International Journal of Electronics Communication and Computer Engineering
Volume 8, Issue 5, ISSN (Online): 2249–071X

# Process for Predicting Stock Prices in the Financial Market using Big Data

**MSc André Hideo Hayashi*** **and Ph.D. Alexandre José Barbieri de Sousa**
*Corresponding author email id: andre.hideo@gmail.com

*Date of publication (dd/mm/yyyy): 07/10/2017*

***Abstract*** **– The financial market is very fickle and investors have the difficult task of following and trying to predict the market oscillations so that their strategies result in better financial returns. With the use of *Big Data* and *Bayesian* mathematical statistics based on prior knowledge and training examples to determine the likelihood of a hypothesis, financial news can be tracked continuously and affecting key financial indicators, actions and assisting the investor. The objective of this work is to propose a stock price prediction process of a listed company in the *Ibovespa*, helping an investor to better buy and sell stocks using a predictive system with *Big Data* and Natural Language Processing) to collect real-time information from news sites that may influence stock prices.**

***Keywords*** **– Big Data, Naive Bayes, Financial Market, Natural Language Processing, Ibovespa.**

## I. INTRODUCTION

The financial market, even being controlled by regulatory agencies, is very fickle. Its inconstancy can be influenced by several factors such as economic crisis, political and governmental changes, changes in the foreign market, mergers and acquisitions of companies, natural or induced catastrophes and other economic factors. Another factor that causes oscillation of stock prices is market speculation. [5] Investors have the difficult task of continually monitoring market fluctuations so that their strategies result in better financial returns. With *Big Data* you can automatically collect various information from the Internet, such as political and economic news, monitor real-time stock price changes, and many other relevant news items that can influence financial indicators and stock prices. A model widely used for predictions is the *Bayesian* model, which is based on the *Bayes*' theorem. *Naive Bayes* uses a probabilistic classifier [1] with *Bayes* rules to make predictions.

With the *Naive Bayes* classifier [2], it is possible, for example, to predict the values of the London stock exchange index FTSE100. The natural processing language contributes to analyze direct and indirect states if the stock prices of texts extracted from the Internet will rise, fall or remain at the same value. Analyzing the feeling of the news [2] involves recognizing and defining the "emotional state" expressed in the text.

This work contributes to the use of the *Naive Bayes* classifier to predict the *Ibovespa* stock market, especially in the use of *Big Data*, little explored in the bibliography raised helping an investor in financial decisions presenting the best moment to buy and sell stocks. In addition, we will deepen prediction studies with *Naive Bayes*, use *the Naive Bayes* language and *R* statistics environment, classify texts for the process of predicting stock prices with the natural processing language and use of *Big Data* and that can serve as a tool for financial analysis or even adapted to another segment.

## II. THEORY OF BAYES

*Bayes*' theorem has been used in areas such as finance, biology, vehicle insurers, fraud-avoidance systems, and others in which probabilistic predictions are desired. *Bayes*' theorem is based on conditional probability and joint probability. For the realization of probabilistic predictions will be used the *Naive Bayes*, which is a probabilistic classifier that uses the *Bayes* rules.

$$P(c|x) = \frac{P(x|c).P(c)}{P(x)}$$

Formula 1. *Naive Bayes* Classificator [1]

In the Formula 1:
P(c|x) represents posterior probability
P(x|c) represents likehood
P(c) represents class prior probability
P(x) represents predictor prior probability

*Naive Bayes* uses *Bayesian* learning to train classifiers. It is possible to improve the prediction of the hypotheses with the training of the *Naive Bayes* classifiers.

For the calculation of the prediction, three tables, data table, frequency table and probability table are required. The problem attributed *to Naive Bayes* as the main question is the prediction of the value of the Abev3 stock is: Will the Abev3 stock price for the fourteenth day rise in the period of analysis?

| Date | Abev3 | Will Up? |
|---|---|---|
| 03/05/2016 | 19.43 | Keep |
| 04/05/2016 | 19.32 | No |
| 05/05/2016 | 19.33 | Yes |
| 06/05/2016 | 19.02 | No |
| 09/05/2016 | 18.77 | No |
| 10/05/2016 | 18.93 | Yes |
| 11/05/2016 | 19.00 | Yes |
| 12/05/2016 | 19.02 | Yes |
| 13/05/2016 | 19.05 | Yes |
| 16/05/2016 | 18.74 | No |
| 17/05/2016 | 18.59 | No |
| 18/05/2016 | 18.59 | Keep |
| 19/05/2016 | 18.51 | No |
| 20/05/2016 | 18.59 | Yes |

Table 1. Abev stock prices

| Abev3 | Yes | No | | |
|---|---|---|---|---|
| Up | 5 | | 5/13 | 0.38 |
| Down | | 6 | 6/13 | 0.46 |
| Keep | 2 | | 2/13 | 0.15 |
| Total | 7 | 6 | | |
| | 7/13 | 6/13 | | |
| | 0.53 | 0.46 | | |

Table 3. Probability

| Abev3 | Yes | No |
|---|---|---|
| Up | 5 | |
| Down | | 6 |
| Keep | 2 | |
| Total | 7 | 6 |

Table 2. Frequency

Tables 1, 2 and 3 were created by the author. In the data table, until record number 13 represents the prices of the stock Abev3 collected from 05/05/16 to 05/19/16. Record number 14 in red represents the result of the prediction. If the current quotation value is equal to the previous value, the column was filled with the result "Hold", if the current quotation value is greater than the previous value, the

column was filled with the result "Yes" and if the value of the current quotation is less than previous value, column was filled with result "No". By the occurrences of yes and no, the tables of frequency and probability are constructed. Substituting the values in the *Naive Bayes* classifier formula results in a 97% probability that the record 14 is positive for the prediction. These calculations are performed automatically by the E1071 language library and *R* statistics environment by the historical price of the previously collected stocks.

$$P(Yes|Will\ Up) = \frac{P(Will\ Up|Yes) * (P(Yes)}{P(Will\ Up)} = 0.97$$
$$P(Will\ Up|Yes) = 5/7$$
$$P(Yes) = 7/13$$
$$P(Yes|Will\ Up) = 5/13$$

Formula 2. Predictive calculation using the *Naive Bayes* classifier. Source: Author.

The prediction with *Naive Bayes* only tells you whether the stock will go up, down, or hold. For the calculation of the value is shown in the formula 3.

$$Prediction\ result = Current\ Value + (Current\ value - Penultimate\ value)$$
OR
$$Prediction\ result = Current\ Value - (Current\ value - Penultimate\ value)$$

Formula 3. Prediction calculation. Source: Author.

If the prediction is positive, one must add the difference of the last value by the penultimate value and divide by two. If the prediction is negative, one should subtract the value of the share by the penultimate share value and divide by two. After performing the calculation, the result is inserted from the *Dataframe*.

## III. Comparison Of The *Naive Bayes Classifier, K-Nearest Neighbor, Support Vector Machine*

In the experiment [7], the *Naive Bayes* classifier was compared to other *K-Nearest Neighbor* (KNN) classifiers, *Support Vector Machine* (SVM). The three classifiers were subjected to tests of data traffic on the Internet in different types of services. The following services were analyzed: *http, smtp, ftp, pop3, nntp, msn, edonkey, ssh, bittorrent and https*. For the experiment, a total of 13,536 data streams with 50 flows of training classes for the 3 classifiers were used.

| Services | KNN | SVM | Naive Bayes |
|---|---|---|---|
| http | 90.9 | 99.85 | 94.1 |
| smtp | 97 | 85.95 | 91.95 |
| ftp | 96.35 | 87.05 | 95.4 |
| pop3 | 90.35 | 78.3 | 97.35 |
| nntp | 98.2 | 97.95 | 97.05 |
| msn | 98.36 | 80.6 | 98.36 |
| edonkey | 93.8 | 66.34 | 91.58 |
| ssh | 93.5 | 72.4 | 98.37 |
| bitorrent | 92.15 | 84.9 | 96.25 |
| https | 80.6 | 59.91 | 89.65 |
| **Media** | **93.18** | **81.32** | **94.4** |

Table 4. Comparison of individual efficiency of *KNN, SVM and Naive Bayes* classifiers. [7]

Table 4 presents a comparative of individual efficiency of the classifiers *KNN*, *SVM* and *Naive Bayes*. The author did not provide further details, but the average results were 93.18% for *K-Nearest Neighbor*, 81.32% for Support *Vector Machine* and 94.4% accuracy for *Naive Bayes* with the highest score.

## IV. *Big Data* And Natural Language Processing

Currently the high volume of data such as navigational records, texts and documents, commercial transactions, bank records, financial charts, medical images, surveillance videos, marketing, telecommunications, social media data can be easily collected or generated from different sources, in different formats in different real-time applications in the *Big Data* era. [3]

Many of the data is collected by sensor networks, giga pixel cameras, Internet of Things, surveillance and monitoring devices and social media by multiple vendors. The data needs to be cleaned to be used. [6]

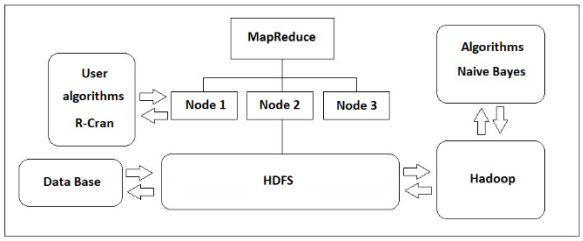


Fig. 1. Example of *Big Data* architecture with *Hadoop*, *R-Cran* and *Naive Bayes*. [4]

The texts collected by *Big Data* do not present useful information by themselves. It is necessary to extract information from collected texts that help predict the values of actions.

The natural processing language allows the texts extracted by *Big Data* to be analyzed and classified, allowing a positive and / or negative perception of the news, indicating if the value of the stock will rise or fall. This perception happens in real time and automatically, adjusting the value of the prediction and presenting graphically.

With this real-time adjustment, the system seeks to keep up with the various fluctuations in the financial market, bringing as close a prediction as possible to reality and assisting the investor.

## V. Evaluation Method

The objective of the proposed research is an exploratory research. The exploratory research aims at providing the researcher with a greater knowledge about the subject or problem under analysis. Therefore it is appropriate for the beginning of the investigation when the knowledge, the understanding and familiarity of the researcher are insufficient or nonexistent. The work method was divided into 5 phases. Being:

| Phase 1 | Phase 2 | Phase 3 | Phase 4 | Phase 5 |
|---|---|---|---|---|
| Bibliographic research | Part 1<br>News Collection<br>Part 2<br>Collecting stock prices during the electronic trading session | Experiment Specification | Part 1<br>Prediction test of the data collected in phase 2 part 2<br>Part 2<br>Prediction tests of the data collected i phase 2 part 1 | Analysis of result |

Table 5. Phases of the working method. Source: Author.

Phase 1: In this phase, research was conducted on books and articles related to the prediction of the London stock exchange [2], and the use of the natural processing language [8] to classify the information extracted from the sites. These references are used to justify this work.

Phase 2: Two types of data were used in the experiments: collection of information extracted from financial sites and collection of stock price values during trading hours of the Ibovespa, from Monday to Friday, from 10 to 17: 00 hour by minute.

Phase 2 part 1: The first type of data, refers to the collection of information from different financial sites, news about the company analyzed, economy, policy and other relevant news. The objective of this collection is the classification of texts using the natural processing language and that can impact on stock prices.

Phase 2 part 2: The second type of data to be collected is the quotation, in short periods of one minute, of the share price of a company listed on the *Ibovespa*. The objective of this collection is to observe the variations in the price of the stock during the trading session. This information will be stored in a database and will compose a history for the performance of the prediction of the share price.

The company chosen for analysis is Companhia de Bebidas das Américas (Ambev), whose share code on *Ibovespa* is represented by Abev3. Ambev is a Brazilian publicly traded company that produces consumer goods. In addition to being a solid and well positioned company in the Brazilian and international markets, the collection of information of classification of texts will be facilitated by being a multinational company with a good reputation.

Phase 3: For the data collection of phase 2, part 1, the *Big Data Hadoop* is used in a virtualized environment with 2 nodes, being 1 master computer and 1 slave computer. *MapReduce* is responsible for the process of getting the pieces of information on the slave computers, processing them, and joining them on the master computer. Data from this collection is stored in the *Hadoop Distributed File System* (HDFS) or unstructured *Hadoop database*.

For the data of item 2, part 2, a *Java* program was developed to capture the values of the actions and to insert them into a structured database.

After the information is collected in item 2, it is possible to use it in a test environment. This environment consists of the master computer being properly installed and configured the *R-Studio* application, in which it is possible to use the language and *R* statistics environment in the client / server environment.

With the language and statistics environment *R* configured and enabled the *Naive Bayes* E1071 package and its *Naive Bayes* and Predict functions it is possible to perform the predictive mathematical calculations with the collected data.

The purpose of using *Big Data*, R-Studio, and language and *R* statistics environment is to provide an appropriate environment for collecting, processing, analyzing and presenting the best possible predictions for an investor.

Phase 4: The tests were divided into two parts.

Phase 4 part 1: In this phase the prediction tests will be carried out with the variations of the stock price history with *Rstudio.*

Phase 4 part 2: In this phase will test the texts extracted by *Big Data* for analysis of the classification of texts in a program developed in *Python* with the natural processing language.

Phase 5: In this phase the results of the analysis obtained after the experiment will be presented. The results presented by the prediction process will be analyzed and compared with the value obtained from the actions during the electronic trading session and analyzed its measurement. If the verification of the predictive process presents many different results from the values of the electronic trading session, its possible causes will be investigated and improvements will be proposed for a better gauging of the results.

The metric proposed for this work is divided into two parts:

1. Average percentage of the prediction reached in relation to the real value of the share.
2. Percentage of correct classification of texts.

For the calculation of item 1, all the records of the real prices and the prediction will be available in tables. The average daily price of the two tables will be calculated. The calculated percentage of the prediction will be calculated in relation to the actual price of all registrations. After that, all values will be counted, where the percentage of the prediction reached less than and equal to 1% positive and negative.

For the calculation of item 2, all texts collected by *Big Data* will be analyzed manually and compared to the result of the natural processing language generating a percentage of correctness.

## VI. Experiment

For the experiment, a virtualized environment architecture was created with a main and secondary computer for Apache *Hadoop* 2.7.3, 1 GB RAM, 1 processor and language and environment for statistical computation *R* for prediction generation. For the extraction of stock prices on the Internet, a Java program was developed, in which the site is captured in *HTML*, extracted information and stored in *MySQL* database.

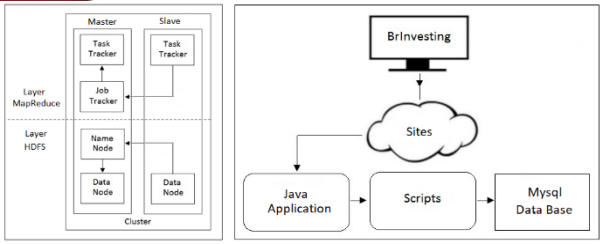


Fig. 2. The left is presented the *Hadoop Big Data* architecture with main and secondary computer and the right is presented the data extraction architecture for collecting Internet stock prices. Source: Author.

The collection of price data began on 03/05/2016 to 03/29/2017 or 374 days of records that store stock values from minute to minute of the company abev3 during the electronic trading session Monday to Friday 10:00 a.m. to 5:00 p.m.

The prediction is performed in the R-Cran daily according to figure 1, always until the day before the electronic trading, generating the prediction for the next day.

With the Naive Bayes and Predict functions of the statistical tool R, a predictive chart of the company's Ambev stock prices was created the next day, analyzing the history to its previous day, as shown in figure 3. In this graph it is possible to identify the maximum price and minimum that the value of the action will reach, and also its best moment for buying and selling.

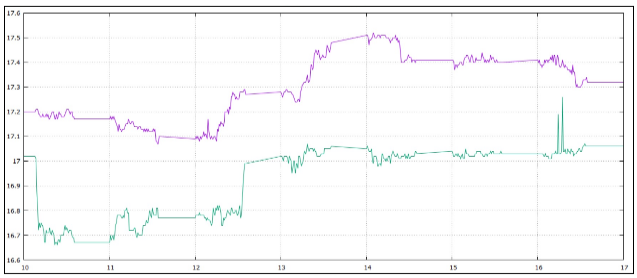

Fig. 3. Ambev's predictive and real stock price chart on 01/18/2017. Source: Author.

The purple line represents the stock prices obtained during the electronic survey, and the green line represents the prediction made before and before the trading day. The prediction in the example, was about $ 0.20 cents below the real price throughout the day. This difference in price is minimized by using *Big Data.*

In *Big Data* and with the use of *Hadoop Streaming* a program was developed in Python for the use of Natural Processing Language (NPL).

Data collection from financial sites started on Jan 30, 2017 until 03/29/2017 and has 21,778 texts with a total size of 2.9 Gbytes of occupied space. News from 07 financial sites was stored at 15-minute intervals.

Texts taken from financial news sites need to be prepared for NPL to be able to identify whether prices will change during the trading session. This process includes converting all uppercase to lower case, removing punctuation, removing numbers, using stopwords or stop words, typing texts and creating a specific dictionary to identify key words that indicate the conditions of the stock.

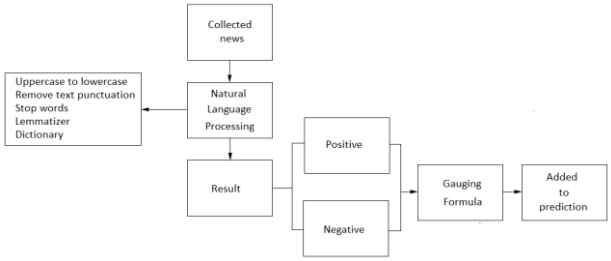


Fig. 4. Data flow from news gathering to prediction. Source: Author.

Figure 4 shows the data flow of the collection until it is added in the prediction.

After processing the LPN, a positive or negative result will be displayed informing if the stock prices will rise or fall. If the result is positive, the result of the formula of measurement of figure 3, if negative subtracted, will be added. Finally the result is added to the prediction.

$$\frac{total\ words}{total\ words\ .total\ documents}$$

Formula 4. *Big Data* price reference formula. [8]

In figure 5, the blue line as a result of *Hadoop*'s word processing, approached 5% more of the real value of the stock price (purple line).

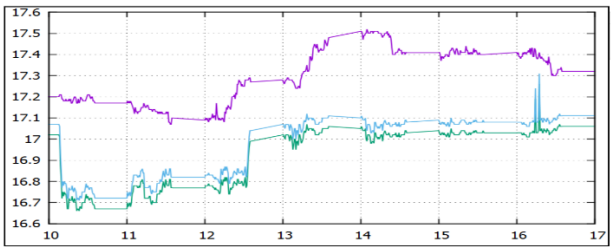

Fig. 5. Ambev's predictive and real stock price graph and the inclusion of the Hadoop result on 01/18/2017. Source: Author.

At that moment the analysis of the feelings of investment analysts is not being carried out because there are many speculative factors that are published freely on the Internet and often do not express reality. Usually reliable opinions from financial experts are not published for free on the Internet, usually paid services and offered by banks and brokers.

## VII. CONCLUSIONS

Big Data is inefficient to handle large amounts of small files up to 15 kbytes. Multiple files can be concatenated into a single file with the merge function of Hadoop. By processing a single file, there will be a speed gain.

Some financial sites on which the news was extracted were taking more than fifteen minutes to update, which

undermined the dynamism of the system and the adjustment of the real stock price by the classification of texts.

The prediction in the overall mean performed at 374 measurement days remained close to the real value of the stock. The red line in the graph of Figure 7 represents the average of all values for each day of the actual value of the action, and the green line represents the prediction reached.

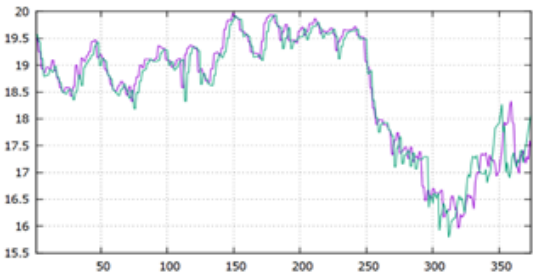


Fig. 6. Graph of the predictive and real mean of 374 days. Source: Author.

- Total records: 158 202
  Total of measured days: 374
- Average prediction hits considering up to 1% above or below the actual value of the share: 55.92%
- Average of the prediction in relation to the real value of the share: remained 0.20 cents above the real value
- Maximum value of the average of the variation of the difference between the real and the prediction: from 1.31 and -0.96
- Prediction value equal to the real value: 2 443 hits or 1.54%

The classification of texts with the natural processing language obtained 48.34% or 10 527 of 21 778 correctly classified texts. There are various forms of writing and sense of phrases extracted from sites that have often undermined the correct result of classifying texts. Phases like "Ambev had a 2% increase." Are identified correctly. Phrases like "Ambev closes away from high maximum" returned incorrect results.

The system met the goal by providing a prediction close to the actual value. Big Data and natural language processing helped improve prediction by approaching real value. There is no certainty that a history will always repeat itself and each day is a different chart.

Although this article presents the research using only one company, this study can be applied to other member companies of the stock exchange or even in other segments.

## REFERENCES

[1] R. Sandhya ., S. Sonali ., K. Siddhant., B. Deepti., Experiments on Content Based Image Classification using Color Feature Extraction. Communication, Information & Computing Technology (ICCICT), IEEE International Conference. Mumbai. 2015. p.1-6

[2] A. Shihavuddin., S. M., Prediction of Stock Price analyzing the online financial news using Naive Bayes classifier and local economic trends. Advanced Computer Theory and Engineering, Volume:4. 3rd International Conference, Chengdu. 2010. p.V4-22-V4-26

[3] C. Leung C. K., F. Jiang., A Data Science Solution for Mining Interesting Patterns from Uncertain Big Data. IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, NSW 2014. p.3-5

[4] X. Brian, A. Sathish K., Big Data Analytics Framework for System HealthMonitoring. IEEE International Congress on Big Data.USA 2015. p.408

[5] B. Graham. O investidor inteligente. Rio de Janeiro, Nova Fronteira, 2015. p.32

[6] Rao D, Gudivada Venkat, Raghavan Vijay, Data Quality Issues in Big Data. Big Data Conference. IEEE International Conference on Big Data. North Carolina USA. 2015. p.2654-2660.F. Mattar, N. Pesquisa de Marketing. São Paulo, Atlas, 1994.

[7] F. Ghofrani., A. Haddad K., Jamshidi A. Internet Traffic Classification Using Multiple Classifiers. IEEE 7th International Conference on Information and Knowledge Technology. 2015. p. 1-5

[8] Y. Kim, S. Jeong R, I. Ghani, Text Opinion Mining to Analyze News for Stock Market Prediction. Int. J. Advance. Soft Comput. Appl., Vol. 6, No. 1, March 2014.

## AUTHORS' PROFILES

**Hideo Hayashi André**, Bachelor in Computer Science, Post-Graduate in Computer Networks, Post-Graduate in Information Security, Post-Graduate in Project Management, MSc in Software Engineering by IPT. He has more than 15 years I.T. area acting in Management, Infrastructure and Systems. Work in Telecom and I.T. companies. Currently Manager of I.T. in São Paulo, Brazil.

**Barbieri J. Sousa Alexandre**, 20 years of experience in service strategy and delivery. Barbieri has developed a reputation for leading and motivating senior teams. Graduated in Computer Engineering, with a Master degree in the same subject and a Doctor degree in Electric Engineering by USP-POLI. Actually, works as Senior Service LATAM Leader (IT, Internet, Datacenter) for Banks, Service Provider and Enterprise, leading Professional Services Consulting organization for LATAM region. Alexandre Barbieri speaks English, Portuguese, Spanish. email id: abarbieris@hotmail.com