Expert Systems with Applications 39 (2012) 10674–10680

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Predicting published news effect in the Brazilian stock market

P.S.M. Nizer *, J.C. Nievola

PPGIa, Pontifical Catholic University – Paraná (PUC-PR), Curitiba, Paraná, Brazil

ARTICLE INFO

Keywords:
Text mining
Volatility forecast
Stock market
News effect

ABSTRACT

The Efficient Market Hypothesis states that the value of an asset is given by all information available in the present moment. However, there is no possibility that a single financial analyst be aware of all published news which refers to a collection of stocks in the moment they are published. Thus, a computer system that applies text mining techniques and the GARCH model for predicting the volatility of financial assets may helps analysts and simple investors classifying automatically the news which cause the higher impact on stock market behavior. This work has the goal of creating a method for analyzing Portuguese written news's content about companies that have their stocks negotiated in a stock market and trying to predict what kind of effect these news will cause in the Brazilian stock market behavior. Also, it was demonstrated in this study that it is possible to find out whether certain news may cause a considerable impact on prices of a negotiated stock.

1. Introduction

The work of a financial analyst in the stock market consists in suggesting his client to buy or sell stock according to the expectation of the rise or fall of the stock value. This expectation, as any information, is incorporated into the stock value and also defines the price which the stock is negotiated. Determining what is the real price of a stock is crucial to experts in the stock market.

Based on the Efficient Market Hypothesis (Fama, 1970), the value of a financial asset is given by all information available in the moment. However, there is no possibility that a single financial analyst be aware of all published news which refers to a collection of stocks in the moment that they are published. Thus, a computer system that applies text mining techniques to analyze the content of real time news, simultaneously with econometric techniques for predicting the volatility of financial assets may help analysts and simple investors to choose which news cause the higher impact on stock market behavior.

There are available to final analysts (and they are vastly used) models which try to predict market behaviors. These models help the analysts to determine the risk when negotiating stocks. But only a few studies today try to accommodate in this models another type of data generated and used by the market to define its prices: the content of news published by online media.

The first goal of the research presented in this work is to implement a method that automatically classify important news related to companies which have their stocks negotiated in the market. The news have to be classified as "interesting" if they bring information that may cause a meaningful impact in the market behavior, and as "not interesting" if they do not cause this type of impact.

2. Method

2.1. Database

The data used in this study is composed of the content of textual news and the temporal series of stock prices which are negotiated in the market. It contains the BOVESPA indexes, textual information and numerical information.

The BOVESPA Index (IBOVESPA) is a portfolio of stocks organized by the Stock Market of São Paulo, Brazil (BOVESPA). This portfolio is composed by the stocks believed to be the most important during its elaboration. This index is not static: there are periodic changes either in its belonging stocks list as in the weight of each stock inside the portfolio.

The news collected were the ones published during the year of 2009 and the ones that quote the following companies which have their stock belonging to IBOVESPA (and it is the main reason for choosing these companies): ALL, Ambev, Bradesco, Brasil Telecom, Braskem, Celesc, Cemig, Cesp, Comgas, Copel, Cosan, Cyrela, Eletrobras, Eletropaulo, Embraer, Gerdau, Gol, Itaú, Klabin, Light, Natura, Net, Petrobrás, Sabesp, Siderúrgica Nacional, Souza Cruz, Tam, Telesp, Tim, Usiminas, Vale and Vivo.

The collected news was the ones that quote at least once the company's name, and have their published time mentioned. It was made use of the news written in Portuguese and published in an online media using web portals that publish news continuously.

* Corresponding author.
E-mail addresses: philippe.nizer@ppgia.pucpr.br (P.S.M. Nizer), nievola@ppgia.pucpr.br (J.C. Nievola).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.02.162

For searching news of a given period of time it was used the Google News Archive search mechanism. It was also developed a program that, with the links list generated by Google News Archive, download automatically all HTML files which have the news content. For each HTML file, it was separated and saved into a XML file the main text, the published date, the headline and the online media vehicle's name. The online news portals used for getting textual data were: Folha Online, O Estadão, O Globo and Valor Online. Table 1 shows the number of news collected for each company.

It was also collected the prices series of the above mentioned stocks, with 5 min interval for each quotation, regarding the months of September, October, November and December of 2009. The source for these data was the Bloomberg Terminal for monitoring and analyzing the real-time financial market, which have a historic database of several stock markets. The prices (with date and time of the quotation) series were saved in CSV (comma-separated values) format files.

2.2. Preprocessing

The news stored in HTML code have gone through a preprocessing stage which removed undesired elements like advertisements, HTML tags of images, links, tables and text formatting. For this stage, it was developed software that read all XML files which contain the news texts, remove all HTML tags, and also replace all HTML escape codes for especial characters, common in Portuguese words, for the Unicode characters.

After this process, a pure text without formatting, images or effects remained from the original news stored in HTML, which could be used in the following stage (removing stop words and stemming).

Stop words are the ones filtered out prior to a text mining process, because they are regarded as irrelevant or undesired for the task at hand. Commons stop words are those which do not have semantic signification by themselves like articles, prepositions, etc. For this stage, it was used a default stop word list of Portuguese language available in the LINGUATECA site (http://www.linguateca.pt/).

After that, the news texts were submitted to a stemming process, which intend to group all words by their stem in a unique representation. Thus, after this processes, all variations of words which have the same root or base, are considered the same. Moreover, all word infections, like number, gender or conjugation, for verbs, are removed from them (Vieira & Virgil, 2007).

There is a few stemming algorithms for Portuguese language. One of them is a modification of Porter (1980) algorithm, and another one is Orengo and Huyck (2001). For this work, it was used a Java implementation of Orengo algorithm known as PTStemmer (Oliveira, 2010).

Table 1
Database of news by company.

Company's name	News count	Company's name	News count
ALL	62	Gol	640
Ambev	212	Itaú	1.525
Bradesco	1.472	Klabin	184
Brasil Telecom	484	Light	53
Braskem	215	Natura	379
Celesc	79	Net	25
Cemig	184	Petrobrás	5.414
Cesp	99	Sabesp	446
Comgas	71	Siderúrgica Nacional	159
Copel	97	Souza Cruz	82
Cosan	212	Tam	900
Cyrela	257	Telesp	98
Eletrobras	535	Tim	1.172
Eletropaulo	348	Usiminas	562
Embraer	661	Vale	147
Gerdau	611	Vivo	156
Total	17.541		

2.3. News labeling and information retrieval

For training a classification method it is necessary a previous classified data set. However, the texts given from news websites do not bring with them a label that can be used for classification purposes. Based on a method proposed by Robertson (2008), it was decided to label the texts in two classes: "interesting" and "not interesting". Such as Robertson, for determining whether a news text is "interesting" or not, it was analyzed the temporal prices series of the company's stock quoted in one news text.

If the news text causes considerable change in the way which stocks are negotiated, it is supposed that in the next few moments after the news is published, one finds an abnormal activity. This not expected activity can be observed by the growing of the stock return's volatility, that is, in the following moments it is expected that the variation between positive and negative stock's returns be intensified in some way.

The volatility can be calculated as the variance of return (rise or depreciation) of one stock, being calculated by Eq. (1), where R_t is the return in time t. For this work, it was used $n = 1$.

$$\upsilon = \sqrt{\frac{1}{n}\sum_{j=0}^{n-1}(R_{t-j})^2}. \quad (1)$$

There are econometric models which try to predict the volatility of financial assets. Some of these methods, like GARCH (Bollerslev, 1986), are based on the idea that a period of time that follows another one of high volatility tends also to be of high volatility.

For deciding whether some news is important or not, first it is necessary that one knows whether the published news caused or not some modification in market behavior. Applying the GARCH model in the temporal prices series, and then comparing the predicted volatility with the one observed effectively, it is possible to know whether the market was operating in an abnormal manner. If it happens due to the Efficient Market Hypothesis (Fama, 1970), it can be supposed that the market is assimilating the new information. Hence, if there is a published news minutes before the market behavior modification, it is possible that the information assimilated by market is within the text of this published news. Thus, this news should be marked as "interesting". The other news is marked, by default, as "not interesting".

Fig. 1 shows how the labeling process works. From the historic temporal price series of each stock, it was extracted two others series: the effective volatility temporal series, and the predicted volatility temporal series (using a volatility model). The next step is to extract a third temporal series which represents the difference between the effective and the predicted volatility. For each published news it is possible to know whether there is some modification in the market by analyzing the time near the news publishing time. Then, if the error between the predicted and the effective volatility is bigger than some threshold the news is labeled as "interesting".

Thus, a way to evaluate the news' importance is by the error (difference) between a volatility model and the effective volatility. The error can represent the impact of an external event, because it excludes the volatility normal behavior described in the model.

After labeling the news, it is necessary to select data for training an automatic classifier. It is necessary to select which are the words of the published news that will be used as representative attributes. It is very important to be careful which word will be chosen, because important words should not be dismissed, and, in the same way, irrelevant ones should be ignored.

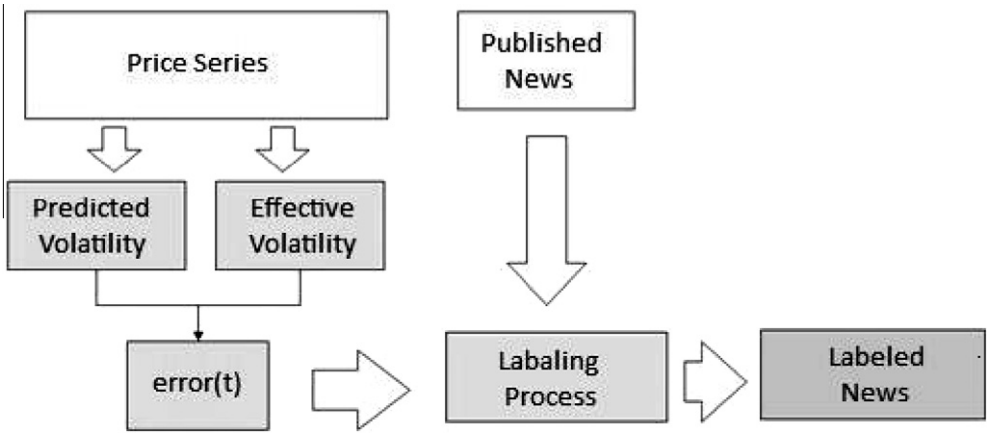

Fig. 1. News labeling.

Two methods for selecting attributes were tested: Information Gain and ADBM25. A function TF-IDF was used to represent the values of each attribute of the news. The performance of the automatic classifiers was measured for each variance of the methods of attributes selection.

2.4. Classifier training

The classification techniques used in this research were Naïve Bayes and SVM, chosen for being the most suitable to text classification (Duda, Hart, & Stork, 2002). It was used the Weka system, which contains the mentioned algorithms, allowing the training and testing of the chosen classifiers (http://www.cs.waikato.ac.nz/ml/weka/).

2.5. Time window

Robertson, Geva, and Wolff (2007a) demonstrated that the error between the prediction of the volatility of a stock using the GARCH model and the effective volatility has a high correlation with the publication of news about that stock. In another work, Robertson, Geva, and Wolff (2007b) categorized news by the error of GARCH model in a given time window. If the error is bigger than the average of the error history added to its standard deviation, the published news is defined as "interesting". The average was calculated in the beginning of each day by the error of the last 20 days of negotiation.

So, this study also uses the error between the volatility prevision and the value. For each month when there are published news which one wants to classify, the GARCH (3,3) model is used to predict the volatility based on the previous month data. It is calculated the average error and its variance with the effective volatility and its prediction (obtained by the use of the model) from the previous month. To label published news, it is necessary to identify a period in which the error between the effective and the expected volatility is greater than the average of the previous error added to its variance. Thus, if the news is in a period $\Delta\tau$ in which this abnormal behavior was detected, it is labeled as "interesting". Fig. 2 shows a stock chart with a news published by a certain time; if an abnormal volatility behavior is detected during the "analysis time", the news is labeled as "interesting".

The performance of the automatic classification of news was associated to some $\Delta\tau$ periods. The values used for $\Delta\tau$ were 5, 10, 15, 20 and 30 min. These values were chosen considering that Robertson et al. (2007a) indicated that the market reaction to publication of news reflects quickly on the volatility.

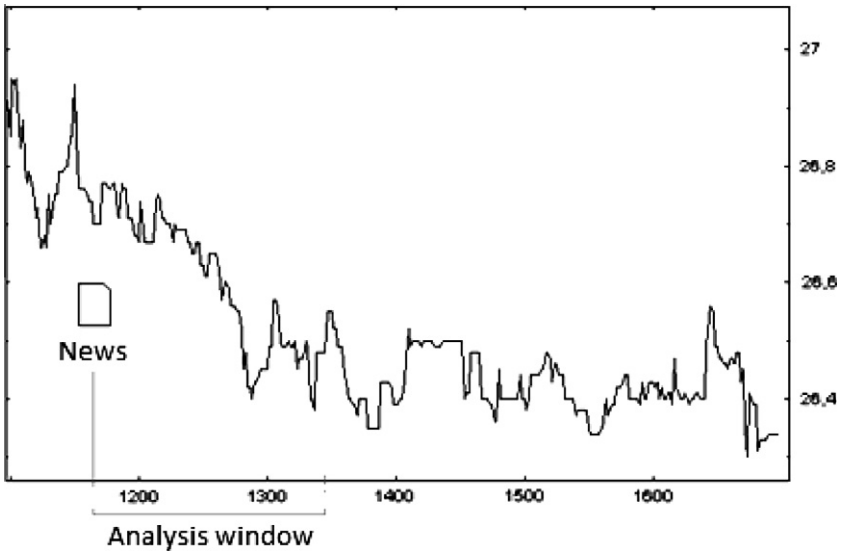

Fig. 2. Published news and price series.

2.6. Measuring results

For measuring the results of the classifiers, it was used the confusion matrix and also the Sensibility, which is the proportion of "interesting" documents correctly classified; the Specificity, which is the proportion of "not interesting" documents correctly classified; and the Accuracy, which is the amount of documents correctly classified.

$$\text{Sensibility} = \frac{\#TP}{\#TP + \#FN}, \quad \text{Specificity} = \frac{\#TF}{\#TN + \#FP},$$

$$\text{Accuracy} = \frac{\#TP + \#TN}{N}$$

3. Results

3.1. News labeling

Before the labeling process, it was excluded from the news list all the ones which was published out of the business hours. This avoids large news groups, which may contain news with opposing content, to be labeled equally, what shall contribute to create more uniform groups, and consequently to create a better classifier. Table 2 shows the news count after the filter.

The news labeling was done by a Java Software created for this specific purpose. For each $\Delta\tau$ value, it was observed a modification in the number of news classified as "interesting" and "not interesting". Table 3 shows the number of news for each value of $\Delta\tau$.

It is important to observe that the number of news labeled as "interesting" is very low compared to the number of "not interesting" news. With $\Delta\tau = 30$ the proportion between these two classes is 0.221 ("interesting" to "not interesting"). For $\Delta\tau = 5$ this same proportion reaches 0.049. It was an expected result, because it is known that just few news are the ones which can modify the way stocks are negotiated changing the normal variation of its volatility. The greater number of published news does not bring so relevant information.

Table 2
Number of news used for creating the classifiers.

Company	News	Company	News
ALL	6	Gol	99
Ambev	37	Itaú	201
Bradesco	216	Klabin	28
Brasil Telecom	48	Light	17
Braskem	37	Natura	59
Celesc	20	Net	4
Cemig	31	Petrobrás	761
Cesp	18	Sabesp	91
Comgas	20	Siderúrgica Nacional	29
Copel	31	Souza Cruz	9
Cosan	31	Tam	143
Cyrela	49	Telesp	35
Eletrobras	64	Tim	148
Eletropaulo	48	Usiminas	51
Embraer	108	Vale	10
Gerdau	81	Vivo	22
Total	2552		

Table 3
"Interesting" and "not interesting" news for each time window ($\Delta\tau$).

Class	$\Delta\tau = 5$	$\Delta\tau = 10$	$\Delta\tau = 15$	$\Delta\tau = 20$	$\Delta\tau = 30$
Interesting	120	211	286	355	463
Not interesting	2432	2341	2266	2197	2089

In preliminary tests with the generated classifiers, whose training set kept the proportion between "interesting" and "not interesting" news of the complete set, it was verified a bias classifying almost all news as "not interesting". Because of that, it was used in all training set generated for this study a fixed proportion between both classes. The proportion selected is 2:1, that is, two "not interesting" for each "interesting" news.

3.2. Classification

The documents used in this study present a relatively small set of words after their stemming. The number of words never exceeded 4500. However, it was tested more restrictive sets of words to verify how it affects the results in some way. From the many term selecting methods, *Information Gain* is the more usual. Another used method is the one proposed by Robertson (2008) named ADBM25, which was developed for his research. For this study, it was tested both methods: *Information Gain* and ADBM25.

For performing the training and evaluation of classifiers, the news was separated in two groups. The first one, being the training set, contains $2t_i$ documents (where t_i is the total number of "interesting" documents from database), where $2t_i/3$ is the number of "interesting" documents and $4t_i/3$ is the number of "not interesting" documents, both generated randomly. The second group, used as test set, contain the remaining documents.

In the following, it was selected features with both methods of term selection. The values used for the number of selected terms were: 100, 200, 500, 1000, 2000 and 4000. For each variation of term selection methods, $\Delta\tau$ and number of selected terms, it was executed five training and evaluations of a classifier, each one with a training set generated randomly. The results presented in Table 4 are the mean of the five executions (the other results bring sensibility near to zero, thus they were omitted).

We note that the sensibility (correctness in classifying "interesting" news) have a low value. Only in one unique case the sensibility is greater than 50%. In the following charts, it is possible to observe how results change while the arguments, with which is generated the classifiers, vary.

It is possible to observe in charts of Figs. 3, 9 and 12 a general trend, in which the sensibility is low when the number of terms is also low. The exception is when *Information Gain* is used with a Naïve Bayes classifier (Fig. 6): with 100 terms and time window of 15 min, the classifier has its best result of sensibility, a mean of 65.07% with standard deviation of 3.0. Despite being the best sensibility result, the specificity ("not interesting" classified correctly) is very low comparing with other results, only 36.87%. However, if the objective of the classifier is to filter news for showing to user only relevant ones, it is important to not lose important news. It is necessary not to lose "interesting" and reduce the minimum the total of retrieved news. In tests realized with the Naïve Bayes classifier with *Information Gain*, one observes that there is a less expressive change in sensibility and specificity while the number of terms increase (Figs. 6 and 7), comparing it with other classifiers

Table 4
News classifiers results.

	Classifier	Δt	Terms	Sens. (%)	Specif. (%)	Precision (%)
Gain	SVM	10	100	2.25	97.12	**93.96**
	SVM	5	200	**44.50**	60.39	60.11
	NB	10	100	**65.07**	36.87	37.80
ADBM25	SVM	15	100	1.50	86.36	**83.03**
	SVM	30	4.000	**45.16**	68.49	66.27
	NB	15	100	30.21	73.63	**71.52**
	NB	10	1.000	**48.73**	54.64	54.45

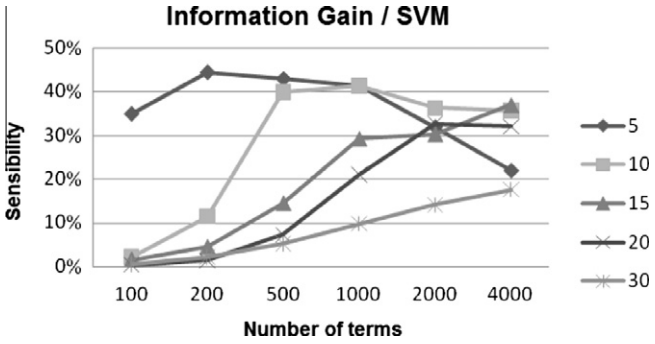

Fig. 3. Sensibility – *Information Gain* and SVM.

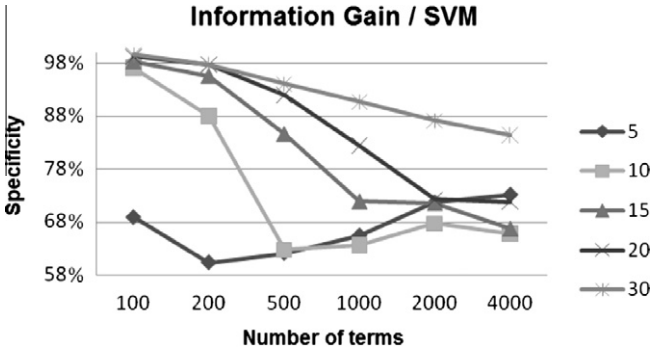

Fig. 4. Specificity – *Information Gain* and SVM.

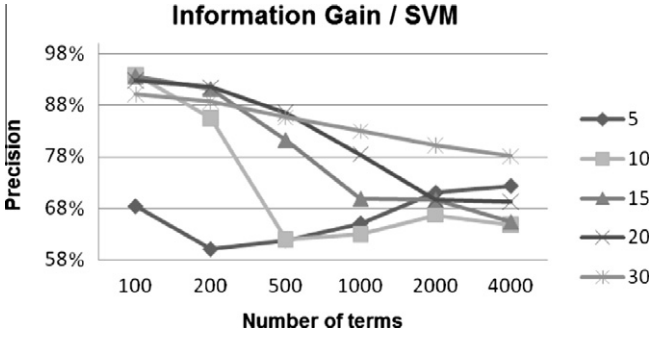

Fig. 5. Precision – *Information Gain* and SVM.

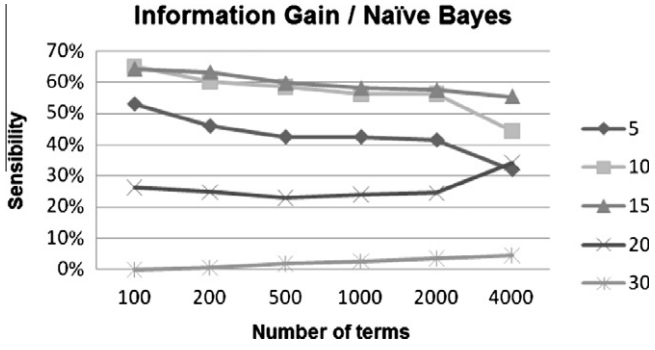

Fig. 6. Sensibility – *Information Gain* and Naïve Bayes.

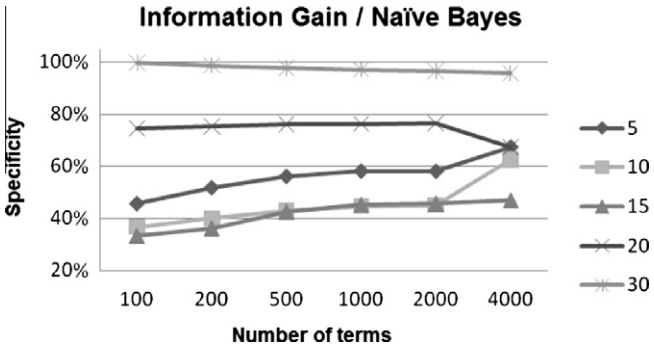

Fig. 7. Specificity – *Information Gain* and Naïve Bayes.

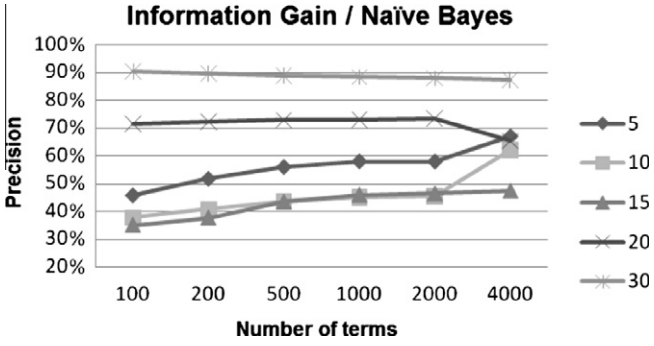

Fig. 8. Precision – *Information Gain* and Naïve Bayes.

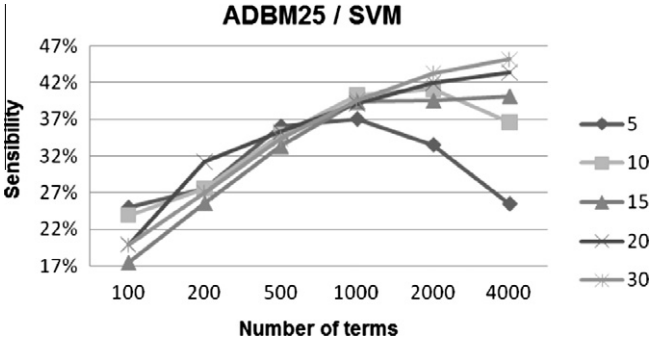

Fig. 9. Sensibility – ADBM25 and SVM.

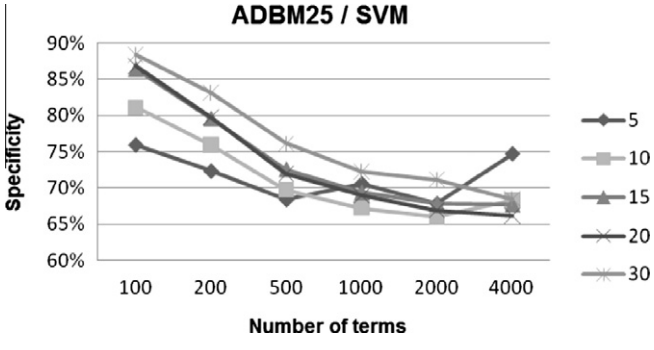

Fig. 10. Specificity – ADBM25 and SVM.

and methods of term selection (Figs. 3, 4, 9, 10, 12 and 13). When the sensibility is greater than 50% (*Information Gain* e Naïve Bayes – 5 and 10 terms), the specificity was always the lower than that.

In tests realized with the method *Information Gain*, when number of selected terms is increased, there is a decreasing trend in sensibility (Figs. 3 and 6). One can observe the inverse with the

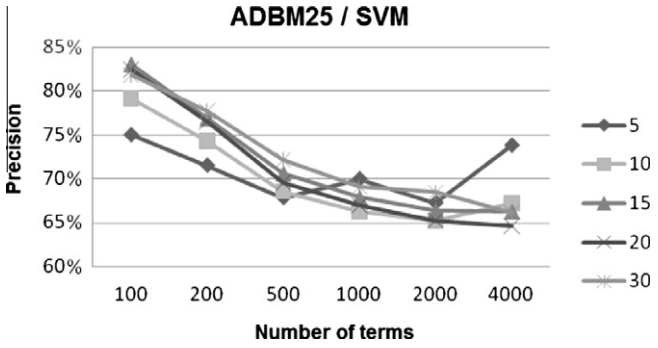

Fig. 11. Precision – ADBM25 and SVM.

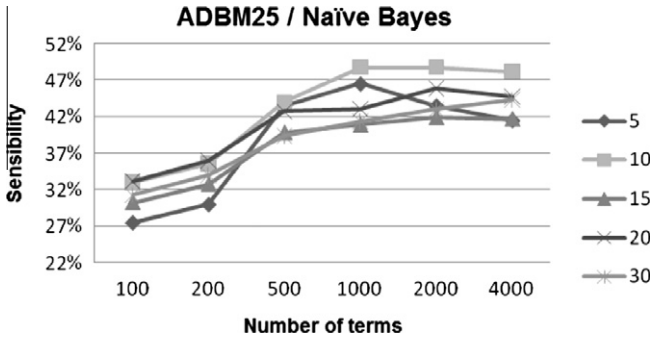

Fig. 12. Sensibility – ADBM25 and Naïve Bayes.

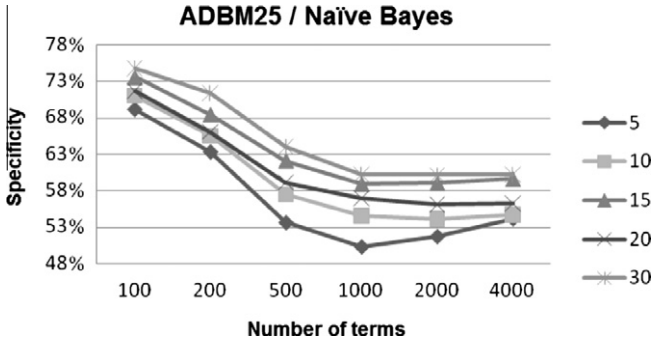

Fig. 13. Specificity – ADBM25 and Naïve Bayes.

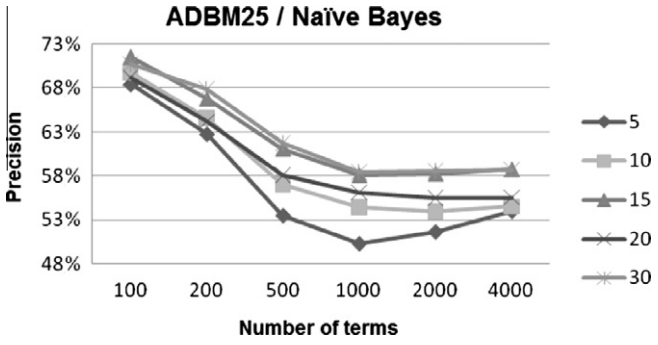

Fig. 14. Precision – ADBM25 and Naïve Bayes.

ADBM25 method: in both cases (SVM and Naïve Bayes), there is an increasing trend in sensibility when the number of terms increases too (Figs. 9 and 12). With the method ADBM25, we also observe a smoother evolution in sensibility while the number of terms increases (beside this index it never surpasses 50%). Specificity ever behaves in the inverse trend of sensibility.

Table 5
Results with NB classifier, ADBM25 and $\Delta t = 10$.

Terms	Sens. (%)	SD Sens.	Spec. (%)	SD Spec.	Precision (%)	SD Prec.
100	32.96	0.073591	71.05	0.071339	69.78	0.068151
200	35.49	0.091495	65.56	0.080833	64.56	0.076644
500	43.94	0.037793	57.51	0.068949	57.05	0.066215
1000	48.73	0.027456	54.64	0.0645	54.45	0.061938
2000	48.73	0.025586	54.16	0.06025	53.98	0.057684
4000	48.17	0.018364	54.73	0.060238	54.51	0.057818

SD = Standard deviation

In all tests, there is no single case in which both classes are correctly classified in more than 50% of cases. Even with classifier's precision being greater than this value in almost all variations, there is not possible to conclude that a classifier was succeeded. This is because of the big difference between the number of "interesting" and "not interesting" news. Thus, the correct classification of "not interesting" news, which is majority, contribute almost exclusively for classifiers precision (Figs. 5, 8, 11 and 14).

If one considers the best case the one with the biggest sensibility from those which the specificity is greater than 50%, it is possible to assign this quality to the ones with Naïve Bayes classifier, ADBM25 term selection and time window of 10 min. In those tests with 1000 and 2000 terms, the sensibility is 48.73%. Table 5 shows results of the referred classifiers.

Thus, we highlight the following variations of classifiers, term selection methods and time window:

- Naïve Bayes, *Information Gain*, 100 terms and $\Delta t = 10$ min, which reaches the biggest sensibility (65.07%). However, the specificity is only 36.87%;

- Naïve Bayes, ADBM25, 1000 terms and $\Delta t = 10$ min, with 48.73% of sensibility and 54.64% of specificity.

4. Conclusion

This study described a method which aim to identify which published news that quotes a company has the property of causing some changes in the way company's stocks are negotiated (these changes are measured by the prices' volatility). According to the research by Robertson (2008) with published news and companies from United Kingdom, United States and Australia, the precision of his method is around 80%. However, as we consider sensibility the most important result (for this method to be used as a news filter for a final user), that is, the success rate of classifying "interesting" news (those ones which may cause changes in stock market), the Robertson's study reached 42.26% as sensibility (with specificity 80.77% and precision 80.31%).

The results obtained in this study, for news published in Portuguese language, can be compared with those obtained by Robertson. The best values of precision obtained by this study surpass 80%. However, in these cases, the sensitivity is only 20%. This proposed method's results are similar to those of Robertson. Therefore, a filter that applies this technique would reduce the number of news for a financial analyst to analyze.

It was also verified the efficiency of ADBM25 selection terms method. By the results obtained with this method, it is possible to verify a line with a smooth variation while the terms number increase.

Further studies in this area may consider whether it is possible to improve some volatility model through the use of this classifier, or try to apply a different method of classification which may give better results.

References

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*, 307–327.

Duda, R., Hart, P., & Stork, D. (2002). *Pattern classification* (2nd ed.). New York: Wiley Interscience.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. Papers and proceedings of the twenty-eighth annual meeting of american finance association. *Journal of Finance, 25*(2), 383–417.

Oliveira, P. (2010). *PTStemmer: A Stemming toolkit for the Portuguese language*. Retrieved from: <http://code.google.com/p/ptstemmer>.

Orengo, V. M., & Huyck, C. (2001). A stemming algorithm for the portuguese language. In *8th International symposium string processing and information retrieval (SPIRE 2001), proceedings, 13–15 November 2001, Chile* (pp. 186–193).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Robertson, C. S., Geva, S., & Wolff, R. C. (2007a). *The intraday effect of public information: Empirical evidence of market reaction to asset specific news from the US, UK, and Australia*. SSRN Working Paper Series.

Robertson, C., Geva, S., & Wolff, R. C. (2007b). Can the content of public news be used to forecast abnormal stock market behaviour? In *Seventh IEEE international conference on ICDM* (pp. 637–642).

Robertson, C. (2008). *Real time financial information analysis*. Queensland University of Technology.

Vieira, A. F. G., & Virgil, J. (2007). Uma revisão dos algoritmos de radicalização em língua portuguesa. *Information Research, 12*(3).