Estimating financial words' negative-positive from stock prices

Keiichi Goshima[*] Hirohi Takahashi[†] Takao Terano[‡]

Abstract

In practical asset management business, institutional investors make their investment decisions by utilizing various kinds of information, including textual information and numerical data. This study analyses the relationship between textual information and financial markets in Japan. In analyzing the textual data, the subjectivity of the researchers play a significant role and could have a substantial impact on the accuracy of textual analyses when classifying articles into positive and negative. In this study, we propose the objective way to estimate financial text's and words' negative-positive. We analyze news articles written in both English and Japanese through the proposed method, which is one of the novelties of the work. As a result of intensive analysis, we made the following results: (1) we succeed in making new word lists which fit to the financial context by utilizing stock price information. Furthermore, (2) we construct classification system which predicted future stock prices by using new word lists and machine learning technique such as support vector regression. These results are suggestive from both academic and practical viewpoints.

1 Introduction

Many analyzing methods using the numerical information such as past asset prices or financial statements have been developed with the development of the financial theory. However, text information as well as numerical information is an important decision-making materials for investors. Text information might contain information that is not reflected in the numerical information, and the text analysis may be able to clarify market mechanisms which are not uncovered by only the numerical analysis. Therefore, since the mid-2000s, many researches have attempted to use a variety of text information for the asset pricing analyses[8].

When applying the text analysis to the asset pricing, it is necessary to judge positive-negative of textual contents. There are two major approaches to classify text information into negative and positive. One is the dictionary-based approach and the other one is the machine learning approach. The dictionary-based approach uses a mapping algorithm in which a computer program reads text and classifies the words, phrases or sentences into groups based on pre-defined dictionary categories[5]. The choice of words have a great influence on the result. The financial field text tend to use their unique vocabulary, and in order to create a dictionary specializing in finance, researchers need to choose words manually[12][7]. The machine learning approach apply the machine learning algorithm such as Support Vector Machine and Naive Bayes model to the classification text into negative and positive. The training set needs to be manually classified as positive-negative and the judging of training set have a great influence on the analysis results. Either part of the approach needs to proceed manually, and the subjectivity of the researchers has a significant

[*] PhD student, Tokyo Institute of Technology.

[†] Professor, Graduate School of Business Administration, Keio University, E-mail : htaka@kbs.keio.ac.jp.

[‡] Professor, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology.

impact on the accuracy of textual analyses. As one of solutions to these problems, there is the method for estimating negative-positive of text contents from the actual asset prices. Healy and Lo (2011) attempted to evaluate the news articles by using the foreign exchange[6].

In this study we developed Healy and Lo (2011) and estimated words' positive-negative from the stock prices. Furthermore, we constructed the classification system which predicts future stock prices by using past stock prices and the machine learning algorithm.

Our paper is organized as follows. Section 2 refer to data. Section3 datails our proposed method. In Section 4, we provided analysis results Some concluding remarks are offered in Section 5.

2 Materials

2.1 Market Data

In order to estimate news articles' negative-positive from stock prices we employed total return daily data and Japanese market factor's return data. The primary source for total return data is the Thomson Reuters Datastream. Japanese market factor's return data (Rm, Rf, HML, SMB) are token from the Nikkei Needs.

2.2 News Data

We employed the news data offered from Thomson Reuters which is one of the largest news agencies in the world. Our study focused on English and Japanese news articles in relation to Japanese stock market. We also used tag information relating to Reuters news (date time, Ticker symbol).

Leinweber (2009) distinguished three broad classifications of news: *News*, *Pre-news*, *Social media news*[11]. Each types of news have different information about the market. Reuters news is categorized into *News*. Journalists and Analysts of each media process the primary information. Processed information is broadcasted via TV, radio and newspaper as *News*. Therefore, *News* is screened by journalists and analysts, it might contain important information about the society and the market compared with *Pre-news*.

Reuters news is the media that many institutional investors in the Japanese stock market is viewing in real time. They have a small lag from the event to news announcement and might have more information that has not been reflected on asset prices than TV, radio and newspaper. Using Reuters among many text data, we could analyze between stock price fluctuations and text information. We focused on Reuters news written in English and Japanese related to companies listed to Tokyo Stock Exchange 1st Section from 2005 to 2011.

3 Methods

3.1 The analysis procedure

In this section, we discribed an outline of our analysis procedures. Fig.1 shows a flowchart of our analysis.

Firstly, we assigned scores (negative-positive) to each news article from stock prices. Using stock prices enable to classify articles into positive and negative objectively. Secondly, we computed vector representations of news articles based on bag-of-words. Thirdly, we made a machine learning classifier from $t-1$ year's news

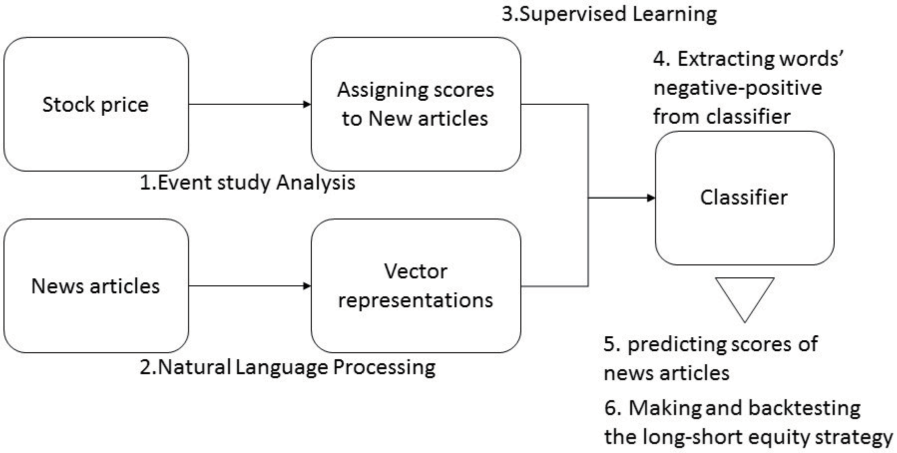
3.Supervised Learning
Stock price
Assigning scores to New articles
4. Extracting words' negative-positive from classifier
Classifier
1.Event study Analysis
News articles
Vector representations
2.Natural Language Processing
5. predicting scores of news articles
6. Making and backtesting the long-short equity strategy

Figure 1: The outline of our analysis procedure

articles attached scores to predict scores of t year's news articles. Fourthly, we attempted to extract words' negative-positive information from a machine learning classifier. Lastly, using scores predicted by a machine learning, we made the long-short equity strategies and backtested our supposed methods. We analyzed at the above procedures. In the following sections, we described the details for each the analytical method.

3.2 The way to assign scores

In our study we assigned score (negative-positive) to each news article from stock prices by the event study analysis[3].

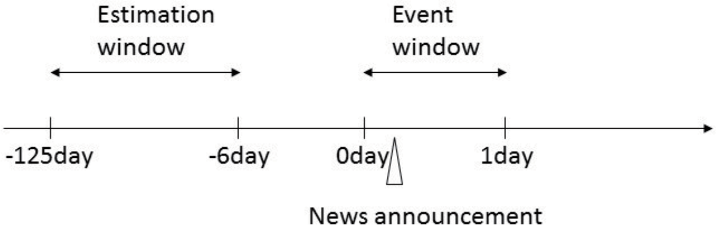

Figure 2: Time line for an event study

Fig.2 shows time line for an event study. The abnormal return is the actual ex post return of the security over the event window minus the norinal return of the firm over the event window. The normal return is defined as the expected return without conditioning on the event taking place. For news i and event date t, the abnormal return is

$$AR_{it} = R_{it} - E[R_{it}|X_t], \tag{1}$$

where AR_{it}, R_{it}, $E[R_{it}]$ are the abnormal, actual, and normal returns respectively for time period t. X is the conditioning information for the normal return model. We choiced Fama-French three-factor model for modeling the normal return. We estimated the factor model over the 120 days prior to the event. In this study the event window included the day of the news announcement and the day after the news announcement. We assumed that the cumulative abnormal returns for a 2-day event window (CAR(0,1)) were caused by news articles information. We analyzed news articles announced after 15 o'clock and news articles announced in the market closing day as the next market day's news articles.

Furthermore, we standardized abnormal returns in order to make supervised scores which follow the normal distribution. For news i and event date t the standardized abnormal return is

$$SAR_{it} = \frac{AR_{it}}{\sigma_{e_i}}, \tag{2}$$

where SAR_{it} is the standardized abonormal return for time period t, σ_{e_i} is a Residual standard deviation in the estimation window. We calculated the standardized cumulative abnormal returns for a 2-day event window (SCAR(0,1)) as supervised scores.

3.3 The way to compute vector representations of news aritcles

Our study computed vector representations of news articles based on bag-of-words model. We noted the processing of both English and Japanese text. In regard to the English news articles' analysis, firstly we removed stop-words, numbers and punctuation. Secondly we did stemming by Porter's stemming algorithm. Finally we weighted news articles' term by the tf-idf and the cosine normalization. In regard to the Japanese news articles' analysis, firstly we divided Japanese text into meaningful words because Japanese text had not already been divided by writers. Secondly we extracted noun, verb and adjective from text data. This processing provided the same effect as removing stop-words in the English text analysis. Finally we weighted news articles' term by the tf-idf and the cosine normalization.

3.4 The way to predict scores and to estimate words' negative-positive

We predicted scores to next year's news articles by support vector regression (SVR). SVR is the method of applying the support vector machine to a regression problem. SVR formulation is

$$\begin{aligned} \min_{w,\xi,\hat{\xi}} \ \tfrac{1}{2}||w^2 \quad & || + C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) \\ \text{subject to} \quad & w\phi(x_n) - y_n - \epsilon \leq \xi_n; \forall n \\ & y_n - w\phi(x_n) - \epsilon \leq \hat{\xi}_n; \forall n \\ & \xi_n \geq 0; \forall n \\ & \hat{\xi}_n \geq 0; \forall n. \end{aligned} \tag{3}$$

where $\phi(\,)$ is Kernel function, ξ_n and $\hat{\xi}_n$ are slack variables. The constant $C > 0$ determines the trade of between the fitness of function and the amount up to which deviations larger than ϵ are tolerated.

In this study we computed without Kernal methods bacause we regarded the primal variable w as words' negative-positive information and attempted to extract w. We chose parameters for SVR by the grid search.

3.5 The long-short equity strategy

Finally we developed and backtested the long-short equity strategy to verify the validity of our proposed method. When Scores attached by SVR were higher than $z_{0.975}(1.96)$, our method could predict that news articles have positive information.

We bought stocks related to news articles which were classified into positive by SVR. We bought stocks at the end of the news announcement day and sold at a next day. When our method predicted that more than one news articles were positive on the same day, we calculated arithmetic means of the stock returns. While when there was no positive news, we did not buy stocks.

When Scores attached by SVR were lower than $-z_{0.975}$(-1.96), our method could predict that news articles have negative information. We made the short selling stocks related to news articles which were classified into negative by SVR. We made the short selling stocks at the end of the news announcement day and bought back at a next day.

In this study supposing that we managed the same amount funds about long positon and short position, we calculated indices by the long-short strategy.

4 Results

4.1 The results of estimation words' negative-positive

We remark the results of estimation words' negative-positive. We regarded the primal variable w in SVR as word' negative-positive information and attempted to extract w. We made SVR classifier from all period news articles. The information of words' negative-positive may be able to apply qualitative analyses and other textual data.

Table.1 shows words lists which were gotten by our proposed method[1]. In this paper we put on only top 30 words as positive words and low 30 words as negative words. We could get some positive words : "heated", "speedy", "resurgent", and also could get some negative words : "falsified", "bottleneck", "depressing". These words generally have negative or positive information.

On the other hand, we also got individual company names and words which were difficult to hold the negative-positive information in the context of the asset pricing analysis. It may be able to increase the accuracy by improving of the machine learning and of the text mining. Improvement of these techniques, we want to do as future works.

4.2 The results of backtesting the long-short equity strategy

In this section we remark the results of backtesting the long-short equity strategy. We measured indices performance using Fama-French three-factor model. Table2 shows the results of indices performance which were made from Japanese news articles.

Indices from 2006 to 2009 could earn α. α of 2006, α of 2007, α of 2008 and α of 2009 were 0.274, 0.490, 0.620 and 0.439, respectively. Each α was statistically significant at the 1% level. Our proposed methods could get the excess earning even after taking factor returns into consideration. For Indices in 2010 and 2011, even though significant levels went down to 5% level, Indices could earn α. These results shows the high possibility of getting the excess earning through analyzing Japanese news articles.

Table3 shows the results of indices performance which were made from English news articles. Only index in 2007 could earn α (at the significant 10% level). Other indices could not earn α. These results shows the difficulty of getting the excess earning through analyzing English news articles.

[1] We uploaded Japanese words lists onto http://labs.kbs.keio.ac.jp/htakalab/word/.

Table 1: The results of estimation words' negative-positive

positive words	negative words
gainer	dating
whistleblower	birthrate
speedy	reacting
dubious	lofty
scraps	accelerators
acknowledge	falsified
delisted	bust
downs	averaging
boding	pages
disappeared	championed
botched	folded
kongs	trillions
surely	santa
resurgent	fourfold
eos	wellknown
hindered	perfect
leapt	halfowned
grapple	defaults
heated	bottleneck
forthcoming	cloudy
standpoint	strains
exacerbated	kicks
steer	doubted
toptier	halving
braking	retailings
jackets	abandon
featured	depressing
overcrowded	specifications
saddled	businessmen
haul	diluting

The cause of different results came from the difference of the way to write news articles. In this study we focused on news articles in relation to Japanese market, so English news articles might have lower information than Japanese news articles. We will try to analyze other countries' news articles.

5 Summary

Text information might contain information that is not reflected in the numerical information, and the text analysis may be able to clarify market mechanisms which are not uncovered by only the numerical analysis. Therefore, since the mid-2000s, some researches have attempted to use a variety of text information for the asset pricing analyses. In analyzing the textual data, the subjectivity of the researchers play a significant role and could have a substantial impact on the accuracy of textual analyses when classifying articles into positive and negative. In this study, we propose the objective way to estimate financial text's and words' negative-positive. We analyze news articles written in both English and Japanese, which is one of the novelties of the work. As a result of intensive analysis, we made the following results: (1) we succeed in making new word lists which fit to the financial context by utilizing stock price information. Furthermore, (2) we construct classification system which predicts future stock prices by using new word lists and machine learning technique such as support vector regression. These results are suggestive from both academic and practical viewpoints.

References

[1] Bishop, C. M. : Pattern Recognition and Machine Learning, Springer (2006).

[2] Bollen, J., Mao, H. and Zeng, X. : Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8 (2011).

[3] Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. : The Econometrics of Financial Markets, Princeton University Press (1997).

[4] Fama, E. F. and French, K. R. : Common risk factors in the returns on stock and bonds, *Journal of Financial Economics*, Vol. 33, pp. 3–56 (1993).

[5] Li, F. : The information content of forward-looking statements in corporate filings—A Naive Bayesian machine learning algorithm approach, *Journal of Accounting Research*, Vol. 48, pp. 1049-1102 (2010).

[6] Healy, A. and Lo, A. W. : Managing Real-Time Risks and Returns: The Thomson Reuters NewsScope Event Indices. In: Mitra, G. and Mitra L. (eds.), The Handbook of New Analytics in Finance, John Wiley & Sons, West Sussex, UK (2011).

[7] Henry, A. : Are investors influenced by how earnings press releases are written? , *Journal of Business Communication*, Vol. 45, No. 4, pp. 363-407 (2008).

[8] Kearney, C. and Liu, S. : Textual Sentiment in Finance : A Survey of Methods and Models, *International Review of Financial Analysis*, Vol. 33, pp. 171-185 (2014).

[9] Mitra, L. R., Mitra, G. and Bartolomeo, D. D. : Equity portfolio risk (volatility) estimation using market information and sentiment, *Quantitative Finance*, Vol. 9, No. 8, pp. 887–895 (2009).

[10] Tetlock, P. C. : Giving Content to Investor Sentiment:The Role of Media in the Stock Market, *Journal of Finance*, Vol. 62, No. 3, pp. 1139–1168 (2007).

[11] Leinweber, D. J. : Nerds on Wall Street, John Wiley & Sons (2009).

[12] Loughran, T. and McDonald, B. : When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *Journal of Finance*, Vol. 66, No. 1, pp. 35–65 (2011).

[13] Yamashita, Y., Jotaki, H. and Takahashi, H. : Analyzing the Influence of Head-Line News on the Stock Market in Japan, *International Journal of Intelligent Systems Technologies and Applications*, Vol. 12, No. 3-4, pp. 328–342 (2013).

Table 2: The result of backtesting the long-short equity strategy through analyzing Japanese news articles

	indexRt - Rf (2006)	indexRt - Rf (2007)	indexRt - Rf (2008)	indexRt - Rf (2009)	indexRt - Rf (2010)	indexRt - Rf (2011)
α	0.274***	0.490***	0.620***	0.439***	0.219**	0.271**
	(3.055)	(4.351)	(4.392)	(3.755)	(2.295)	(2.275)
Rm - Rf	0.049	-0.038	-0.023	0.006	0.187*	0.032
	(3.055)	(-0.335)	(-0.307)	(0.047)	(1.740)	(0.368)
SMB	-0.033	-0.541**	0.093	0.138	0.250	0.471**
	(-0.194)	(-2.089)	(0.530)	(0.565)	(0.943)	(2.394)
HML	0.057	-0.336	-0.093	0.478	0.116	0.475
	(0.186)	(-0.978)	(-0.364)	(1.631)	(0.410)	(1.432)
$adj.R^2$	-0.010	0.010	-0.007	0.001	0.001	0.024
Obs	247	244	244	242	244	244

***, **, *** denote two-tailed significance at the 10%, 5%, and 1% levels, respectively.**

Table 3: The result of backtesting the long-short equity strategy through analyzing English news articles

	indexRt - Rf (2006)	indexRt - Rf (2007)	indexRt - Rf (2008)	indexRt - Rf (2009)	indexRt - Rf (2010)	indexRt - Rf (2011)
α	0.020	-0.098*	-0.003	-0.065	-0.009	-0.051
	(0.517)	(-1.739)	(-0.039)	(-0.828)	(-0.2169)	(-1.213)
Rm - Rf	0.073*	-0.058	0.088*	0.074	0.039	-0.021
	(1.858)	(-1.013)	(1.894)	(0.886)	(0.834)	(-0.699)
SMB	0.090	-0.110	-0.039	0.197	-0.028	-0.308***
	(1.242)	(-0.843)	(-0.359)	(1.200)	(-0.237)	(-4.433)
HML	-0.186	-0.179	0.273*	0.151	-0.042	-0.072
	(-1.438)	(-1.037)	(1.737)	(0.769)	(-0.336)	(0.617)
$adj.R^2$	0.065	-0.006	0.035	-0.005	-0.006	0.072
Obs	247	244	244	242	244	244

***, **, *** denote two-tailed significance at the 10%, 5%, and 1% levels, respectively.**