Big Data for Stock Market by Means of Mining Techniques

Luciana Lima, Filipe Portela, Manuel Filipe Santos,
António Abelha, and José Machado

Algoritmi Research Centre, University of Minho, Portugal
a58315@alunos.uminho.pt, {cfp,mfs}@dsi.uminho.pt,
{abelha,jmac}@di.uminho.pt

Abstract. Predict and prevent future events are the major advantages to any company. Big Data comes up with huge power, not only by the ability of processes large amounts and variety of data at high velocity, but also by the capability to create value to organizations. This paper presents an approach to a Big Data based decision making in the stock market context. The correlation between news articles and stock variations it is already proved but it can be enriched with other indicators. In this use case they were collected news articles from three different web sites and the stock history from the New York Stock Exchange. In order to proceed to data mining classification algorithms the articles were labeled by their sentiment, the direct relation to a specific company and geographic market influence. With the proposed model it is possible identify the patterns between this indicators and predict stock price variations with accuracies of 100 percent. Moreover the model shown that the stock market could be sensitive to news with generic topics, such as government and society but they can also depend on the geographic cover.

Keywords: Text Mining, Stock Prediction, Big Data.

1 Introduction

Currently the business and consumer's behaviour is changing faster than ever, turning the market even more unpredictable. Facing this, organizations must assess the alternative business strategies and implement them with optimal technology business solutions [1]. Thus, raises a new challenge, organizations are now in front of a huge amount of data but they do not know how to get value from it. Most of information is obtained from raw or unstructured data, being hard to know how to recognize what is relevant and how to interpret it [2].

In order to explore one of the several Big Data business opportunities in Financial Services Industry, this project focus is the relation between daily news articles and the market variation. The main purpose is predict the stock market variation based on the events stated in daily releases.

News articles written about companies serve the purpose of spreading information about them. This information can influence people either consciously or unconsciously in

Á. Rocha et al. (eds.), *New Contributions in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 353, DOI: 10.1007/978-3-319-16486-1_67

their decision process when trading in the stock market [4]. Consequently, the impact it will be bigger when a given information is unexpected by the companies or investors.

This paper addresses Big Data and the forecasting of Stock markets by means Text and Data Mining. It introduces a proposal to improve the existent stock work prediction with news sentiment analysis. We believe that this process can be enriched with other relevant variables that can be extracted from the news articles.

Identifying relevant indicators to add the sentiment analysis and measuring the document impact with higher accuracy leads to more valuable predictions. Sentiment analysis algorithms are well developed and have shown great results. However the "sentiment "of an article can be classified with some ambiguity. If sharing the sentiment weight with other indicators such as the market (geographical) that the news "covers" and the issue that is being addressed to it (government, society, company) probably the prediction accurateness will increase significantly.

Following this proposal, it was designed a knowledge discovery engine and it was induced a set of Data Mining Models. As main results it was possible predict the stock price variations with accuracies of 100% to a particular case of USA stock market.

The document is divided in seven sections, the introduction followed by the concept background and related work. The third section presents the materials and used methods and the fourth shows the big data architecture and data solution. Next, in the section five it is explained the developed modelling and the achieved results in the section six. The last section presents the conclusions of the project

2 Background and Related Work

2.1 Big Data

The way of society and the people communicate each other changed the way how data is produced and consumed. In an organizational point of view, the information, as an important asset, is generated and stored in big amounts (rounding Zettabytes). Organizations collect large volumes of information of their clients, suppliers, operations and millions of sensors linked to the network and implemented in the physic world in devices such as mobile phones, computers, cars, etc. [3].

Big Data came as a dataset with a size and complexity that is beyond the ability of conventional tools of managing, storing and analysing the data [3]. Now, in this context, data also includes large amounts of structured, semi-structured and non-structured schemas that can be collected from call logs, social networks, weblogs, emails and documents. In addition, there are other sources of unstructured data that continuously spew digital exhaust and contribute to what is known as 'Big Data'– blogs, online news, weather, Twitter, YouTube and Facebook.

2.2 Big Data in Financial Market

The banking and financial management business is rife with transactions, conducting hundreds of millions daily, each adding another row to the industry's immense and growing ocean of data [5]. In order to better understand what is forcing Big Data technology adoption in Financial Services, the Oracle white paper [6] detailed some drivers that increased the need of Big Data architecture: Costumer Insight, Regulatory Environment, Explosive Data Growth and Technology Implication.

The Financial Service sector, the focus of this project, is one of the most data-driven industries and most of the data that exists within a bank's datacenter is not analyzed [6]. Financial services organizations are leveraging Big Data to transform their processes, their organizations and soon, the entire industry. Big data is especially promising and differentiating for financial services companies. With no physical products to manufacture data, the source of information became one of their most important assets.

2.3 Text and Data Mining

Text mining (TM) is known as text data mining or knowledge discovery from textual databases. TM is the process of discovering hidden useful and interesting pattern from unstructured data [7]. This concept is very similar to Data Mining (DM), defined as extraction of knowledge, hidden predictive information from huge amounts of data. With DM technologies is possible to find patterns and trends on large relational databases [8]. According to Ah-Hwee Tan [7] TM can be visualized as consisting of two phases: text refining and knowledge distillation. The process starts with a pre-processing phase where the text is Cleaned, Tokenized, and it is analyzed Part of Speech Tagging (POS) which means assign word class to each token. Finally, using DM algorithms the results are interpreted and evaluated.

At a DM level this work is recognized as a classification problem. Classification assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

2.4 Related Work

In the last few years data mining and prediction algorithms have been an appeal subject for the most several industries. The basis of this analysis is the text mining and sentiment extraction of the articles, which means, define if they have a positive, negative or neutral impact to the reader, in this case companies or investors. Trying to predict the exact variation in a shorter period after the article released is the main focus.

The AZFinText System [9] combines several textual representations: Bag of Words, Noun Phrases, and Named Entities to analyses the articles with past stock pricing information. Other works filtered the news articles according to their relevance to a specific company or market [9]. The impact can be also predicted by the behavior of the stock prices during the time window of influence of the article [10] According to Aase [4] it is possible to predict stock price changes after news articles publications. When the stock trading is done from signals generated from sentiment analysis of news articles, then the profit is better compared to what a random trader gives. A training set of news articles for the sentiment classifier might be automatically created and labeled by looking at how the price of a certain company changes after the article is published.

3 Materials and Methods

Although the project was supported by Deloitte Consultores SA, the target for the study was more generic because the access to the Financial Industry transactional data was considered limited. The goal was to use what is available on the web to explore some of the concepts above presented. For that reason it was decided develop a workflow that will analyze news articles and their impact over the stock market. The idea is follow some of the opportunities suggested for the industries and apply some of the insight from the earlier investigation.

News Articles Influence on Stock Market was the chosen scope and it is based on the Kim-Georg Aase work [4]. The main purpose is to find patterns between the financial news articles and the stock value variation in the Financial Services Market.

The project management and development followed two methodologies, SCRUM methodology and a pipeline called "The Big Data analysis pipeline" [11]. This pipeline presents a set of phases for Big Data projects that goes from data acquisition to interpretation of results. To achieve this goal it was designed architecture able to support the use case and text mining techniques to the data received. After process the data it was induced several data mining models using Knime engine and Weka module.

4 Architecture and Data Sources

In order to support this use case, it was defined an architecture (Fig 1) based on the literature review and the analysis of the existent tools in the market. The established architecture supports two data sources, non-structured (news web site) and structured data (NYSE stock price history). The data from the news web sites was collect using a web crawling provided by KNIME. In the Storage Layer each type of structure is deposited into a specific database. Articles news were storage into Hbase managed by Cloudera. The historic stock price was stored into a relational database from Teradata.

Finally, these data were submitted to some transformations in order to be analyzed and accessed by the users.

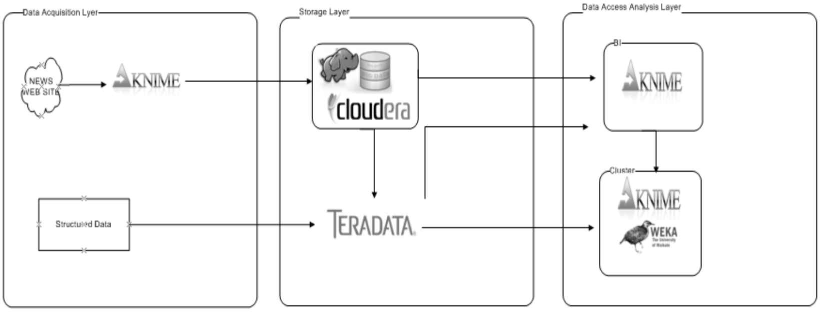


Fig. 1. Use Case Architecture

Taking into account the business requirement, it was predicted stock variation based on news articles. The data sources, filters and technique to use were defined. The unstructured sources are the economics and general news websites such as the ones extracted using the Web Crawl technique:

- http://www.reuters.com/
- http://www.washingtonpost.com/
- http://www.businessweek.com/

The news websites mentioned above were chosen by their popularity combined with their releases, economics and general subjects.

The structured source, also from the web, was the NYSE Stock Exchange which provides the stock value history from every company quoted on the stock exchange.

As use case it was chosen a specific analysis - the USA market, Bank and Energy Industries. As result part of this process, the data stored into Hbase is column-oriented and presents the different schemas defined on Knime. The attributes in the column-family File of News Crawl table were the follows:

- **Summary:** Summary of the article retrieved according to Google's API criteria.
- **Date:** Date of publication.
- **HTML:** Html code from the web page collected.
- **Title:** Title of the news article
- **URL:** Source of the article

At the same time, structured data was imported from csv files with the stock exchange history and has the following attributes:

- **ID:** Key of the company.
- **Symbol:** Symbol of the company.
- **DateSE:** Date of the stock exchange release.
- **OpenSE:** Daily opening value of the stock exchange.
- **HighSE:** Highest stock value of the day.
- **LowSE:** Lowest stock value of the day.
- **CloseSE:** Daily closing value of the stock exchange.
- **VolumeSE:** Daily total volume of stock.

5 Modelling

5.1 Extraction/Cleaning/Annotation

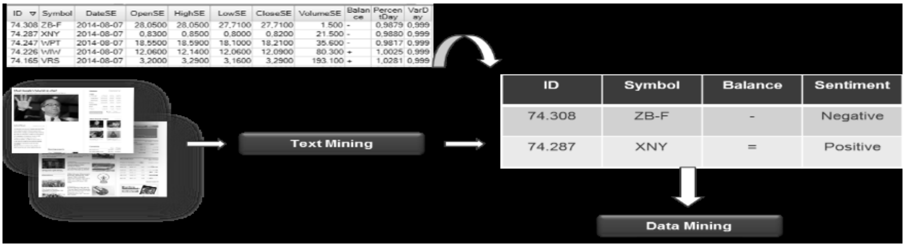
In this phase the objective was extracting the interest content (cleans it from all the HTML and JavaScript code and registers the categories of the particular news). Then the entities are filtered and passed through the Keygraph Keyword extractor [12] [13]. The tags were converted into string and pass through an inference rule engine defining the tags from a dictionary with -1 and the other with 1.The market, company and sentiment indicators are calculated with the mean of the tags score. The result is a dataset containing each article categorized according the mean of it correspondent scored tags.

Therefore, these categories are: **Market** (If the news refers to the local (USA) or foreign market), **Company** (If the news refers to a specific company (from our chosen list) or other subjects such society and government) and **Sentiment (**If the news could be classified as Positive, Negative or Neutral).

In the engine results, most of the values are absolute, for example, the document has a positive or negative sentiment. However in some of the documents the resulted were inconclusive, with the decimal values close to 0. For those values, it was decided round to 0 in order to represent a neutral sentiment.

5.2 Integration/Aggregation/Representation

In order to analyze and predict the impact of news articles in a set of companies' stock exchange, it was critical integrate those two information sources. The process of merging different types of data can improve the data analysis and make it more interesting. The approach for the integration process (Fig 2) was collected and transformed, separately both structured and unstructured data. From each source it was stored the processed information. Finally the both data types were merged and prepared to be analyzed.



ID	Symbol	DateSE	OpenSE	HighSE	LowSE	CloseSE	VolumeSE	Balance	PercentDay	VarDay
74.308	ZB-F	2014-08-07	28.0500	28.0500	27.7100	27.7100	1.500	-	0.9879	0.999
74.287	XNY	2014-08-07	0.8300	0.8500	0.8000	0.8200	21.500	-	0.9880	0.999
74.247	WPT	2014-08-07	18.5500	18.5900	18.1000	18.2100	35.600	-	0.9817	0.999
74.226	WIW	2014-08-07	12.0600	12.1400	12.0600	12.0900	80.300	+	1.0025	0.999
74.165	VRS	2014-08-07	3.2000	3.2900	3.1600	3.2900	193.100	+	1.0281	0.999

ID	Symbol	Balance	Sentiment
74.308	ZB-F	-	Negative
74.287	XNY	=	Positive

Fig. 2. Text Integration Approach

In this phase it was also integrated the stock exchange history with news articles structured information. This integration was made by the date reference, in other words, it was assumed a daily time window of influence. Each news article will be related with the all stock exchanges provided they are between the opening and closing stock market. The final dataset had the following scheme: **ID**; **Symbol**; **DateSE**; **Sector(**Industry of the company); **SubSector(**Sector inside of industry); **First (URL) (**URL of the news article); **VarDay(**Average of PercentDay variation grouped by Symbol);**OutOfRange(**Identifies which company in certain day has its Percent of variation different from the average of VarDay); **Market(**Text Mining Category);**Company(**Text Mining Category);**Sentiment(**Text Mining Category).

5.3 Data Analysis

Before looking at the predicted data it was important to know well the dataset and their particularities. At this point it was possible to conclude the following. More than half of the companies show that at the end of the day the price was lower than at the opening, being more precisely, 51% of that stock had a negative balance, 36% a positive balance and 2% remains the same price at the end of the day.

In the distribution of the news articles sentiment, most of the news, 73%, were classified as negative because the words with negative meaning are more expressive and easier to “find”. The rest was classified as neutral (without any impact), 20% and 7% as positive. Also, 51% of the documents are related to news from USA cities, but almost 100% of the news does not identify any listed company. This means that the subject was mostly related to government and society issues.

5.4 Data Mining Model

In this phase it was induced DM models in order to discover if there was a pattern between the news articles and the stock exchange variation. To perform the classification algorithms they were used several Weka nodes available in the Knime tool. Also, from the entire dataset, it was used 70% for training and 30% for test.

To better understand, the DM workflow contained the algorithms nodes for each scenario. After the algorithm node it is performed the Weka prediction node will use the result model into the test data set. After the prediction it was used a scorer node to analyze the confusion matrix and accuracy statistics. By option, the dataset was trained with four classification algorithms provided by Weka. These four algorithms were chosen based in their success level in similar studies and they were: J48, Multilayer Perceptron, LibSVM and Naïve Bayes.

5.5 Scenarios

To perform the classification algorithms, some scenarios with different variables combination were defined. The Balance, the variable data which indicates either the stock value increased or decrease, is the target of the model. The other variables will help the algorithm to find a pattern between the news article indicators and the target. For testing, they were explored five different scenarios (Table 1), resulting in a total of 20 Models (5 scenarios x 4 techniques x 1 target).

Table 1. Test Scenarios

	Scenario	**1**	**2**	**3**	**4**	**5**
Variables	Target	Balance	Balance	Balance	Balance	Balance
	ID	X	X	X	X	X
	DateSE	X				
	Balance	X	X	X		
	Percent	X				
	Sector	X			X	
	SubSector	X	X			
	Varday	X				
	Market	X	X	X	X	
	Company	X	X	X	X	
	Sentiment	X	X	X	X	X

6 Results

6.1 Performance Measures

For the classification techniques, the performance is evaluated through the analysis of the capability that the model has to correctly predict a value. The confusion matrix (Table 2) is the basis for several measures and presents the number of correct classifications *vs* predicted classifications for each class.

Table 2. Confusion Matrix (example)

Target	True Positives	False Positives	True	False
-	12589	292	7596	0
+	7308	4	13034	131
=	268	16	20012	181

According to the relation between True/False Positives and True/False Negatives it was possible to measure the performance as follows: Precision, Sensitivity, Specificity, F-measure, Accuracy and Cohen's Kappa.

To complement the evaluation, the Receiver Operating Characteristic (ROC) curve was also designed. ROC can be plotted as a curve on an X-Y axis. The false positive rate is placed on the X axis. The true positive rate is placed on the Y axis [14].The area under the ROC curve (AUC) measures the discriminating ability of a binary classification model. The larger the AUC is, the higher the likelihood is.

6.2 Algorithm Evaluation – Best Results

After executing each Weka node in the Knime interface, Table 3 presents the top 3 of classification results:

Table 3. Classification Algorithm Results

Scenario	Algorithm	Class	Precision	Sensitivity	Specifity	F-measure	Accuracy	Cohen's Kappa
1	J48	"-"	1	1	0,999	1	1	0,999
		"+"	1	0,999	1	1		
		"="	1	0,998	1	0,999		
1	LibSVM	"-"	0,989	0,991	0,982	0,990	0,987	0,973
		"+"	0,983	0,982	0,999	0,982		
		"="	0,990	0,929	1	0,959		
5	LibSVM	"-"	0,999	0,999	0,997	0,999	0,998	0,997
		"+"	0,998	0,998	0,999	0,998		
		"="	1	0,980	1	0,998		

Analyzing the measures chosen it was possible to verify which it was the best model capable of predicting the three classes with a high precision. The best model was induced using the classifier J48 for the scenario 1 which presents the highest accuracy value. Just reminder that the scenario 1 predicted de balance target using all existent variables. Although there are great models with strong precision in some predictions, this is the model that will be used to perform data analysis with the news prediction.

After submitting the training algorithm is possible to see the resultant prediction made to a test dataset (Table 4). Here the objective is to produce some statistics, in order to get to know the model better and perform some others improvements.

Table 4. Predicted Data Set Scenario 1 J48 algorithm (example)

ID	DateSe	Balance	PercentDay	Sector	SubSector	Varday	Market	Company	Sentiment	Prediction (Balance)
3.487	14-07-09	-	1.003	Energy	Gas	0.997	1	-1	-1	-
19.719	14-07-16	-	1.013	Bank	Bank	0.997	1	-1	-1	-
3.489	14-07-09	+	1.003	Bank	Bank	0.997	-1	1	1	+
3.487	14-07-16	-	1.013	Energy	Gas	0.997	1	-1	0	=

7 Conclusion

This use case was a practical example of the capabilities attached to Big Data. The objective was to be aware of the existence of other types of data, with the challenge of volume and velocity required. At the end of this use case, it was possible to conclude that, actually, merging the sources and types makes the data analysis more interesting and complete. The analysis of text content is very complex, it requires an extra effort for the machine to identify the characteristics in a document, such as sentiment with the same sensitivity of a human (being).

In the classification of the news, according to its sentiment, it was defined a set of words and combinations of positive and negative words. Luckily, in most cases, those words were enough to classify the sentiment but, at the same time, the occurrence of “undefined” classification and the false classification indicate that the dictionaries must be more embracing and capable of detecting the particularities of natural language like context and double meaning and, in a more advanced form, they should be capable of detecting irony. For that reason, we believe that the stock exchange prediction is better when combined with the history stock exchange, not only with the sentiment of an article but also with other indicators that could be extracted and, in some manner, determinant for the market variations.

Concluding and in an abstract perspective, general subjects like politics, health, governance, war or even climacterics changes or tragedies could also bring some impact for the different industries and influence the stock market.

Acknowledgements. This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

References

[1] SanthoshBaboo, L., RenjithKumar, P.: Next Generation Data Warehouse Design with Big data for Big Analytics and Better Insights. Glob. J. Comput. Sci. Technol 13(7) (February 2013)

[2] Lima, C.A.R., de Calazans, H.C.J.: Pegadas Digitais:'Big Data' E InformaÇão Estratégica Sobre O Consumidor (2013)

[3] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity (2011)

[4] Aase, K.-G.: Text Mining of News Articles for Stock Price Predictions (2011)

[5] Turner, D., Michael, S., Rebecca, S.: Analytics: The real-world use of big data in financial services (2013)

[6] Mathew, S., Halfon, A., Khanna, A.: Financial Services Data Management: Big Data Technology in Financial Services (2012)

[7] Tan, A.-H.: Text Mining: The state of the art and the challenges (2000)

[8] Linoff, G.S., Berry, M.J.A.: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd edn. Wiley, Indianapolis (2011)

[9] Drury, B.: A Text Mining System for Evaluating the Stock Market's Response to News. University of Porto (2013)

[10] Gidofalvi, G., Gidófalvi, G.: Using News Articles to Predict Stock Price Movements (2001)

[11] Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J.: Challenges and Opportunities with Big Data. A community white paper developed by leading researchers across the United States (December 2, 2012), http://cra. org/ccc/docs/init/bigdatawhitepaper. pdf

[12] Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor. In: Proceedings of the Advances in Digital Libraries Conference, Washington, DC, USA, p. 12 (1998)

[13] Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. Tools

[14] Oracle, O.: Data Mining Concepts (December 03, 2011), https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004 (accessed: November 23, 2014)