

# CLASSIFYING NEWS ANNOUNCEMENTS USING NAÏVE BAYES METHOD TO PREDICT EURO / DOLLAR VOLATILITIES

Janita S Meliala SE MT\*, Dr. Petrina Faustine SE MM MSc\*\*, Luxky Wijaya\*\*\*

## Abstract

*Commonly used analysis of price movements in foreign exchange (forex) market are fundamental analysis and technical analysis. One among many indicators which influences the forex price is news articles. In this study, it was selected and classified news announcements which affected euro/dollar return volatilities. By using the Naïve Bayes theorem, the news was “weighted” to become the predictor of the forex price movements. The post-announcement reactions were highlighted and analyzed. They were classified and labeled as: “up”, “down”, or “unchanged”. The study revealed a significant predictive power of news announcements over the forex price movements of euro/dollar return volatilities.*

*Keyword: Naïve Bayes, forex, news announcement*

## INTRODUCTION

Information or news gives us description about the future or the past happenings. When information as data is processed into a better useful data for the real beneficiaries, it can become very valuable, in forms of understandable values in current or future decisions ( Gordon B. Davis, 2000). Information is something which reveals the results of processed data which is organized and useful for the beneficiaries (Barry E. Cushing, 1982). The influence of news can affect many aspects: economy, politics, cultures, forex, stocks and many others.

---

**Janita S Meliala, S.E., MM.** Vice Director of Management Department and Lecturer of Faculty Business & Management at Widyatama University.

**\*\* Dr. Petrina Faustine, SE, MM, MSc.** Deputy Rector for International Office, Lecturer & Researcher of Faculty Business & Management at Widyatama University in Bandung /Indonesia, Board Committee Member of International Family Enterprise Research Academy (IFERA)

**\*\*\* Luxky Wijaya,** Students of Information Technology Department, Widyatama University

---

The influence of news towards forex transaction volatilities has been discussed by many researchers such as Evans and Lyons (1999); Melvin and Yin (2000) and Cai, Cheung, Lee, and

Melvin (2001). Broadly spoken, the analysis of forex trading can be classified into two groups, they are: Fundamental Analysis and Technical Analysis.

As it was discussed earlier, jumping into forex without knowing how to do analysis can be equal to gamble in it. It is clear, that gambling will end at loss. But, what a pity is, that this phenomenon is not just experienced by new players in the forex. There are too many people who experienced loss at tens to hundreds of billions dollars, before they realized that without good analysis, such forex trading is equal to gambling. Gain without a good analysis is generally because of luck factor, which will not sustain, and all funding can vanish in just weeks or months because of the minimum knowledge in analyzing. Therefore, this study tries to emphasize the importance of an analysis in forex trading, so that one can expect to reach the targeted profitability in the long run. The know-how on analyzing price movements is quite an absolute skill of the forex "players".

In general, analysis in forex trading can be classified into two methods, i.e.: Fundamental and Technical Analysis. When one heard on television, news on increase of world oil price, or American Central Bank's decision, that the Fed increased the interest rate, this news is called Fundamental news. Such news can be very influential on the price movements in the market, and can be very influential towards the portfolio of an investor in forex world. Analysis which is dependant on these sorts of news is called Fundamental Analysis.

Fundamental Analysis concludes that price movements are resulted from government's announcements and policies, as well as response feed-backs of the market to the announced news. In fact, such fundamental news have been the triggers of price movements in forex market. Any news which will, or has appeared publicly, triggers mutual reactions to the traders, both the sellers and the buyers, which resulted in price changes. It means, announced news in a country can cause a change on this country's currency. This change, at the end, will push the government, as the decisive monetary authority, to set an economic policy. The consequence is: it causes a publication of novel fundamental news. So it happens continuously.

A fundamental analyst, it is vital that he should definitely possess: timeliness, preciseness of the news, ability to forecast the market's reaction towards the published news. Without the possessions of these factors, it will be hard for a fundamental analyst to take advantage of the available news to gain profit on forex. For an easy example, say, when one hears news that the Fed has just increased their interest rate into 25 basis points (which is equal to 0.25%) a couple of minutes ago (in this business the time is counted not in hours anymore, but in minutes! Being left behind of the news means that opportunities passed already!). Generally, US Dollar currency will become stronger after the news, and Buy position can be preceded. But, if one hears the news two days later, this important news has become useless as the market has stopped reacting, and has probably even come into a corrected situation. Thus, in this forex trading, timeliness in obtaining news is crucial. So does the source of news one hears. It does not matter whether the news heard is not valid, as long as the market has the same perception with everybody else. Therefore, the standard is the market itself, not the reliability of the news.

In the forex world, there are over 50 different fundamental news issued by every country, and each of them affects differently to the price movements. Now the question is: what are these 50 news? And how the published news influences the currencies' movements?

One of the text classification method can be used is Naïve Bayes, which is often called as Naïve Bayes Classifier (NBC). The advantages of NBC are: it is simple, and yet it has a high accuracy. NBC uses probability theory as the fundamental theory. There are two steps in text classification process. The first step is: exercising example. The second step is a process of classifying documents which categories are unknown. This second step can be called Behavioral Finance, which currently is seriously looked into and studied.

Behavioral finance is a science on aspects of attitude, or psychology, of the decision makers, especially the investors in stock markets. Previously, the decision makers have been assumed as being rational and consistent. Behavioral Finance tries to oppose the assumptions and clarify the bias by using many psychological terminologies. Behavioral finance becomes popular for the first time when a psychologist, Daniel Kahneman with his prospect theory, was awarded Economy Nobel Price in 2002. Thenafter appeared also Microstructure Market Theory, which is a branch of finance or investment science, which studies the influences of capital market regulations on outcomes, such as: return, volatility, liquidity, efficiency, and transaction cost. Using news as an assessment variable of Behavioral Finance Theory or Microstructure Market Theory, to assess or to decide the movements of euro/dollar in foreign currencies, which is most likely has positive value. From this assessment, a financial algorithm can be produce, to be implemented on the application which has been connected to the tick-chart of Euro/Dollar forex or full forex transaction application. The existence of this application, or finance algorithm, can support investors or brokers in deciding the direction of transaction: increase, decrease, or balance to a transaction.

## Literature

### Macroeconomics Market

Empirical proofs show that bias or misinformation in a market can lead forex price to be far from the normal market price. Bias examples which can cause distortion are: Lack of information about a certain condition which can create false hope, world market will then be avoided by international investors, and research area will become less. Market is bias towards loss-sales; after 12 month market performance will drastically decrease, the investors suffer from losses. Political stability is an important factor. Any issues arise in the market, are dependant on it.

Using past performance as a guidance in the future? The world in general depends upon the current happenings which have great effects in the future. The element "following the main stream" and bias towards "good" forex because they play save. Foreign and local investors will both buy enormous forex which is relatively safe, which includes in a perception as a safe foreign currency.

We believe that investors can exploit market inefficiency. In short term, forex prices react to the market inefficiency. Whereas in a mid-term, market will become semi-efficient (in the

developing/emerging market), and will give good reward to the investors whom have been exploiting price inefficiency, or in other words: those who have identified currencies which are “undervalued” or “overvalued”.

The phenomena of “light to quality” show an alternative measurement to prevent risks, one among many which generally is related to the investment style. The need to “safely invest” will become stronger when the investors feel unsure about the situation. Because the boiling market reflects and may enhance the investors’ worries, “light to quality” often happens during the volatility period. During the period, investors’ risks are increasing, they tend to pursue the strong currency, and they will sell the currencies which they think are risky. Style characteristic is very much related to quality perceptions. Therefore, the premium of uncomfortable style or contrarian should be high after the boiling period (Financial Analyst Journal, March-April 1995).

Conventional macroeconomics theories on exchange rate decision have been accepted; nevertheless they are limited to empirical learning. Now there is a wide comprehension that the “main” macroeconomic variables play relatively unimportant roles in deciding short-term exchange rate movements. A comprehension which has to be altered in the dynamic exchange rate needs an examination which is closer to the function of forex sales or exchanges. A conference on “Microstructure in Foreign Currency Exchange Sales” which is related to the National Bureau Economic Research (NBER) and Banca d’Italia was held in Perugia, on 1-2 July 2008, which considered how the theories of “microstructure market”, which have been developed to study the equities and to take important roles to explain exchange rate behaviors. The conference was organized by: Jeffrey Frankel, a Senior Fellow at the Institute for International Economics in Washington DC, a Professor in Economics at California University, and NBER Director untuk International Finance and Macroeconomics; Giampaolo Galli, a Co-Director of Research Department of Banca d’Italia; Alberto Giovannini, a board member of Treasury Ministry of Italian Experts; Jerome A. Chazen, a Professor International Business at Columbia University, and also a Co-Director of International Macro Economy Program (CEPR).

It was explained that in the microstructure, there is relationships between trade volume and risk; between information and trade intensities which are related to the exchange rate issues in more details, such as the behaviors of the market participants during the EMS crisis since 1992, the mechanisms predict speculatively, and capital affectivities are designed to control the crisis. [15]

In order to highlight the remaining theoretical gaps and briefly document the poor empirical performances of: the purchasing power parity (PPP) theory, flexible price and sticky price (overshooting) monetary models, also the portfolio balance and equilibrium models, Mark Taylor of University of Liverpool and CEPR, has written with Robert Flood, in “Exchange Rate Economics: What's Wrong with the Conventional Macro Approach?”, of their surveying the traditional literature, that their estimations which used pooled annual data for 21 industrialized countries from the floating-rate period, revealed that - although very poor over even a one-year horizon- the explanatory power of the fundamentals in applications of both the PPP theory and a simple flexible price model, was considerably greater for data averaged over periods of five, ten or twenty years. Traditional approach can therefore still play a useful role in explaining longer-run exchange rate behavior [15].

The contribution of Taylor & Flood findings is in setting an agenda for future research was welcomed by both Andrew Rose (University of California, Berkeley, NBER and CEPR) and Lars Svensson (Institute for International Economic Studies, Stockholm, NBER and CEPR). Although minute in comparison with daily turnover, exert significant effects. Studies adopting this microstructure approach should focus on how and why speculative attacks begin and why official interventions. Besides, the importance of such interventions' secrecy, the roles of rumors and heterogeneous beliefs, the effects of exchange rate volatilities on hedging for trade, and the design of monetary policy to take account of markets' microstructure, they all should also be assessed. The reason why the data usually reject the PPP theory was noted by Jeffrey Frankel, that prices exhibit insufficient variance in the sample periods, and that PPP theory is stated as tends to work better during hyperinflation periods [15].

## Text Mining

Keyword-Based Association Analysis is an analysis done by collecting keywords or terms, which often appear at the same time. Then try to find the association relationships as well as the correlations between the keywords or the terms.

Like any other analysis in text databases, in the beginning, association analysis is started by doing pre-processing of text data through parsing, stemming, removing stop words, and so on. Afterwards then the algorithm association mining is run. In a database document, each document can be taken as a transaction, whereas a group of keywords in the document can be treated as a group of items in a transaction, so the database will have a format as follow [28]: {document\_id, a\_set\_of\_keywords}. This kind of format is usually called as Compact Transaction format.

## Parsing

In grammar and linguistic terms, parsing is a process which is done by someone, to make a sentence become more significant or meaningful, by dividing the sentence into words or phrases [29].

## Stop words, Stop list, Stop words Removal

Generally a formal language in any country has function words and connecting words, such as articles and prepositions which almost always appear in text documents. Usually these words do not have more meaning in fulfilling a researcher needs in searching for information. The words are (e.g.: a, an, the, on - in English) which are called Stop words [30].

A Text Retrieval System is usually accompanied by a stop list. Stop list contains of a group of words which are “irrelevant”, but very often appear in a document. In other word, Stop list contains a group of Stop words [28].

Stop words removal is a process to omit “irrelevant” words at parsing result of a text document, by comparing them with the available Stop list.

## Stemming

Stemming is a mapping process and separating various variants of a word into the stem word [31]. This process is also called as conflation [27]. Stemming process has been widely used in Information Retrieval, to increase the quality of the information obtained [26, 27].

#### Naïve Bayes Method for Text Classification

The outline of Naïve Bayes Method for text classification is as displayed below:

At NBC, every news document is presented in paired attribute ( $a_1, a_2, \dots, a_n$ ) where  $a_1$  is the first word,  $a_2$  is the second word, and so on. Whereas  $V$  is the group of news category (sport, science technology and so on). At classification phase, Bayes approach will result in category labels which probability is the highest (VMAP), by including attribute ( $a_1, a_2, \dots, a_n$ )

$$v_{MAP} = \arg \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n) \quad (2.1)$$

Bayes theorem said:

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)} \quad (2.2)$$

Using this Bayes Theorem, the equation (2.1) can be written as

$$v_{MAP} = \arg \max_{v_j \in V} \frac{P(a_1, a_2, \dots, a_n | v_j) P(v_j)}{P(a_1, a_2, \dots, a_n)} \quad (2.3)$$

$P(a_1, a_2, \dots, a_n)$  values are constant for all  $V$  so that this equation can be written as follows:

$$v_{MAP} = \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j) P(v_j) \quad (2.4)$$

Difficulty level to calculate  $P(a_1, a_2, \dots, a_n | v_j)$  becomes high, as term quantity  $P(a_1, a_2, \dots, a_n | v_j)$  can be very big. It is due to the total term is equal to all word position combination times the number of categories.

Naïve Bayes Classifier simplified this by assuming that in each category, every word is independent to one and another. In other words:

$$P(a_1, a_2, \dots, a_n | v_j) = \prod_i P(a_i | v_j) \quad (2.5)$$

Substituting this equation with equation 2.4 will result in:

$$v_{MAP} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j) \quad (2.6)$$

$P(v_j)$  and probability for word  $w_k$  for every category  $P(w_k | v_j)$  calculated during exercise:

$$P(v_j) = \frac{|docs_j|}{|Contoh|} \quad (2.7)$$

$$P(w_k | v_j) = \frac{n_{kj} + 1}{n_k + 1}$$

Where  $|docks|$  is number of words at  $J$  category, and  $|Canto|$  is number of document used in exercise. Whereas  $n_k$  is the number of word appearance  $w_k$  at  $VJ$  category,  $n$  is all of the words in  $v_j$  category, and  $|kosakata|$  (= vocabulary) is the number of words which is unique (distinct) at all of the exercise.

Algorithm summary for Naïve Bayes Classifier is as follows:

A. Exercise Process. Inputs are sample documents which category is known.

1. Group of Vocabulary  $\beta$  are all unique words from the sample documents
2. for each category  $v_j$  do:
  - a) Docs  $\beta$  Group of documents which are at category  $v_j$
  - b) Calculate  $P(v_j)$  with the equation 2.7
  - c) For every word  $w_k$  at the vocabulary, do:
    - i. Count  $P(w_k | v_j)$  with equation 2.8

B. Classification Process. Inputs are documents which category is unknown:

Produce  $vmap$  accordingly to the equation 2.6 by using  $P(v_j)$  and  $P(w_k | v_j)$  which has been obtained from the exercise.

## Research Methodology

The research method done was as follows:

### Data set

300 web documents were taken as data for each of the 9 categories from the website <http://www.bloomberg.com/>. This website acts as financial news givers, as well as others in trustworthy international world. This website contains of about eight millions (8,000,000) documents or so, in correspondence to all news program; which many people browse the news document at the website. Nine (9) category used are: business, Economic Monetary System (EMS), Cost of Living, Average Earning, Federal Reserve Fund, Gross National Product (GNP), Prime Rate, Non Farm Payrolls, and Discount Rate.

Actually news has about 46 categories, but 9 categories are taken as samples for this study. Euro /Dollar data are two currencies taken as application or implementation to find the effect obtain. The data were taken from the platform of Millennium Trader, quoted from 07May 2007 to 16 June 2008.



Graph. 1 Euro/Dollar Currency Exchange Movements  
Source: Millenium Trader Platform (07 May to 06 June 2008)

### News Category

In this study, news will be classified in 9 categories, as are shown at Table 1 below. These categories are classified from the effects arise and the influence towards forex in Euro/Dollar markets. These categories were chosen as the news are more often published.

Table 1. Categories in Forex

| NO. | Economic Indicator                | Increase/ Decrease<br>US\$ |          |
|-----|-----------------------------------|----------------------------|----------|
| 1   | Business                          | Decrease                   | Stronger |
| 2   | Economic Monetary<br>System (EMS) | Increase                   | Stronger |
| 3   | Cost of Living                    | Increase                   | Stronger |
| 4   | Average Earning                   | Increase                   | Stronger |
| 5   | Federal Reserve Fund              | Increase                   | Stronger |
| 6   | Gross National Product<br>(GNP)   | Increase                   | Stronger |
| 7   | Prime Rate                        | Increase                   | Stronger |
| 8   | Non Farm Payrolls                 | Increase                   | Stronger |
| 9   | Discount Rate                     | Increase                   | Stronger |

Furthermore, the analysis needs to be classified into the various news, so the study segmentation is focused and the directions are more organized, not randomly taken. As it was stated by Situngkir (2003), social system can be differentiated into several descriptive levels as the objects of the study. There from of the 46 categories available, only nine categories were taken to be studied on their influences which affected the market.

## Document Extraction

The objective of extracting process is to produce terms which will be used as prototypes of each document. Each term is sought for its stems of words. The procedures are based on dictionary of language stems of words. The purpose is to prevent the usage of words which have the same stems of words, but have different affixes and prefixes. Besides, filtering was done to words which were not proper to be treated as the variants. These groups of words are usually called as stop lists. As they are not readily available, thus in this study stop lists had to be found manually.

After extraction, the news was grouped into three parameters, i.e.: “good”, “bad”, and “no increase”. In this study, the three parameters are called labels, they are meant for observing clearly the effects on the currency exchange movements influenced by the news

## Calculation of TF-IDF Weight

In this phase, each document was transformed into vector, with many elements, as many as terms which can be recognized from the document extraction stage above. The vector consisted of weight of each term, which was calculated based on TF-IDF method. This TF-IDF method is a weighing method in a the method integrated within Term Frequency (TF) and Inverse Document Frequency (IDF) [7][15].

The formula is:  $w(t,d) = tf(t,d) * \log_2(N/nt)$ . The Symbol  $w(t,d)$  is the weight of term  $t$  in document  $d$ , whereas  $tf(t,d)$  is the term frequency in the document ( $tf$ ).  $N$  is the size of data exercise which is used to calculate IDF, and  $nt$  is the number of documents which are exercised and has  $t$  value.

The function of this method is to find the representative value of each document in a group of exercise set. There from a vector between documents with terms used to find similarities of the documents and their clusters, which will be based on the vector prototype, known as cluster centroid [6].

## Classification

Similarities resulted were then evaluated to decide paired documents which were stated as similar, based on a specific threshold values. News documents were classified by Naïve Bayes Classification Algorithm.

## Classification Evaluation

This Evaluation was done for the purpose of understanding the performance of classification algorithm at the trial stage. The measurement was based on two parameters: recall and precision.

## Findings

The findings of manual evaluation of the documents are put into documents in the clusters. These Clusters are the 9 categories discussed above. They are then classified into three groups, they

are: “good”, “bad, and “no movement”. After that, 2 parameters are used to evaluate the classifications as follows:

1. Recall: a success rate in recognizing an event of all the events available. The formula is:  $r = a/(a+c)$  for  $a+c > 0$ . Other than that is not defined.
2. Precision: an accuracy of clarification result to an event. Which means knowing the result percentage of the correct clarified documents? The formula is  $p = a/(a+b)$  when  $a+b > 0$ . Other than that is not defined.

Table 2. Means of Recall and Precision

|               | Good News |        | No Movers |      | Bad News |      | Overall<br>(Weighted<br>Recall) |
|---------------|-----------|--------|-----------|------|----------|------|---------------------------------|
|               | Prec.     | Rec.   | Prec.     | Rec. | Prec.    | Rec. |                                 |
| <i>Avg.</i>   | 70%       | 72%    | 40%       | 30%  | 60%      | 64%  | 58%                             |
| <i>Min.</i>   | 62%       | 60.40% | 40%       | 44%  | 55%      | 55%  | 54%                             |
| <i>Max.</i>   | 80%       | 80.30% | 70%       | 61%  | 60%      | 61%  | 60%                             |
| <i>StDev.</i> | 7%        | 7%     | 4%        | 3%   | 6%       | 6%   | 5%                              |

Notes;

Avg = Average

Min = Minimum

Max = Maximum

Prec. = Precision

Rec. = Recall

StDev = Standard Deviation

From the results above, it can be seen that the average memorizing of this algorithm is around 72% and 64%, as are shown at Table 2. Based on the results, the ability of the algorithm to recognize an event or news documents are very good. The existence of good recognition or good memorizing, the next news documents will easily find the level or value for the next calculation or next analysis.

Furthermore, the recalls showed very good results: about 70% for “Good News” and 60% for “Bad News”. The accuracy of this algorithm was very good to recognize or classifying the news into label “good news” and “bad news”. Thus, it is easier to decide the following actions in purchasing or selling Euro/Dollar. The effects which will be gained from this result can be up to +7.2%, which shows a significant influence in deciding final price. The result is in accordance with the macroeconomics theory, as it is explained earlier in the the literature part of this paper, that there is relationships between trade volume and risk; between information and trade intensities. In more detailed, the exchange rate related issues, such as the behaviors of the market participants during the EMS crisis since 1992, the mechanisms predict speculatively, and capital’s affectivities are designed to control the crisis.)

The results are applied in the basic trading, and the outcomes are as follows:

Table 3. Average Buy and Sell Recommendation

|             | Buy<br>Recommendations | Short<br>Recommendations |
|-------------|------------------------|--------------------------|
| <i>Avg.</i> | 45%                    | 43%                      |
| <i>Min.</i> | 19%                    | 17%                      |

|        |     |     |
|--------|-----|-----|
| Max.   | 40% | 35% |
| StDev. | 6%  | 6%  |

From the table above, buy recommendation in Euro/Dollar around 45% in all transaction result is not a recommendation to buy at the next transaction, but the average buy which was analyzed by Naïve Bayes Method.

The average Sell Recommendations for the dates of 07 May to 06 June 2008 were about 43%. The above results explained that the recommendation gave about 50% of the Euro/Dollar transaction movements.

## Conclusions

The study revealed that the news calculation using Naïve Bayes Method has a significant influential power of about 80-90%, after news classifications and its testing. These results were obtained manually, as there has been no specific system available which could handle this matter yet. Therefore the calculation may still have to be improved to avoid the possible human error in it.

Nevertheless, usage of Naïve Bayes Method for text classification and influence analysis in fores trading was very simple, yet the results obtained was also about 50% for recommending buy or sale. Moreover, categorizing and labeling usage for every news have the levels or grouping which was very effective and efficient in getting the results to be processed.

This research can be very prospective for the creation of software using Naïve Bayes Classification with a better digitalized algorithm computerization, so that human error can be minimized much well than the manual system. Thus, having the application or trading algorithm distributed software will enable investors as well as brokers to analyze more easily, while the speed and the accuracy could be enhanced.

## References

- [1] S. Lawrence, I. Burns, A. D. Back, A. C. Tsoi, and C. L. Giles, 1998, "Neural Network Classification and Prior Class Probabilities", G. Orr, K.-R. Müller, and R. Caruana, (eds.), *Tricks of the Trade*, Springer, Heidelberg, pp. 299-314.
- [2] H. Brücher, G. Knolmayer, and M.-A. Mittermayer, 2002, "Document Classification Methods for Organizing Explicit Knowledge", *Proceedings of the 3rd European Conference on Organizational Knowledge, Learning, and Capabilities*, ALBA, Athens,
- [3] J. Bollen, Text Operations, URL: [http://www.cs.odu.edu/~jbollen/spring03\\_IR/cs695\\_lect6.pdf](http://www.cs.odu.edu/~jbollen/spring03_IR/cs695_lect6.pdf) [2003-03-20].
- [4] Yiming Yang, Jaime G. Carbonell, Rulf D. Brown, Thomas Pierce, Brian T. Achibald, Xin Liu., 1999, "Learning Approaches for Detecting and Tracking News Events". *IEEE Intelligent Systems*, Language Technologies Institute, Carnegie Mellon University

- [5] Ron Papka, 1999, "On-Line New Event Detection, Clustering, and Tracking". Ph. D dissertation on University of Massachusetts
- [6] J. Allan et al, 1998, "Topic Detection and Tracking Pilot Study : Final Report." Proc. DARPA Broadcast News Transcription & Understanding Workshop, Morgan Kaufman, San Francisco, pp194-218
- [10] Ricardo Baeza-Yates, Berthier Ribeiro- Neto, 1999, "Modern Information Retrieval" ACM Press, Addison Wesley Longman Limited, New York
- [11] Anil K. Jain, Richard C. Dubes, 1988, "Algorithms for Clustering Data", Prentice-Hall Advances Reference Series
- [12] Pakana, Fitrio, 2001 "Perancangan Dan Pembuatan Aplikasi Pencarian Dokumen Berbasis Web Dengan Penerapan Metode Suffix Tree Clustering Pada Result Set", Tugas Akhir, Teknik Informatika, Institut Teknologi Sepuluh Nopember Surabaya,
- [13] Weiss, Mark Allen, 1999, "Data Structures & Problem Solving Using Java", Florida International University, Addison Wesley
- [15] <http://www.cepr.org/pubs/bulletin/meets/1231.htm> Acces at : June 14 2008
- [16] B. Wüthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, and W. Lam, 1998, "Daily Stock Market Forecast from Textual Web Data", Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, IEEE Computer Society Press, Los Alamitos, CA, pp. 2720-2725
- [17] E. F. Fama, 1995, "The Behavior of Stock-Market Prices", The Journal of Business 38, The University of Chicago Press, Chicago, IL, pp. 34-105.
- [18] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, 2000, "Language Models for Financial News Recommendation", Proceedings of the 9<sup>th</sup> International Conference on Information and Knowledge Management, ACM Press, New York, NY, pp. 389-396.
- [19] G. Gifdofalvi, Using News Articles to Predict Stock Price Movements, URL: [http://www.cs.ucsd.edu/users/gyozo/studies/cse\\_254\\_AI/stock\\_price\\_prediction.pdf](http://www.cs.ucsd.edu/users/gyozo/studies/cse_254_AI/stock_price_prediction.pdf) [2001-06-15].
- [20] T. Joachims, SVM Light Classifier, URL: <http://svmlight.joachims.org/> [2002-07 05].
- [21] S. Lawrence, I. Burns, A. D. Back, A. C. Tsoi, and C. L. Giles, 1998, "Neural Network Classification and Prior Class Probabilities", G. Orr, K.-R. Müller, and R. Caruana, (eds.), Tricks of the Trade, Springer, Heidelberg, pp. 299-314.
- [22] C. W. J. Granger and O. Morgenstern, 1970, "Predictability of Stock Market Prices", Heath Lexington Books, Lexington, MA

- [23] J. Roman and A. Jameel, 1996, "Backpropagation and Recurrent Neural Networks in Financial Analysis of Multiple Stock Market Returns", R. H. Sprague (ed.), Proceedings of the 29th Hawaii International Conference on System Sciences, Vol. 2, IEEE Computer Society Press, Los Alamitos, CA, p. 454.
- [24] R. J. van Eyden, 1996, "The Application of Neural Networks in the Forecasting of Share Prices," Finance and Technology Publishing, Haymarket, VA
- [25] V. Niederhoffer and M. Osborne, 1996, "Market making and Reversal on the Stock Exchange", Journal of the American Statistical Association 61, 316, ASA, Alexandria, VA, pp. 897-916.
- [26] Frakes, Bill, 1999, What is Information Retrieval, Computer Science Department. Virginia : Tech.
- [27] Frakes, W.B., R. Baeza, 1992, Information Retrieval, Data Structures and Algorithms. Prentice Hall.
- [28] Han, Jiawei, Micheline Kamber, 2001, Data Mining : Concepts and Techniques. Morgan Kaufmann.
- [29] Parsing, Yotor.com, 2005. <<http://www.yotor.com/wiki/en/pa/Parsing.htm>>
- [30] Chakrabarti, Soumen, 2003, Mining the Web: Discovering knowledge from hypertext data. San Francisco: Morgan Kaufman.
- [31] Porter, M. F. ,1980, An algorithm for suffix stripping, Program 14(3), p. 130-137.

