. H., Guillen, A., Cerna, E., Valenzuela, J., Landeros, J.: Análisis de Regresión Lineal Simple para Predicción. Revista Daena (International Journal Of Good Conscience), 8(1) (2012) http://www.spentamexico.org/v7-n3/7(3)67-81.pdf
9. Aguilar, D., Batyrshin, I., Pogrebnyak O.: A Survey on Computer Science Techniques in the FOREX Market: Models and Applications. Research in Computing Science, 138, pp. 89–98 (2017)
10. Vinuesa, P.: Correlación: teoría y práctica. No 1, UNAM (2016) https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema8_correlacion.pdf
11. Matich, D. J.: Redes Neuronales: Conceptos básicos y aplicaciones. Universidad Tecnológica Nacional, México, 41 (2001) https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf
12. Garzón, J. I.: Cómo usar redes neuronales (LSTM) en la predicción de averías en las máquinas. (2018) https://blog. gft. com/es/2018/11/06/como-usar-redes-neuronales-lstm-enla-prediccion-de-averias-en-las-maquinas.
13. Iguarán Cotes, J. M.: Aplicación de redes neuronales para predecir el precio de acciones en la bolsa colombiana. Bachelor's thesis, Uniandes (2019)
14. Chandar, S. K.: Grey Wolf optimization-Elman neural network model for stock price prediction. Soft Computing, 25(1), 649–658 (2021)