Prediction of growth rate of operating income using securities reports

Nakatoh, Tetsuya
Research Institute for Information Technology, Kyushu University

Amano, Hirofumi
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：Proceedings - 2nd IIAI International Conference on Advanced Applied Informatics, IIAI-AAI 2013, pp.84-88, 2013-12-16
バージョン：
権利関係：

Prediction of Growth Rate of Operating Income using Securities Reports

Tetsuya Nakatoh, Hirofumi Amano and Sachio Hirokawa
Research Institute for Information Technology,
Kyushu University.
Email: {nakatoh, amano, hirokawa}@cc.kyushu-u.ac.jp

***Abstract*—Corporate analysis is needed for various purposes such as finding a good business partner or a good employment, as well as choosing a good investment. Conventionally, it has been based mainly on financial figures. Recent advances in natural language processing technology, however, has activated studies on analysis of non-financial, textual data. This paper tries to predict the growth rate of the operating income of a company from text data contained in the security report of that company. It reports that this method can classify profitable companies and loss-making ones at 55% F-measure.**

I. Introduction

Corporate analysis is needed for various purposes such as finding a good business partner or a good job, as well as choosing a good investment. It analyzes various information related to a specific company and tries to estimate that company's prospects, from medium- and long-term perspectives.

In general, corporate information contains qualitative and quantitative data. Conventionally, corporate analysis focuses on quantitative analysis based mainly on financial figures. It judges the financial condition or the business showings of a corporation from the financial statements, points out latent problems, predicts future performance or stock price, and even gives a waring of a possible insolvency. These techniques have been explained in textbooks, and we can buy commercial systems which help such analysis.

These financial figures of companies listed on the stock exchange can be obtained mainly from their security reports. Those listed companies have a duty to report their financial condition every year in the security reports, whose formats are defined by the law. Therefore, security reports are the most fundamental report of a firm. Firms also publish their information such as Investors Relations pages on their web sites. The security reports are, however, the most reliable source of information on that firm since the items to report are defined by the law and firms have a legal duty to submit their reports every year. UFOReader [1] or UFOLenz [2] provided by Lafla Inc. [3] are examples of web services for financial statements analysis.

Large scale analysis by computers is feasible when analyzing financial figures, while qualitative data such as text must be read and judged by men, and thus qualitative analysis was not performed in large scale. Recent advances in natural language processing technology, however, has activated studies on analysis of non-financial, textual data. Now, stock price, business performance or insolvency risk can be predicted based on the analysis of a large amount of text data.

In this paper, we attempts to predict business performance of a company and find out the reason of a change in performance, by analyzing the text data contained in the financial statements.

II. Literature Review

Many research works focus on the immediacy of textual information such as net news, blogs or twitter. For example, if adequate analysis is performed against some tweets mentioning the financial condition of a company, it will give us more timely prediction of its stock price. Some researchers work on stock price prediction by classifying such tweets into positive ones and negative ones and calculating their emotional scores [2], [18]. Similar studies on positive/negative classification are carried out against newspaper articles[14]. There are also works on the correlation between the number/lengths of comments and the stock price fluctuation [4], [5]. [22] focused on themes attached to news articles, and analyzed the correlation between the stock price data and the articles by examining the impacts of articles on the stock price, as well as external factors of the stock price fluctuation. [21] performed the sentiment analysis of stock exchange market news. [16] utilized *SVM* to predict stock prices from news articles, and showed that the proposed approach analyzing both article words and stock prices had achieved the best performance in predicting the stock price 20 minutes after the article release.

These studies focus on the immediacy of the information source and attempt to estimate the short-term change of stock prices. In contrast, we focus on the analysis and the estimation of a company's prospects using text information in the security report, from medium- and long-term perspectives.

Other studies put more emphasis on accuracy rather than immediacy, and analyze non-financial, textual data in annual reports or similar documents. [15] analyzed the correlation between stock prices and IR information on web sites for pharmaceutical and companies. [6], [1] proposed an analyzer system which utilizes insolvency reports. From insolvency articles for companies selected in a specific criteria, it extracts

[1] http://www.uforeader.com/v1/
[2] http://www.yano.co.jp/ufolenz/index.php
[3] http://www.lafla.co.jp/

sentences mentioning the reason of the insolvency and visualizes the reason with the help of feature words contained in those sentences. [13] developed an SVM-based prediction model capable of implementing various feature selection mechanisms from annual reports for 10 years, and showed that the document frequency threshold is effective in reducing the feature space while preserving an equal classification accuracy. They also showed that this approach can predict the financial performance (i.e., return on equity) for a company from its annual report for the previous year. [7] extracted a feature vector from monthly reports of a bank, and performed regression analysis for government bonds, stock prices and foreign exchange. Moreover, [8] compared the proposed approach with support vector regression, and showed that their approach can achieve more accurate prediction of the stock market trend.

There exists some studies focusing on security reports, which are more reliable source of information concerning a company. [12] visualized the relationship among feature words common in the security reports of growing U.S. companies, and analyzed their stoch prices. [17] narrowed the target text to the dividend policy sections of security reports by their empirical knowledge. By text mining, they clarified the difference between the companies which became insolent and those which did not. [19] ranked the relative insolvency risks of companies by machine learning against text information in their security reports. [9] developed a regression model from a security report to predict the stock price volatility (risk) for the subsequent period. To forecast stock price changes in a short period, [11] proposed an approach to develop a prototype combining numerical and textual information in security reports by assigning adequate weights to them. Their experiments showed that their approach achieved better accuracy and average profit than the SVM, the naive-Bayes, or the PFHC approaches. To predict the short-term fluctuation after the release of a security report, [10] proposed the HRK (Hierarchical agglomerative and Recursive K-means clustering) approach which selects feature vectors by clustering. Experiments showed that their approach achieves better accuracy and more average profit than SVM-based approaches.

All these studies using security reports focused on stock prices and their risks. However, little attention has been paid to the growth of a firm such as growth rates of operating incomes. [20] extracted common features from the security reports of companies continuing its growth. [3] developed a new model by combining risk words information into an existing performance prediction method based on numerical information, and showed that it improves the accuracy of performance prediction.

In contrast, our approach utilizes only textual information in security reports to predict the income growth rate of a company, and extracts the reasons of such growth. This challenge is harder than other studies dealing with only extreme cases such as insolvency, since we attempt to estimate a variable growth rate. Moreover, a security report itself may cause the change in the stock price and thus a close correlation can exists inherently between them. On the other hand, the growth rate can be a more objective prediction measuring a company's prospects than the stock price, since it has nothing to do with the security report for the previous year. By putting numerical data aside, our approach can extract feature words from the security report as the reason of an income growth.

III. Security Reports and Growth Rate

A. Security Reports of Pharmaceutical Companies

A security report is a document reporting important items concerning the financial condition and other business profiles, and other items necessary and adequate for public good, both of which the Cabinet Office Ordinance specifies as such, and must be published by every company which issues stocks and bonds. A company has a legal duty to submit the security report to Prime Minister within three months after every fiscal year, and a false description in a security report results in a criminal charge. This makes a security report the most reliable source of information concerning a company, and thus we chose it as the target of analysis.

The format of each security report is divided into two parts. Part I gives us the information about the company which published the report, while Part II contains the information about the surety companies for the publisher of the report. Part I describes the profiles and the risks of the business, and tells us important information in finding out the financial condition of the company. In this paper, items containing enough amount of text for analysis are chosen as the target.

B. Preprocessing

Lafra Inc. provided us security reports in the form of machine-readable text data and numerical data containing 34,720 documents of 4,493 companies in total, in 76 business categories from Year 2006 to Year 2012, .

In the first step of the analysis, we narrowed the business categories into 68 pharmaceutical companies (Category Code: S0300). From 451 documents obtained by this elimination, . we selected 285 reports containing sales and profit information, then further narrowed them into 223 documents which have enough information for calculating an income growth rate. The latest security reports of each company are eliminated at this step since the growth rate cannot be calculated from them.

C. Income Growth Rate

Fig. 1 plots the individual income growth rates obtained from those security reports in their descending order. It shows that approximately half of the profits of those companies during period were growing.

Fig. 2 plots the positive income growth rates in the logarithmic scale. Positions 3 through 20, and Positions 20 through 70 show an almost linear distribution. Though this conjecture is not within the objectives of this paper, we might say that the Power Law applies to income growth rates in a way similar to other social phenomena.

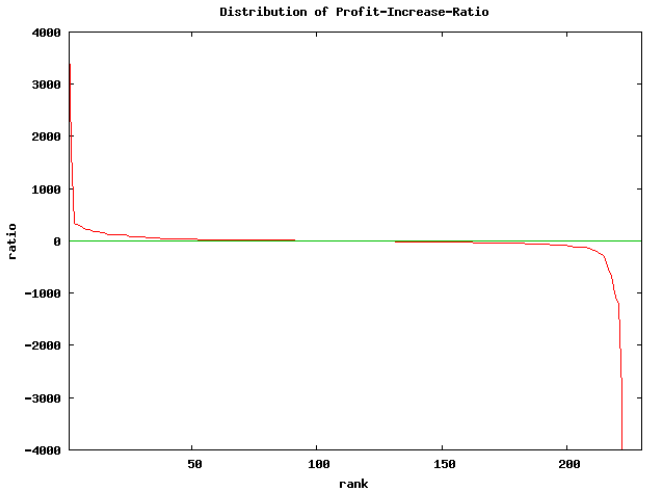

Fig. 1. Distribution of Growth Rate of Operating Income

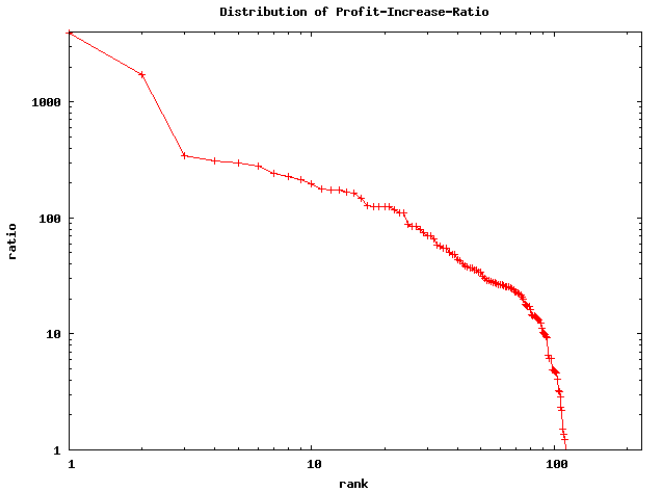

Fig. 2. Distribution of the Growth Rate of Operating Income

IV. Prediction of Growth Rate of Operating Income using SVM

A. 10-fold Cross Validation of Top α Reports

Then, we analyzed 223 security reports of pharmaceutical companies, focusing on words which have 2% to 50% occurrence in those documents. Let α be a threshold value which divides financial reports into two sets by growth ranking. Then, we ran an SVM learning procedure to the sets. We obtained F-measure and accuracy by the 10-fold cross validation.

The security reports of the same company often contain similar descriptions such as those in that company's history. If words specific to a particular company have much occurrence in its security reports, they are very likely to appear as false "feature words" rather than common words which really suggests growth of operating income. To avoid this problem in our experiment, we made sure that security reports of the same company do not appear both in the learning data and the test data. First, we randomly reserved 10% security reports as the test data and identified which company published those reports. Then, we eliminated the security reports of those companies from the remaining 90% reports, and fixed the learning data. After this process, the test data and the learning data never contained the same company in common.

B. Performance Evaluation

Fig. 3 shows the F-measures obtained in the experiment.

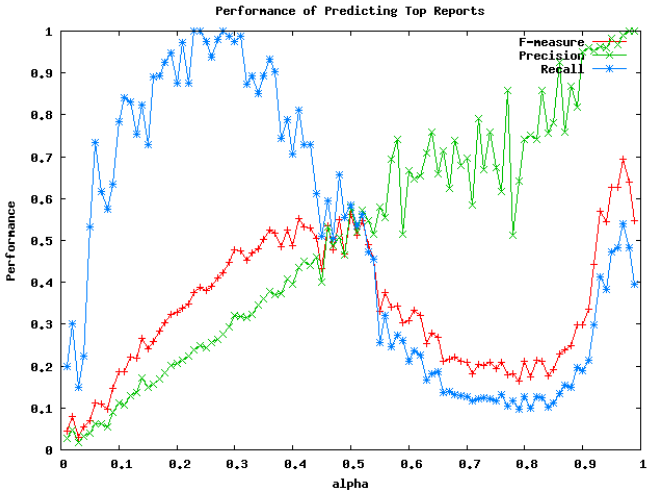

Fig. 3. Performance of Predicting Top Documents

We can interpret Fig. 3 together with Fig. 2 as follows. The precision increases monotonically, simply because the number of reports increase that have the growth rate smaller than α. The move of F-measure is mainly determined by the recall. The recall gains the best performance around $\alpha = 0.25$. Then it drops until $\alpha = 0.8$. As far as reports with positive growth rate, the F-measure is around 0.55, which is not too bad.

V. Feature Words of Reports with Positive Growth Rate

The upper part of Table I shows the top 5 feature words of reports with top 10% growth rates. The lower part of Table I shows the feature words for the reports whose growth rates are in the range between 20% and 90%. The feature words of the Table I are translated in English from the original Japanese words manually by the authors. The words with the star marks are proper nouns. Most of them are the medical-supplies names and the company names.

VI. Stability of Classification

The stability of effectiveness evaluation is guaranteed by the 10-fold cross validation for each threshold α. However, it does not imply that the similar threshold returns almost the same feature words. In this section, we check if the set of feature words obtained from SVM is stable with respect to the change of threshold α. We extracted the feature words in the reports which be selected with the threshold α. Then we analyzed the ranking and the scores of the words in other threshold β.

TABLE I
FEATURE WORDS

α	Feature Word
0.01	Kainos*,KirinPharma*,Bistner*,urea,SkyePharma*
0.02	Paltac*,Kainos*,KirinPharma*,Bistner*,urea
0.03	pentasa,Paltac*,Kainos*,Bistner*,KirinPharma*
0.04	Paltac*,pentasa,Kainos*,OhtsukaChemical*,bridge
0.05	Cephalon*,aggravation,bendamustine,myeloma,OhtsukaChemical*
0.06	Cephalon*,aggravation,bendamustine,myeloma,Paltac*
0.07	Cephalon*,aggravation,bendamustine,myeloma,HSP
0.08	Cephalon*,aggravation,bendamustine,myeloma,pentasa
0.09	Cephalon*,aggravation,bendamustine,University,myeloma
0.10	Cephalon*,aggravation,bendamustine,myeloma,University
0.20	University,cornea,erythropoietin,epithelium,Scientific
0.30	University,cornea,lactamase,erythropoietin,epithelium
0.40	lactamase,Sepracor*,University,erythropoietin,cornea
0.50	Sepracor*,erythropoietin,lactamase,University,Paltac*
0.60	erythropoietin,University,Sepracor*,Paltac*,collapse
0.70	University,Paltac*,norlevo,Therapeutics*,isolation
0.80	University,Paltac*,lumigan,Cmic*,arasenas
0.90	Paltac*,AF,Cmic*,embolism,arasenas

A. Word Rank at Threshold α

Table II displays the ranks of feature words that appear in Table I at each threshold β. A "dot" implies that the word is not in the top 5 with respect to the threshold β. The word "cornea", for example, which is the 2nd rank at $\beta = 0.20$, stays at the 2nd rank at $\beta = 0.30$ and drops to 5th rank at $\beta = 0.40$. Most words appear as the top 5 words in a continuous range of β. The ranks of the same word changes gradually as the threshold β changes. This implies that the set of feature words are stable with respect to the threshold.

TABLE II
RANK OF WORD AT ALPHA

β	.10	.20	.30	.40	.50	.60	.70	.80	.90
University	5	1	1	3	4	2	1	1	.
aggravation	2	.	.	.	.	.	.	.	.
bendamustine	3	.	.	.	.	.	.	.	.
Cephalon*	1	.	.	.	.	.	.	.	.
myeloma	4	.	.	.	.	.	.	.	.
erythropoietin	.	3	4	4	2	1	.	.	.
epithelium	.	4	5	.	.	.	.	.	.
Scientific	.	5	.	.	.	.	.	.	.
cornea	.	2	2	5	.	.	.	.	.
lactamase	.	.	3	1	3	.	.	.	.
Sepracor*	.	.	.	2	1	3	.	.	.
Paltac*	.	.	.	.	5	4	2	2	1
collapse	.	.	.	.	.	5	.	.	.
norlevo	.	.	.	.	.	.	3	.	.
isolation	.	.	.	.	.	.	5	.	.
Therapeutics*	.	.	.	.	.	.	4	.	.
lumigan	.	.	.	.	.	.	.	3	.
arasenas	.	.	.	.	.	.	.	5	5
Cmic*	.	.	.	.	.	.	.	4	3
embolism	.	.	.	.	.	.	.	.	4
AF	.	.	.	.	.	.	.	.	2

B. Move of Word Score with respect to Threshold α

Figure 4 displays the scores of top 5 feature words (University,cornea,lactamase,erythropoietin and epithelium) that characterizes the top 61 (=30%) security reports. They are in the line of $\alpha = 0.30$ at Table I. The top scored word "University" occurs in the sentences which concerns with the technology introduction contract to a university and with research result by a university. This word is effective to distinguish the reports within top 80% rank. The second ranked word "cornea" and the word "epithelium" of 5th rank co-occurs as "corneal epithelium" to explain applying-eyewash curative medicine. Thus they have the similar moves in the plot. They are independently occurs in the context of corneal regeneration epithelium sheet and the curative medicine of epithelial cancer. The score of the two words drops gradually after $\alpha = 0.40$. The drops implies that they are not effective to distinguish negative growth rate. The third ranked word "lactamase" appears as β-lactamase inhibitor that effects an antibiotic medicine to a drug resistant bacterium. The development and sale of the antibiotics preparations for injection which blended lactamase inhibitor have contributed to profits. The 4th ranked word "erythropoietin" appears in the security reports as the tablet by transgenics. The manufacturing-and-selling application for approval is performed as a renal anemia curative medicine in Japan.

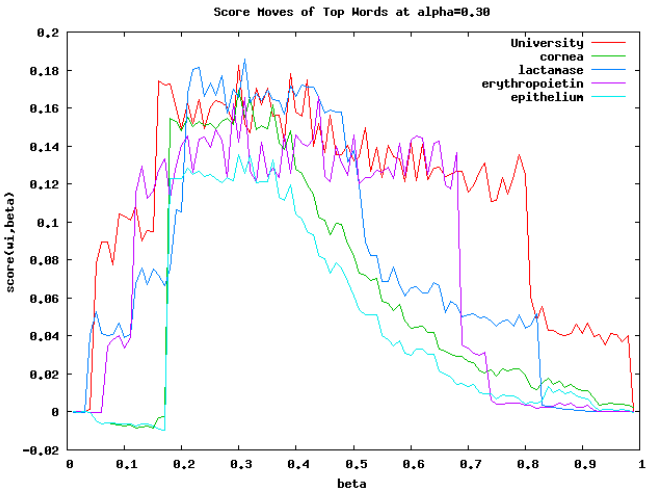

Fig. 4. Score Move of top 5 words at alpha=0.30

VII. CONCLUSIONS AND FURTHER WORK

The present paper applied SVM to estimate the growth rate of operating income from the textual information of security reports. As an empirical study, 223 security reports of pharmaceutical companies are analyzed. The growth rate of the next year is estimated from the security reports of the previous year. The prediction performance for higher increase companies was low. However, the predition performance for the companies with positive growth rate is around 55% which

seems to be not bad. The extracted feature words are used in the sentences that refers to the increase of profit

We used the security reports of the previous year to predict the growth rate. If we use not only one serurity report but also several reports that have been published in preceeding years will improve the performance of prediction. We plan to expand the analysis to many genres other than pharmaceutical companies and would like to compare the differences between genres.

This work was partially supported by JSPS KAKENHI Grant Number 24500176.

REFERENCES

[1] Takahiro Baba, Tetsuya Nakatoh and Sachio Hirokawa, "Text Mining of Bankruptcy Information using Formal Concept Analysis," Proc. of *3rd International Conference on Awareness Science and Technology (iCAST2011)*, pp.527–532, 2011.

[2] Johan Bollen, Huina Mao and Xiao-Jun Zeng, "Twitter mood predicts the stock market." *Journal of Computational Science*, Volume 2, Issue 1, March 2011, pp. 1–8, 2011.

[3] Kuo-Tay Chen, Tsai-Jyh Chen and Ju-Chun Yen. "Predicting future earnings change using numeric and textual information in financial reports," In Proceedings of *Pacific Asia Conference on Knowledge Discovery and Data Mining*, pp.54–63, 2009.

[4] De Choudhury, M., Sundaram, H.,John, A.Seligmann, D.D. "Can blog communication dynamics be correlated with stock market activity?", *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, HT 庁 08 with Creating 庁 08 and WebScience 庁 08*, pp. 55–59, 2008.

[5] Eric Gilbert and Karrie Karahalios, "Widespread Worry and Stock Market", *4th International AAAI Conference on Weblogs and Social Media(ICWSM)*, 2010.

[6] Sachio Hirokawa, Takahiro Baba and Tetsuya Nakatoh, "Search and Analysis of Bankruptcy Cause by Classification Network," *Model and Data Engineering, Lecture Notes in Computer Science*, Volume 6918/2011, pp.152–161, 2011.

[7] Kiyoshi Izumi, Takashi Goto and Tohgoroh Matsui, "Analysis of financial markets' fluctuation by textual information," *Transactions of the Japanese Society for Artificial Intelligence*, 25(3), pp.383–387, 2010.

[8] Kiyoshi Izumi, Takashi Goto and Tohgoroh Matsui, "Implementation tests of financial market analysis by text mining," *Transactions of the Japanese Society for Artificial Intelligence*, 26(2), pp.313–317, 2011.

[9] Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi and Noah A. Smith, "Predicting risk from financial reports with regression," *NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280, 2009.

[10] Anthony J.T. Lee, Ming-Chih Lin, Rung-Tai Kao and Kuo-Tay Chen, "An effective clustering approach to stock market prediction," *PACIS 2010 - 14th Pacific Asia Conference on Information Systems*, pp.345–354, 2010.

[11] Ming-Chih Lin, Anthony J. T. Lee, Rung-Tai Kao and Kuo-Tay Chen, "Stock price movement prediction using representative prototypes of financial reports," *ACM Transactions on Management Information Systems*, 2(3), No.19, pp.1–18, 2011.

[12] Kun Qian, Sachio Hirokawa, Kenji Ejima and Xiaoping Du, "A fast associative mining system based on search engine and concept graph for large-scale financial report texts," Proc. *2nd IEEE ICIFE (Information and Financial Engineering)*, pp.675–679, 2010.

[13] Xin Ying Qiu, Padmini Srinivasan and W.Nick Street, "Exploring the forecasting potential of company annual reports," Proceedings of *the American Society for Information Science and Technology* 43(1), pp.1–15, 2006.

[14] Hiroyuki Sakai, Shigeru Masuyama, "Estimation of Impact Contained in Articles about each Company in Financial Articles," *Information Processing Society of Japan (IPSJ)*, vol 94, pp.43–50, 2006. (in Japanese)

[15] Toshihiko Sakai, Masashi Matsushita, Brendan Flanagan, Jun Zeng and Sachio Hirokawa, "Analysis of influence of Investor Relation Documents to stock price", Proc. of *FSKD2012 : 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 1280–1284, 2012.

[16] Robert P. Schumaker and Hsinchun Chen, "Textual analysis of stock market prediction using breaking financial news: the AZFin text system," *ACM Transactions on Information Systems*, 27(2), pp.1–19, 2009.

[17] Cindy Y. Shirata and Manabu Sakagami, "An Analysis of the 'Going-Concern Assumption': Text Mining from Japanese Financial Reports," The Journal of *Emerging Technologies in Accounting, Strategic and Emerging Technologies Section of the American Accounting Association*, pp.1–16, 2009.

[18] Mike Thelwall, evan Buckley and Georgios Paltoglou, "Sentiment in Twitter Events," Journal of the *American Society for Information Science and Technology*, Volume 62, Issue 2, pp. 406–418, 2011.

[19] Ming-Feng Tsai and Chuan-Ju Wang, "Risk ranking from financial reports," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7814 LNCS*, pp.804–807, 2013.

[20] Shin-ichiro Yoshida, Tetsuya Nakatoh, Shuichi Mitarai and Sachio Hirokawa, "Text Mining of Securities Reports for Discovering Reason of Change, " Proc. of *CAINE-2012: ISCA 25th International Conference on Computer Applications in Industry and Engineering*, New Orleans, Louisiana, USA, 2012.11.14-16.

[21] Liang-Chih Yu, Jheng-Long Wu, Pei-Chann Chang and Hsuan-Shou Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Systems* 41(0), pp.89–97, 2013.

[22] He Zhang and Shigeki Matsubara, "Quantitative Analysis of Relevance between News Articles and Stock Price Change", *Information Processing Society of Japan (IPSJ)*, pp.183–184, 2008. (in Japanese)