A MODEL BASED ON SENTIMENTS ANALYSIS FOR STOCK EXCHANGE PREDICTION - CASE STUDY OF PETR4, PETROBRAS, BRAZIL

Milson L. Lima, Sofiane Labidi, Thiago P. do Nascimento,
Nadson S. Timbó, Gilberto N. Neto, Marcus Vinicius Lima Batista

Post-graduation Program in Electrical Engineering,
Federal University of Maranhão, São Luís, Brazil 65080–805
milsonlima@hotmail.com, soflabidi@gmail.com,
thiagopinheiro.nascimento@gmail.com, nadsontimbo@gmail.com,
gilberto.nunes@ifpi.edu.br, marcus_89lima@hotmail.com

ABSTRACT

Predicting the behavior of shares in the stock market is a complex problem, that involves variables not always known and can undergo various influences, from the collective emotion to high-profile news. Such volatility, can represent considerable financial losses for investors. In order to anticipate such changes in the market, it has been proposed various mechanisms to try to predict the behavior of an asset in the stock market, based on previously existing information. Such mechanisms include statistical data only, without considering the collective feeling. This article, is going to use natural language processing algorithms (LPN) to determine the collective mood on assets and later with the help of the SVM algorithm to extract patterns in an attempt to predict the active behavior. Nevertheless it is important to note that such approach is not intended to be the main factor in the decision making process, but rather an aid tool, which combined with other information, can provide higher accuracy for the solution of this problem.

KEYWORDS

*Sentiment Analysis, Twitter, Prediction of Stock Exchange*s

1. INTRODUCTION

Nowadays, with the advancement of Information and Communication Technologies, there had been developed an enabling environment for widespread use of social networks. Such environment is a favorable place for natural exposure of individuals, their desires, preferences and manifestations in its various forms, which makes the result of this process, more natural and close to reality. In ciberspace, each subject is effectively a potential producer of information [1]. We have as a result of this interaction a source of valuable information and yet barely explored. Their use may be the basis for establishing, certain preferences and calculating the mood of individuals. Nowadays, one of the most popular social media is the Twitter, with over 200 million user who share their opinion through tweets.

Jan Zizka et al. (Eds) : CCSIT, SIPP, AISC, CMCA, SEAS, CSITEC, DaKM, PDCTA, NeCoM - 2016
pp. 131–139, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60112

Bearing in view the large amount of available information, resulting from this process, this paper aims to analyze the content of messages posted on Twitter about certain company in order to establish a relationship between the collective mood of publications and their influence on the price of an asset in the financial market.

The capacity to anticipate the unpredictability of a market, where there are several elements that can influence the value of a stock is desirable on many aspects, especially from a strategic point of view aiming the profit.

2. RELATED WORKS

Frequent exposure of users on social networks, leaves a range legacy of express information in the way that cover a wide variety of topics. And it is on this valuable trace, resulting from the increasingly frantic messages exchange, that many researchers seek to extract valuable information about various fields, making use of various techniques to seek the answers to their questions.

There are several previous works related to sentiment analysis in textual sources using social networks, which goes from predicting the box office revenue for movies [2], consumer opinions about products or services [3], political and policy analysis[4] until disease outbreaks [5].

One who firstly addressed this issue was Pang & Lee [6], which shows in his work an overview of various techniques used for sentiment extraction from the analysis of texts.

Asur and Huberman [7], proposed a model using linear regression, to foresee the box office revenue of a certain movie a few weeks after it is released, it so there would be enough opinions to good analysis. His data set was obtained using Twitter Search API, collected every hour. As research argument they used keywords present in the title of the film, resulting in 2.89 million tweets related to different movies over 3 months. At the end of the task the authors point out the success of their predictions, where it method was more effective than any other used for this domain.

Another highlighted job is from Bollen et al. [8], in it was studied the influence of expressed polarity by Twitter users about particular company and their respective impact on the stock market. Data collection was obtained through Twitter and sentiment analysis was based on the Google-Profile of Mood States (GPOMS) . After analyzing the entire volume of data collected, it was found that variations of collective mood detected, was also observed in the financial market.

3. SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is a branch of mining texts concerned to classify texts not by topic, but by sentiment or opinion contained in a given document. Generally associated with the binary classification between positive and negative sentiments, the term is used more embracing to mean the computational treatment of opinion, sentiment and subjectivity in texts [9].

For Liu [10], The Sentiment Analysis or Opinion Mining is the computational study of opinions, sentiments and emotions expressed in text. The textual information may be classified into two main types: facts and opinions. The facts are objective expressions about entities, events and their properties. The opinions are generally expressions that describe the sentiment and evaluations of people related to a certain entities, events and their properties.

The sentiment analysis has been one of the most active research areas in the field of Natural Language Processing- NLP and aims to obtain and formalize the opinion and subjective knowledge in unstructured documents (texts) for further analysis within a specific domain [11]. The sentimental analysis has been used in different areas, for example:

- Policy - To measure the popularity of a particular candidate for public office
- Industry - To evaluate the acceptance by consumers to a particular product.
- Stock Exchange - To measure the collective mood on certain asset traded on the stock exchange.

Extract sentiment in textual sources is a complex task, since the natural language processing often comes across expressions which are surrounded by neologisms, irony and other linguistic variations that hinder the correct sentiment extraction.

For our particular case we use the feeling of analysis for the financial area, more specifically in the field Stock Exchange.

4. METHODOLOGY

The topics below demonstrate the methodology applied in this work

4.1. The Research Environment

The environment used for this research was Twitter, a microblogging service, which uses short messages up to 140 characters for information transmission, and can on different platforms [12]. Responsible for about 500 million daily posts [13], it has become a propitious place for researchers and companies seeking for information about the most different subjects through techniques of text mining and natural language processing.

4.2. Data Collection

To access the publications made by users, was used the API from Twitter, provided by the site itself. This is a feature that by user credentials as a developer, allows access to messages through OAUTH5 protocol.

To collect the data, it was used as a criterion messages that contained at least one mention of a word selected, which in our case is represented by the name of the object of our study, Petrobras. The choice of the appropriate term is of utmost importance for the result of the collection of information, since it is the main criteria for the search.

They were collected daily about 3,000 tweets, totaling approximately 40,000 messages between 2015-09-01 and 2015-11-20, in a timetable, between 9:00 AM and 17:00 PM (GMT -02:00) time when the Market was in full operation.

The data resulting from the information gathering process were stored in a database and properly addressed in the pre-processing step in order to remove expressions, unnecessary characters and retweets, aiming not interfere with the review process of individual feeling and collective. Figure 1 illustrates the application process and messages storage.

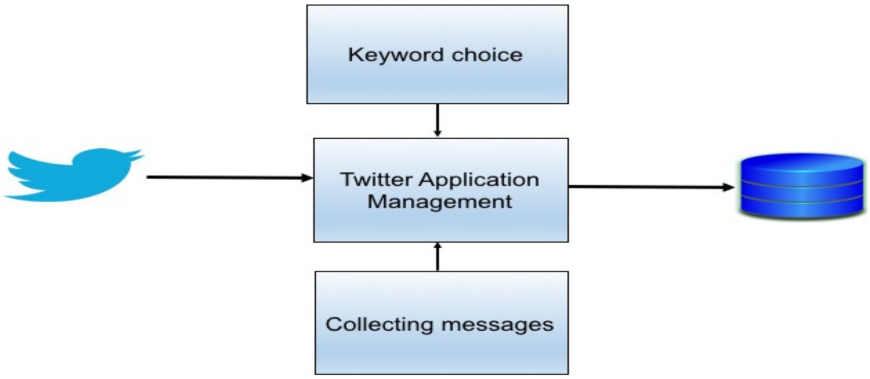


Figure 1. Application Process and storage

4.3. Used Tool

After testing some tools available for sentiment analysis, we chose to use the lexicon Sentiment140 [14], that have shown to be more efficient in the classification of messages with twitter features. The table 1 shows the tools used in the testing phase to find the one best suited to this work, as well as some of its features.

Tabel 1. Ferramentas para Análise de Sentimento

Tools	**Short Description**
SentiWordNet	Lexical dictionary and scores obtained by semi- machine learning approaches
SenticNet	Natural language processing approach for inferring the polarity at semantic level
Sentiment140	API that allows classifying tweets to polarity classes positive, negative and neutral.

In the testing phase were selected 100 tweets randomly within our research field, to be sorted manually and then compared with the results obtained in other tools, as shown in Table 2, where best results are evident when messages were subjected to classification by Sentiment140 lexicon, considering that the polarization done manually in our sample.

Table 2. Hit índex to classify tweets

Manually	**Senticnet**	**SentiWordNet**	**Sentiment140**
100%	55%	52%	**65%**

The Sentiment140 is a lexicon designed specifically to analyze tweets, it corpus was created from a collection of 1.6 million tweets composed of positive and negative emoticons. In it the tweets are labeled in positive or negative according to the respective emoticon. From the auto-labeling was found that words which occur most frequently in positive or negative tweets, yielding a dictionary with more than 1 million terms, distributed over 62,468 unigrams, bigrams pairs 677,698 and 480,010 [15].

The classification of a tweet sentiment "w" is calculated through the value of it score, as shown below:

$$\text{score}(w) = \text{PMI}(w, \text{positive}) - \text{PMI}(w, \text{negative}) \qquad (1)$$

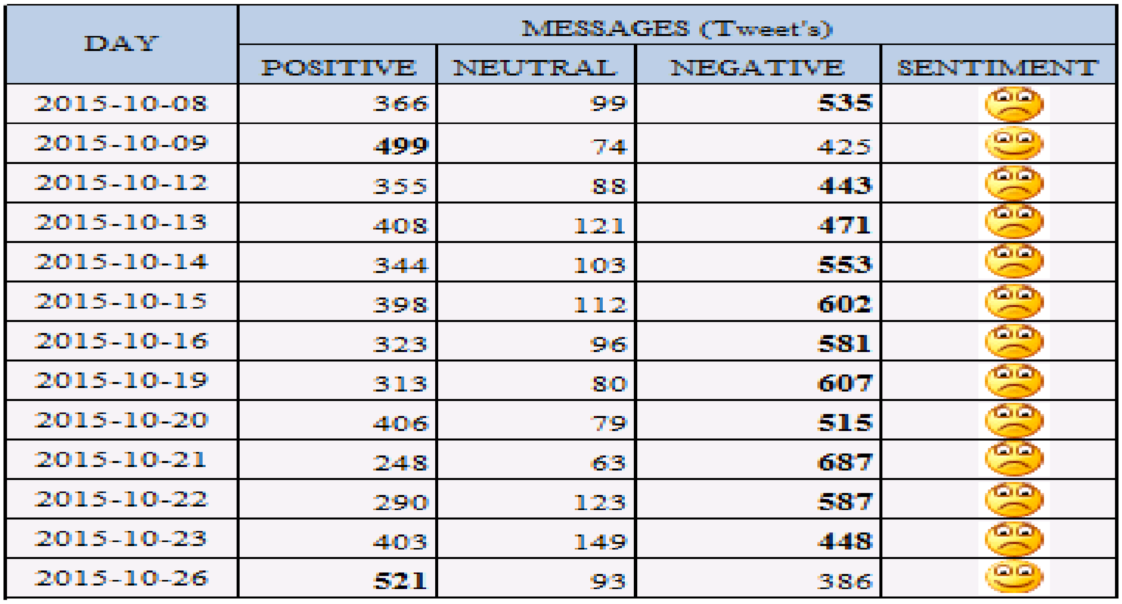
Where, PMI represents pointwise mutual information and receive the default occurrences as positive and negative from the expression, respectively. A positive score indicates association with the positive sentiment, while a negative score indicates association with the negative sentiment.

Like the majority of the methods and techniques available for this purpose contents is available only in the English language.

4.4. Analyzing Sentiment

In order to obtain the daily collective sentiment, which are commented on Twitter about in our dominion that were randomly selected 1,000 tweets per day in Portuguese from Brazil. After the preprocessing steps and data transformation data, the result was arranged as shown in Table 3, which demonstrate a period of the fragment under analysis (2015-10-08 to 2015010-26), where there is clearly a prevalence of "negative" mood over "positive", the latter won only two instances, one on 2015-10-09 and another on 2015-10-26.

Table 3 – Daily Collective Sentiment



DAY	MESSAGES (Tweet's)			
	POSITIVE	NEUTRAL	NEGATIVE	SENTIMENT
2015-10-08	366	99	**535**	☹
2015-10-09	**499**	74	425	☺
2015-10-12	355	88	**443**	☹
2015-10-13	408	121	**471**	☹
2015-10-14	344	103	**553**	☹
2015-10-15	398	112	**602**	☹
2015-10-16	323	96	**581**	☹
2015-10-19	313	80	**607**	☹
2015-10-20	406	79	**515**	☹
2015-10-21	248	63	**687**	☹
2015-10-22	290	123	**587**	☹
2015-10-23	403	149	**448**	☹
2015-10-26	**521**	93	386	☺

The highest occurrence of "negative" mood especially in larger quantities our sample, can be attributed to recent episodes of financial scandals involving Petrobras company which is the subject of our study.

4.5. Machine Learning and Sentiment Analysis

Within the process of sentiment analysis and prediction, a major challenge is to find an efficient way to classify the texts for analysis. The classification process consists to find, through machine learning, a function that expresses the best possible way, data classes involved in the field and thereby, making automatic the classification process for new instances, with reference to the model for which it was trained.

Traditional machine learning approache to text classification, as Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM), are quite effective when applied to sentiment analysis problem [16].

In this article we are going to use a data set previously labeled for the supervised training, these data were divided into two groups, one for training, where we are going to use 70% of the instances and another to test with 30%.

The tool used for the knowledge process discovery was the WEKA and attributes chosen for training and testing were a historical series within the containing period in question: opening price, minimum price, maximum price, closing price and closing situation (high / low). We use two algorithms classification widely used for this purpose, the Support Vector Machine (SVM) and Naive Bayes, in search for the best accuracy for this first stage of our tests. Table 4 shows the results obtained for each.

Table 4 - Comparison of Naive Bayes and SVM

Algorithm	**Correctly Classified Instances**
Naive Bayes	35.29%
SVM	47.05%

In a second step, and now using only the SVM, for been shown better results in the first stage, it was inserted the attribute "sentiment" that can assume two values: positive or negative, at the same set of data and maintaining the same proportionality between training and testing. Table 5 shows the result of the insertion of this attribute for the accuracy of the chosen classifier.

Table 5 - SVM after inserting the attribute "sentiment"

Algorithm	**Correctly Classified Instances**
SVM	88.23%

4.6. Results Obtained

It is noticeable that the insertion of the attribute "sentiment" into the set of data processed by the classifier, resulted in a significant gain in the correct classification of instances using the SVM algorithm.

One of the factors that certainly contributed for this result was not high number of instances that was used to train our classifier, specifically 56. In future work we will repeat the same tests for a much longer period and to seek to ratify the relationship between the collective mood and the financial market.

5. CONCLUSIONS

In this investigation when using Twitter data, it was found that to use them satisfactorily for this purpose, it is necessary a great process to treat the information, since the messages exchanged on this platform are rich in ironies, neologisms and slang which greatly complicates the analysis of the feeling contained in tweets.

Another factor worth mentioning is that not always the collective mood is the true sentiment is about a particular asset, we take as an example the case study in the Brazilian company Petrobras. It is a state-owned joint stock, which in its present time has become entangled in a series of political scandals and misuse of public money. Much of comments where it appears the company name, has an essentially political connotation, that is, there is a shift of focus in the comments, many users write messages containing insults to politicians, for example, which will certainly happen in a scenario where politics the country were bad and the company had its valued stocks, in other words, companies with these characteristics may not be sensitive to the public mood, simply because they have their image associated with the state. The present moment the company with its shares rising devaluation also coincides with the unfavorable political moment which may explain the results presented in the study.

which demonstrate that the mood predominantly negative in the days study, followed by a devaluation of the shares.

Another approach that can be developed from this work is based on the same methodology applied, expand it to companies of different sizes as small caps, which are of low market value companies and study their sensitivity to the collective mood, in other words, it is to evaluate whether the company's capital size is a factor to be considered for this type of study.

Despite the studies focused on sentiment analysis have evolved significantly in recent years, the tools resulted from this process should be seen as something complementary in the decision making process, given the complexity of extracting exactly sentiment textual sources in natural language.

REFERENCES

[1] LEMOS, Lúcia. O poder do discurso na cultura digital: o caso Twitter. Revista de Estudos e Pesquisas em Linguagem e Mídia. São Paulo, v.4. n.1. Janeiro-Abril de 2008.

[2] S. Asur and A. Huberman. Predicting the Future with Social Media. CoRR, abs/1003.5699. 2010.

[3] Eirinaki, M., Pisal, S., and Singh, J. Feature-based opinion mining and ranking. Journal of Computer and System Sciences 78, 4 (July 2012), 1175–1184.

[4] Fang, Y., Si, L., Somasundaram, N., Yu, Z.: Mining contrastive opinions on political texts using cross-perspective topic model. In: Proceedings of the fifth ACM international conference on Web search and data mining - WSDM'12. p. 63. ACM Press, New York, USA (2012).

[5] St Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks? BMJ, 344, 1-3.

[6] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. doi:10.1561/1500000011

[7] S. Asur and A. Huberman. Predicting the Future with Social Media. CoRR, abs/1003.5699. 2010.

[8] BOLLEN, J., MAO, H., AND ZENG, X. Twitter mood predicts the stock market. Journal of Computational Science 2, 1 (2011), 1–8.

[9] PANG, Bo;LEE, Lilian. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, June 2008.

[10] Liu, B. (2010). Sentiment Analysis and Subjectivity. In Nitin Indurkhya & F. J. Damerau (Eds.), Handbook of Natural Language Processing, Second Edition. Boca Raton, FL: CRC Press, Taylor and Francis Group.

[11] Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 5(1), 1–167 (May 2012).

[12] Stevens, Vance (2008). "Trial by Twitter: The Rise and Slide of the Year's Most Viral Microblogging Platform". TESLEJ: Teaching English as a Second or Foreign Language, Vol. 12, N. 1, 2008.

[13] https://blog.twitter.com/2014/the-2014-yearontwitter, accessed: (2015-09-12).

[14] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision. In Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.

[15] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In Proceedings of the International Workshop on Semantic Evalua- tion, SemEval '13, Atlanta, Georgia, USA, June.

[16] Read. J (2005). Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification. Proceedings of the ACL Student Research Workshop. 43-48.

AUTHORS

Milson Louseiro Lima

Postgraduate Diploma in ANALYSIS AND SYSTEMS PROJECT (UFMA, 2007), graduated in Economic Sciences from the Federal University of Maranhão, Brazil-UFMA, (UFMA,2006). Is ERP developer and mobile devices, since 1998. He is currently a Master's student Electrical Engineering course for Computer Science (UFMA, 2014), working in the Intelligent Systems Laboratory at the Federal University of Maranhão(LSI/UFMA).

Sofiane Labidi

Bachelor's at Ciência da Computação from Institut Supérieur Scientifique (1990), master's at Ciência da Computação from Université de Nice Sophia Antipolis Centre National de Recherches Scientifiques (1991) and doctorate at Ciência da Computação from Institut National de Recherche en Informatique et Automatique (1995). He is currently full professor at Universidade Federal do Maranhão. Has experience in Computer Science, acting on the following areas: knowledge management, multi-agent systems, educational technologies, agents, artificial intelligence and business proces modelling.

Thiago Pinheiro do Nascimento

Bachelor's at Ciência da Computação from Faculdade de Ciências Humanas, Saúde, Exatas e Jurídicas de Teresina (2012) and master's at Electric Engineering from Universidade Federal do Maranhão (2015). Has experience in Computer Science, acting on the following subjects: frameworks, software engineering component-based, service-oriented architecture, software reuse and web development.

Nadson Silva Timbó

Has graduation at Ciencia da Computação by Universidade Federal do Maranhão (2013). Currently is of Universidade Ceuma. Has experience in the area of Computer Science. management, multi-agent systems, educational technologies, agents, artificial intelligence and business proces modelling.

Gilberto Nunes Neto

bachelor's at Licenciatura Plena em Computação from Universidade Estadual do Piauí (2006). Has experience in Computer Science, focusing on Computer Science

Marcus Vinicius Lima Batista

Bachelor in Information Systems from the University CEUMA. Researcher at the Center of Research and Extension in Technology and Information Systems (NUSTI). Management studies in T.I; Research online e-Goverment