# Introduction

# The first applications of machine learning to analyze economic and financial data focused on the improvement of algorithms, particularly in the use of artificial neural networks applied to the time series of stock markets (Enke & Thawornwong, 2005; Leight et al., 2002). After exploring the applications of machine learning to numerical data, the researchers were also interested in examining how these applications work with textual data. In this way, the first studies focused on text mining tools based on the calculation of indices and on the analysis of sentiment on news related to the behavior of the financial market (Nikfarjam et al., 2010; Ruiz et al., 2012 ; Schumaker et al., 2012; Wang et al., 2012).

# In one study, the approach known as word-bag (in which words are extracted from their context and examined in isolation) was used by combining the linear and non-linear elements of the time series that were combined with the data. textuals to analyze the return of shares in insurance companies (Wang et al., 2012). In another application of sentiment analysis, the relationship between financial news and stock prices was studied by examining the use of proper names, the tone of the news [objective, subjective and neutral] and its polarity [positive, negative and neutral] (Schumaker et al., 2012). It has also been analyzed to what extent the tweets are related to the change in the price of the shares of those companies listed in the Standard & Poor 500 index, finding that the volume of transactions of a share is correlated with the closing price of the shares (Ruiz et al., 2012).

# Later, the researchers were interested in the combined use of machine learning algorithms, as seen in various studies (Groth & Muntermann, 2011; Gunduz & Cataltepe, 2015; Nassirtoussi et al., 2015). In one investigation, text mining techniques were used to examine corporate news stories to see if they were related to unusual variations in market prices, finding that the news was associated with high levels of volatility, an important finding for managing the market. risk. Of the four algorithms used - Naive Bayes, k-nearest neighbor, vector support machine (MSV) and multilayer neural networks -, MSV achieved the best performance when classifying the news, according to the results of accuracy, precision and recovery ( Groth and Muntermann, 2011). Furthermore, text mining techniques have been employed to transform news from finance websites into attribute vectors using the naive Bayes algorithm to model the relationship between attribute vectors and stock prices to predict changes in the market. (Gunduz and Cataltepe, 2015). Another research sought to model market behavior by analyzing financial news headlines using a three-level text mining approach: semantic analysis of the terms used in the news, sentiment analysis to identify the affective response of investors and the reduction of dimensions of the attributes extracted from the news (Nassirtoussi et al., 2015).

# Data and Methods

## Data extraction

As the first step in our data extraction process, we extracted 100K rated user reviews. These reviews covered very variated items such as films, products and places and were obtained using *web crawling* software in public websites as filmaffinity, tripadvisor, ebay, and eltenedor. The 100K reviews contain texts with opinions and numerical ratings associated

The next step was to obtain tweets from the main business association accounts in Spain and Peru and also the main companies in both countries. We obtained:

* 120K tweets from the 40 main business association accounts in Spain
* 269K tweets from the 35 main companies (ibex 35) in Spain
* 17K tweets from the 30 main business association accounts in Peru
* 72K tweets from the 20 main companies in Peru

These tweets cover the period dated from january 2017 to march 2020.

This two steps can be visualized in figure 1.

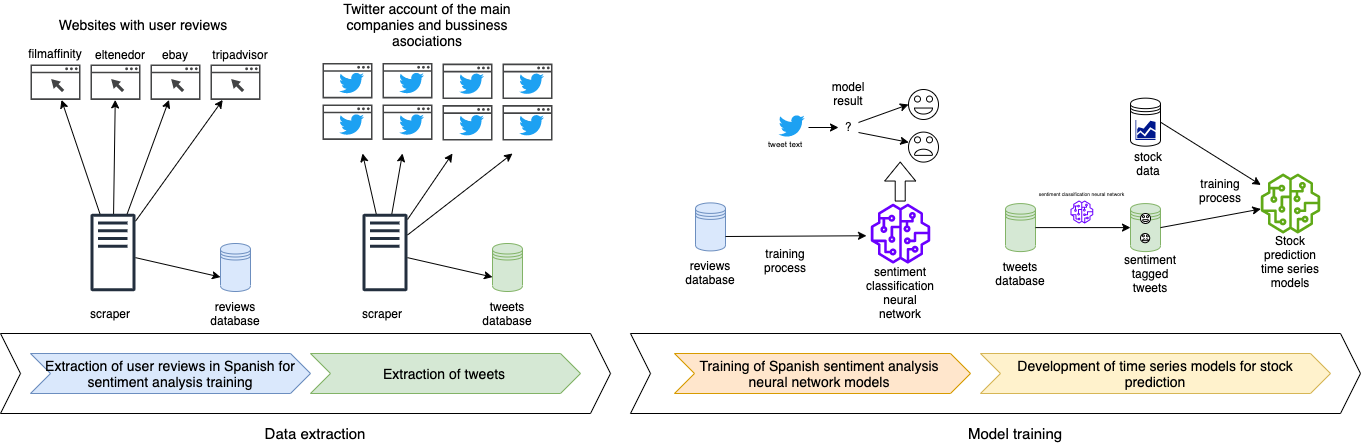


Figure 1.

Spanish sentiment analysis neural networks training using reviews data

As a key component of our analysis we needed to capture the negativity or positivity of the tweets. This is achieved using sentiment classification. Unfortunately, there is not a standard library for sentiment classification in the spanish language. So we build a neural network classification model to achieve this goal.

Using the 100K rated reviews in spanish extracted using web crawling we separated them in two groups: *positive reviews* and *negative reviews*. Since there were more positive than negative reviews, we balanced the two classes removing enough positive ones to have the same size in both groups.

For the neural network architecture we used an embedding layer, four convolutional (1 dimensional) layers, two poolings and a one dimensional output dense layer. Other architectures were tried, but we obtained poorer results in terms of validation accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| **data** | **accuracy** | **precision** | **recall** |
| training data | 0.9988 | 0.9991 | 0.9992 |
| test data (validation) | 0.9869 | 0.9840 | 0.9870 |

## 

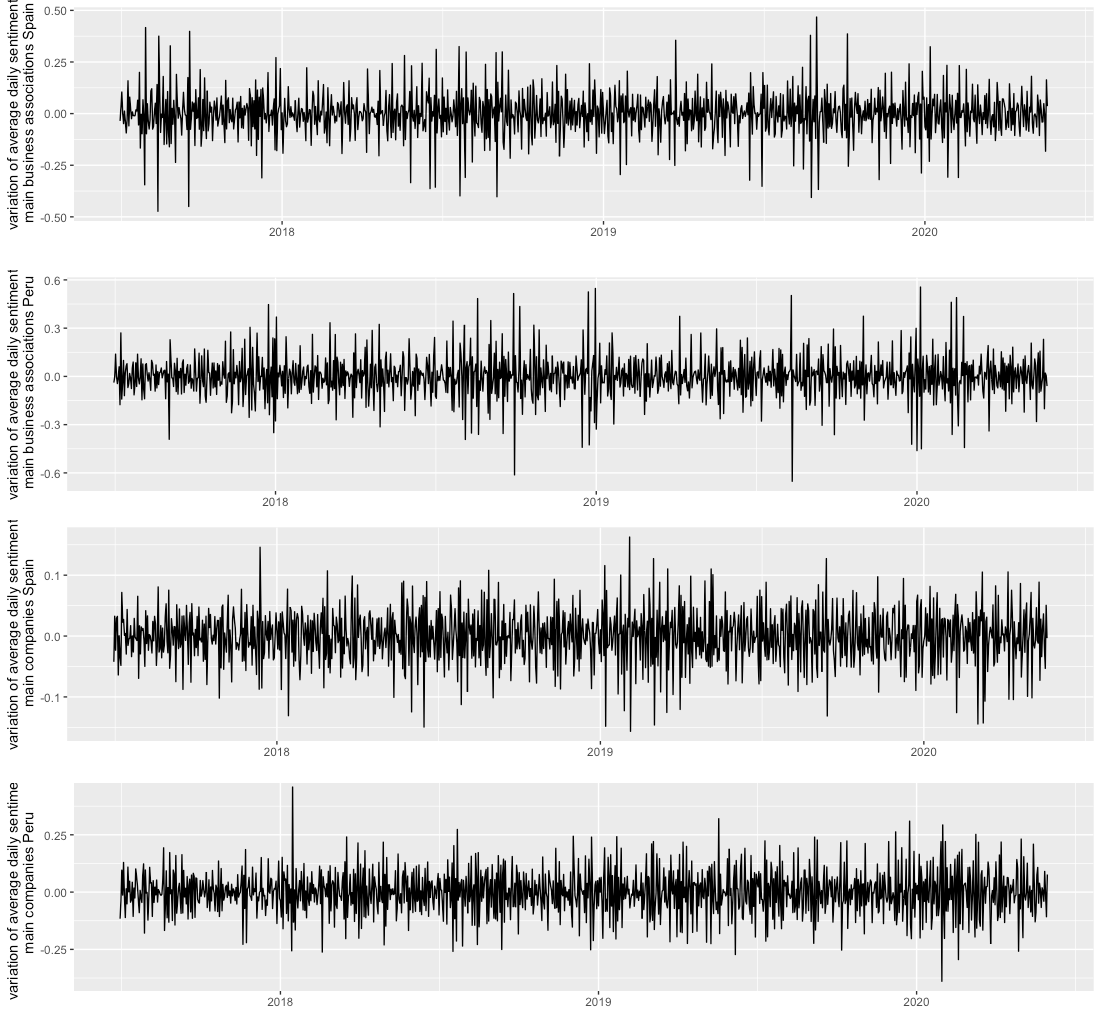
## Tweets processing using sentiment analysis

Using the neural network sentiment model described in the previous section, we proceed with the analysis of the tweets from the business associations and companies using the following steps:

1. We run the sentiment analysis on each tweet. This gives us a number between 0 and 1 expressing the positivity or negativity of the tweet.
2. We calculate the average of all the tweets published on the same day in each group.
3. We calculate the difference between the average sentiment in each given day and the day before to remove the trend component and make the data more comparable.

This transforms the data into a time series that we can study and compare to macro economic data. In particular we obtain four time series: daily sentiment of main companies and business associations both in Spain and Peru.

This four time series are the starting point in our analysis and modeling process. See figure 1



**figure 1** Taking a look to the beginning of 2020, we see how the time series become more anomalous which could complicate the modeling process. We will come back to this later .

# Results

## Prediction of Madrid and Lima stocks change using traditional models

We use the daily average sentiment data described in the previous section to predict the fluctuations of stock values.

First, we see that the daily average sentiment is a stationary time series (both for associations and main companies in both countries). This are the p-values of the Dicker-Fuller test for the four time series

|  |  |  |
| --- | --- | --- |
| **p-value** | **results daily variation sentiment business association** | **results daily variation sentiment main companies** |
| Spain | 3.541674e-23 | 3.280194e-26 |
| Peru | 0.000116 | 0.002758 |

We studied the Granger causality of the variation of the sentiment and the stock values, but we only found significant p-values by restricting the data to the last 5 months of the study (from January 2020 to May 2020). The significative p-values were only obtained for the stocks in Spain.

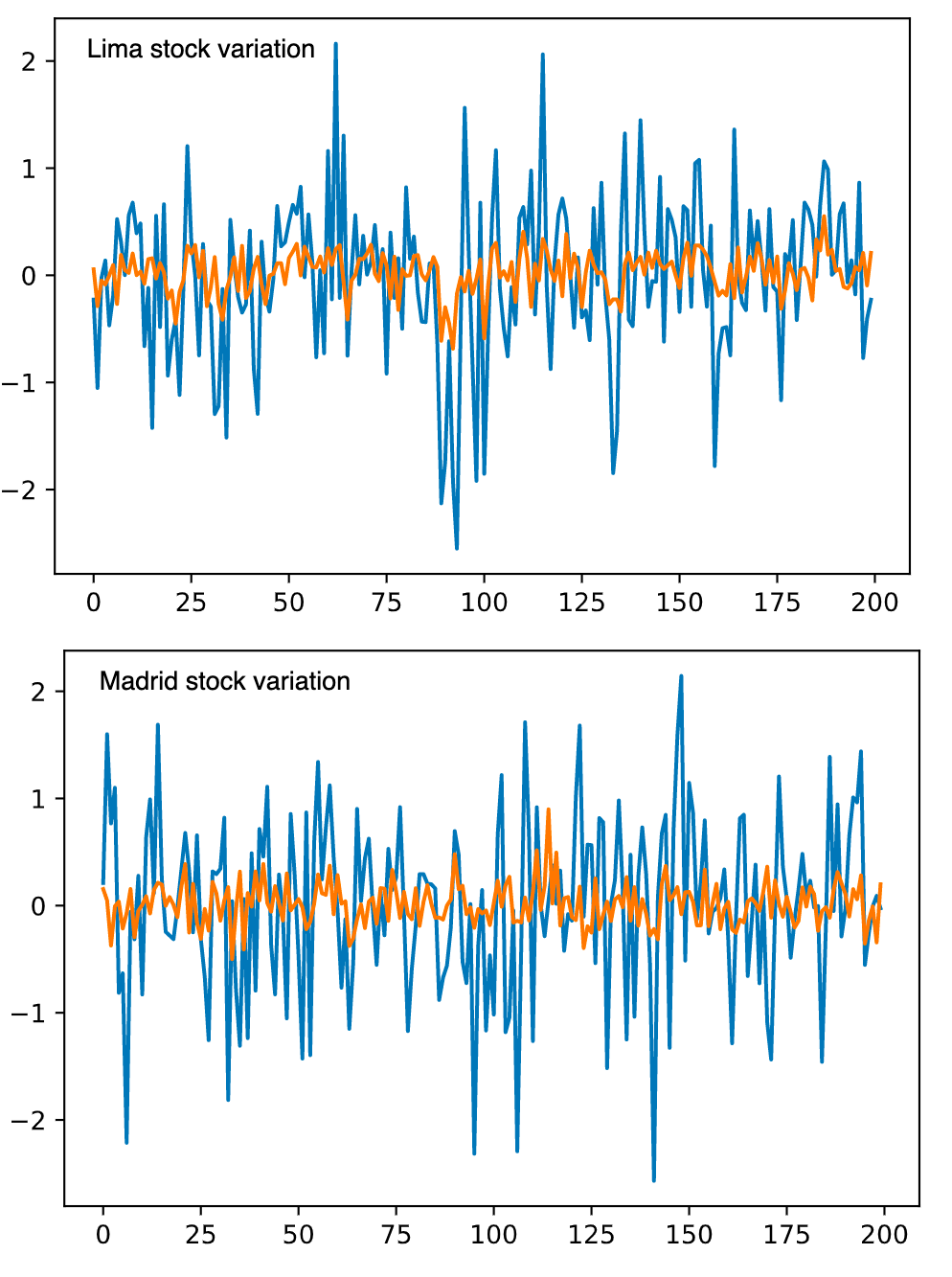
This are the p-values of the Granger causality test of the daily variation of sentiment and the prize change in the stock value (Madrid and Lima respectively)

|  |  |  |
| --- | --- | --- |
| **p-value** | **variation sentiment business association vs variation stock price** | **variation sentiment main companies vs variation stock price** |
| Spain | 0.0282 (lag 5), 0.0530 (lag 6), 0.0378 (lag 7) | 0.0170 (lag 5) |
| Peru | >0.05 (all lags) | >0.05 (all lags) |

Using a VAR model to predict the variation of stocks change one day in advance from the variation in the sentiment both for business associations and companies, we obtain a MAE (mean absolute error) of 0.8735 for Spain and 0.5212 for Peru

Nevertheless, we see that there is a great volatility in the months of COVID-19 pandemic, so removing those three months (february, march and april 2020) we see that the model performs much better in terms of MSE: 0.6931 for Spain and 0.5101 for Peru.

figure 2 shows the sock values (removing the pandemic months) in blue and the prediction in orange.



## Prediction of Madrid and Lima stocks change using neural networks models

Neural network models present a different approach to forecasting time series. We compare LTSM neural networks and Convolutional neural network models.

The model is designed to forecast a day in advance using a window of the previous 7 days as an input. The neural network models were trained taking all the 7 day windows of sentiment data and stock data and splitting them into test data (15%) and train data (85%).

This are the model designs:

* The LTSM model is componed by a 200 size layer of LSTM neurons, a dense layer with 100 neurons and an output layer of 1 neuron.
* The Convolutional model is composed by two one dimensional convolutional layers with 64 filters, a pooling layer, a dense layer with 100 neurons and an output dense layer with 1 neuron.

The outputs have size 1 because we are only predicting one day in advance.

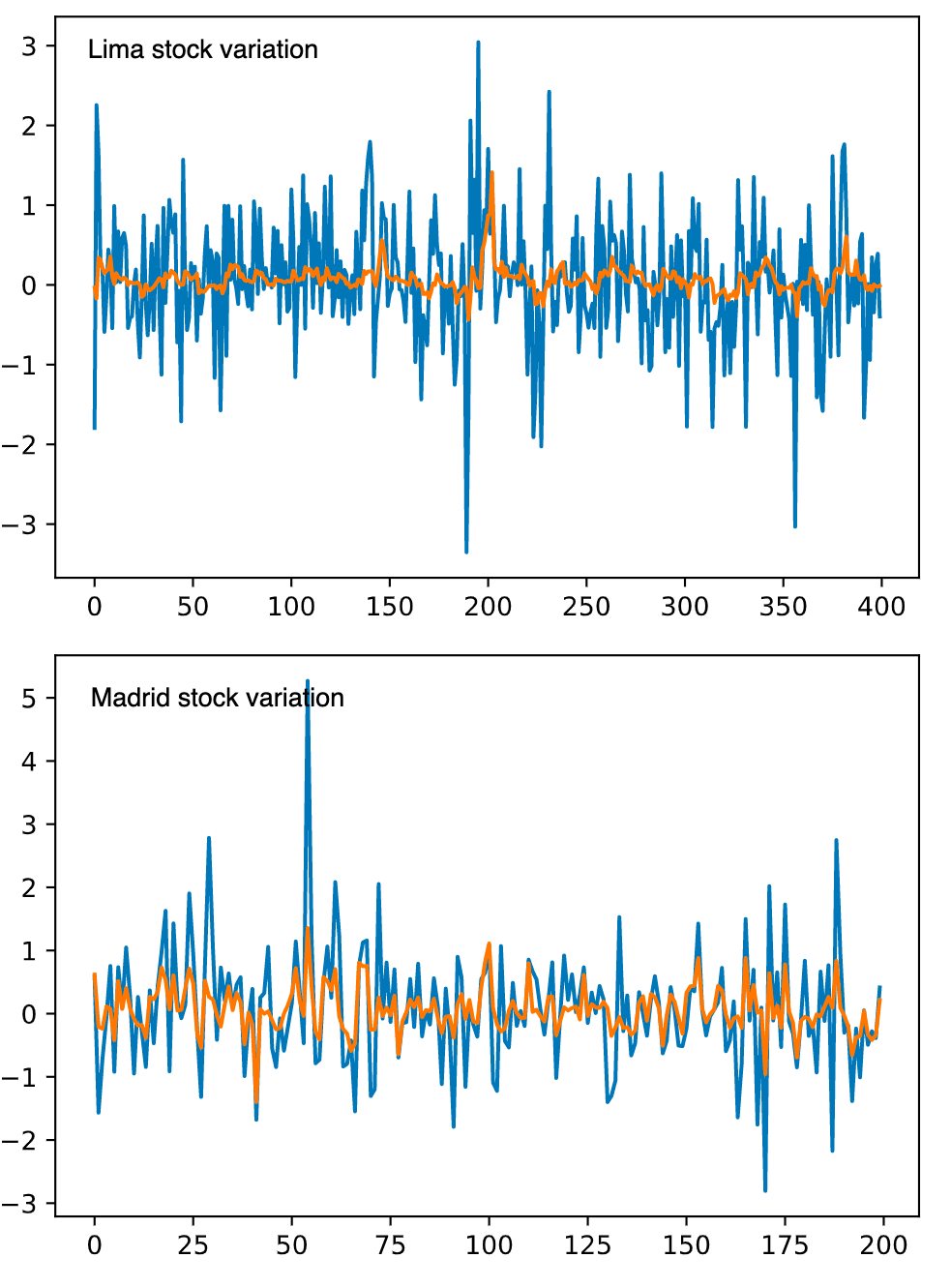
These are the mean absolute errors of the two models

|  |  |  |
| --- | --- | --- |
|  | **LSTM mean absolute error** | **Convolutional networks mean absolute error** |
| Madrid stocks data | 0.7061 | 0.8172 |
| Lima stocks data | 0.7529 | 0.6754 |

As we did with the VAR models, due to the high volatility during the COVID months, it makes sense to repeat the training of the model using only the data from the months prior to the pandemic. These are the results obtained after the restriction

|  |  |  |
| --- | --- | --- |
|  | **LSTM mean absolute error** | **Convolutional networks mean absolute error** |
| Madrid stocks data | 0.5030 | 0.6908 |
| Lima stocks data | 0.5459 | 0.5576 |

figure 3 shows the 200 sock values (taken removing the pandemic months) in blue and the prediction in orange. The prediction in the figure was obtained using the LSTM model.



References

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, *29*(4), 927-940. doi: 10.1016/j.eswa.2005.06.024.

Groth, S. S. , & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, *50*(4), 680–691. doi: 10.1016/j.dss.2010.08.019.

Gunduz, H. , & Cataltepe, Z. (2015). Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. *Expert Systems with Applications*, *42*(22), 9001–9011. doi: 10.1016/j.eswa.2015.07.058.

Leigh, W., Purvis, R., & Ragusa, J. M. (2002). Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. *Decision Support Systems*, *32*(4), 361-377. doi: 10.1016/S0167-9236(01)00121-X.

Nassirtoussi, A. K. , Aghabozorgi, S. , Wah, T. Y. , & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1), 306–324. doi: 10.1016/j.eswa.2014.08.004.

Nikfarjam, A. , Emadzadeh, E. , & Muthaiyah, S. (2010). Text mining approaches for stock market prediction. In *International Conference on Computer and Automation Engineering*. Singapore, Singapore, 26-28 February 2010 (pp. 256–260). doi: 10.1109/ICCAE.2010.5451705.

Ruiz, E. J. , Hristidis, V. , Castillo, C. , Gionis, A. , & Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining* (pp. 513–522). doi: 10.1145/2124295.2124358.

Schumaker, R. P. , Zhang, Y. , Huang, C.-N. , & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464. doi: 10.1016/j.dss.2012.03.001.

Wang, B. , Huang, H. , & Wang, X. (2012). A novel text mining approach to financial time series forecasting. *Neurocomputing*, *83*, 136–145. doi: 10.1016/j.neucom.2011.12.013.