Clustering News to Create Sentiment Indexes that Help Predict Stock Prices

HIROSHI ISHIJIMA[†1] TAKURO KAZUMI[†2]

Abstract: The purpose of this paper is to quantify the market sentiment as three indexes and examine whether they can help predict Japanese stock prices. Sentiment analysis is gaining increasing interest in both academia and business. Along these lines, Ishijima et al. (2014) created a sentiment index that quantifies the positive or negative emotion that might appear in entire headlines and articles of the Nikkei which is the most popular business newspaper in Japan. They concluded that the sentiment index significantly predicts Japanese stock prices three days in advance. We re-examine their results by suggesting a new sentiment index quantified from the headlines and articles limited to the economy-and-business news. To the best of our knowledge, this is the first paper that applies Latent Dirichlet Allocation (LDA) to cluster the Nikkei on an eight-year daily basis. We then explore the implication on how the new sentiment index can help explain Japanese stock prices. Our findings are three-fold: (i) Sentiment index created from the headlines and articles limited to the economy-and-business news significantly allows us to predict the Nikkei 225 and the market trading volume of the next business day. (ii) We cannot observe the return reversal referred to in the literature. (iii) Sentiment index will follow Japanese stock prices.

Keywords: Sentiment Analysis, Stock Price, Japanese Text Mining, The Nikkei

1. Introduction

The purpose of this paper is to quantify the market sentiment as three indexes and examine whether they can help predict Japanese stock prices.

As "sentiment" reflects the atmosphere of economic activities and the psychology of market participants, the sentiment analysis would allow us to understand the economy and stock markets in a more sophisticated way. Hence the sentiment analysis has been intensively studied in the finance and related literatures. Among these, Ishijima et al. (2014) created a sentiment index that quantifies the positive or negative emotion that might appear in entire articles of the Nikkei which is the most popular business newspaper in Japan. They concluded that the sentiment index significantly predicts Japanese stock prices three days in advance.

We re-examine their results by suggesting a new sentiment index quantified from the headlines and articles limited to the economy-and-business news and explore the implication on how the sentiment index can help explain Japanese stock prices. Our findings are three-fold: (i) Sentiment index created from the headlines and articles limited to the economy-and-business news significantly allows us to predict the Nikkei 225 and the market trading volume of the next business day. (ii) We cannot observe the return reversal referred to in the literature. (iii) Sentiment index will follow Japanese stock prices.

2. Creating Sentiment Index

2.1 Data and morphological analysis

In creating sentiment indexes, we used 2,843 daily issues of the Nikkei published during 96 months from January 2007 to December 2014. These archives are supplied by Nikkei Digital Media, Inc. On the other hand, as for Japanese stock prices, we used the daily closing prices of the Nikkei 225 to convert them into log returns. While the Nikkei is published daily and delivered with one no-issue day (Jan 2nd), the Japanese stock market is closed every weekend. To handle such a paired data set that is partially missing and hence inconsistent in frequency, we follow the approach of Mao et al. (2011). That is, we eliminated every Saturday and Sunday from the complete data set before implementing analysis. Hence we have about 21 days per month on average. The total number of headlines or equivalent that of articles is 603,063.

Due to a unique feature of the Japanese language, a *morphological analysis* is a prerequisite before conducting the sentiment analysis. As words are not separated with spaces in Japanese texts including the Nikkei, we first insert spaces to separate words. This is called the *morphological analysis*. To implement the morphological analysis, we employed MeCab 0.996 developed by Kudo et al. (2004).

2.2 Intuition behind the reason why the economy- and-business news is important?

Before implementing the rigorous analysis, we show an intuition why we focus on the news only related to the Japanese economy and business. Intuitively, breaking news such as the bankruptcy of Lehman Brothers might affect the market sentiment to make asset prices very volatile. As a result, economy-and-business news can be regarded a source that might have triggered the 2008 financial crisis.

As one approach to characterize such breaking news, we will count the *tf-idf* score for each word in the entire article that help us to extract the so-called "characteristic words." Let t be a day when the Nikkei publishes articles, $|T|$ the total number of days when the Nikkei is published, $|T_{w_i}: t \ni w_i|$ the number of days when a word w_i appears, and $N_{w_i,t}$ the number of appearances of w_i on day t. We define the tf and idf scores, respectively:

†1 Chuo University
†2 CyberAgent, Inc.

$$tf_{w_i,t} \coloneqq \frac{N_{w_i,t}}{\sum_k N_{w_k,t}} \quad (1)$$

$$idf_{w_i} \coloneqq \log \frac{|T|}{\left|T_{w_i}: t \ni w_i\right|} \quad (2)$$

By taking the product of these two scores, one is able to measure how likely the word w_i is to appear on day t and simultaneously how rarely the word w_i can be observed over the entire horizon.

$$(tf \cdot idf)_{w_i,t} \coloneqq tf_{w_i,t} \times idf_{w_i} \quad (3)$$

We call the above score "$tf \cdot idf$." The higher $tf \cdot idf$ the word w_i scores, the more likely the word w_i characterizes the Nikkei issued on day t. Table 1 shows the characteristic nouns appeared around the collapse of Lehman Brothers. On September 15, 2008, Lehman Brothers went into bankruptcy and Bank of America agreed to acquire Merrill Lynch. In the next day, the stock price of AIG fell sharply. Therefore after the 17th, damaged financial institutions such as Lehman Brothers, AIG, Bank of America and Merrill Lynch emerged as characteristic nouns.

Table 1: Characteristic words appeared around the bankruptcy of Lehman Brothers.

	2008/9/13	2008/9/14	2008/9/15	2008/9/17	2008/9/18
1	麻生氏	モネ	事故米	リーマン	aig
2	事故米	庵	有料道路	事故米	リーマン
3	アウトレット	麻生氏	捕虜	製菓	事故米
4	アーバン	ユダヤ人	分配金	aig	リーマン・ブラザーズ
5	献金	産科	リーマン	リーマン・ブラザーズ	落雷
6	与謝野	検針	共青団	メリル	つゆ
7	小池	ネスレ	内部統制	バンカメ	裁定
8	公開討論会	老い	艦	堂	サムライ債
9	石破	貧血	総合病院	ベアー・スターンズ	MBO
10	資金管理団体	リーマン	秋田県	特別養護老人ホーム	シダックス

Table 1 gives us an intuition that by focusing on the economy-and-business news that contains characteristic words that will appear around breaking news, we have a possibility to create better indexes to appropriately reflect market sentiment.

2.3 LDA clustering to specify the economy-and-business-related news

To specify the headlines and articles related to the Japanese economy and business news from the Nikkei, we employ *Latent Dirichlet Allocation (LDA)* to classify headlines and articles by topics. LDA developed by Blei et al. (2003) is one of the *topic models* which are algorithms for discovering latent topics from a collection of documents.

We conduct LDA using *gensim*[a] which is a software to realize unsupervised semantic modeling from plain text. It is noted that LDA algorithm employed by *gemsim* is an online variational Bayes developed by Hoffman et al. (2010). The reason why we employed this software and algorithm is that we should analyze massive document collections. Since we can interpret which topic each article has by LDA, it is helpful in constructing an SI focusing on a specific topic. Especially, an SI created from the headlines and articles related to the economy-and-business news seems to be more useful to predict the Japanese stock price than other SIs. In subsequent sections, we employ LDA to compute distributions over topics and construct an SI related to the economy-and-business news.

Before implementing LDA clustering, we remove the following articles: (1) only containing quarterly or annual closing figures, (2) only containing the company personnel information. Excluding the above articles, the number of articles used is 603,063.

We then set the number of topics as $K = 30$. Table 2 shows the five words with the highest probabilities of appearing within each topic.

Table 2: Top five words with highest probabilities of appearing in each topic.

Topic	Words				
1	前年	比	増	減	回復
2	億	期	純利益	増	売上高
3	工場	生産	中国	インド	事業
4	店	店舗	商品	出店	販売
5	氏	大統領	選挙	政権	投票
6	北朝鮮	中国	米	ロシア	会談
7	研究	細胞	薬	治療	患者
8	感染	台湾	卸売市場	元代表	ウイルス
9	容疑者	逮捕	容疑	人	捜査
10	首相	民主党	自民党	氏	党
11	株	東証	貸し借り	銘柄	指定
12	原発	東電	福島	稼働	電力
13	配	予	発売	newface	販売
14	スマホ	通信	装置	半導体	パネル
15	さん	私	人	著	作品
16	位	ゴルフ	アンダ	番	打
17	空売り	人事	課長	比率	維新
18	価格	トン	値上げ	原料	輸入
19	数表	先物	原油	相場	残高
20	社長	氏	取締役	就任	入社
21	日銀	金利	国債	応札	物
22	試合	監督	回	戦	チーム
23	発行	格付け	銀	運用	融資
24	サービス	ネット	サイト	情報	企業
25	企業	経済	政府	制度	改革
26	ドル	安	高	銭	上昇
27	空港	日航	便	ホテル	航空
28	死去	歳	さん	人	氏
29	週	馬	特許	gi	公社債
30	位	女子	男子	五輪	大会

Topic 1 might refer to changes in production and export volume and Topic 21 to monetary policy. Topics 16, 22, and 30 might relate to sports. A reasonable consideration leads us to interpret that a cluster of topics, $econ \coloneqq \{1,2,3,11,18,19,21,23,25,26\}$ colored in blue, can be related to the news on the economy and businesses.

2.4 Create three types of sentiment indexes

We construct three types of sentiment indexes (SIs) in this section. The first is $SI^{(econ)}$ created from the headlines and articles that are selected from the econ cluster only. That is, if any document in a headline or an article has a positive probability of being any topic $k \in econ$ that is characterized by a vector of words, the headline or the article is categorized to be economy

a Gensim is an open source application. For details, refer to https://radimrehurek.com/gensim/index.html (accessed March 18, 2017).

and business related. The second is $SI^{(all)}$ created from all the headlines and articles for comparison. The third is $SI^{(non\ econ)}$ created from the headlines and articles that are not categorized to be economy and business related. That is, if all the documents in a headline or an article have zero probabilities of being any topic $k \in econ$, the headline or the article is categorized to be non-economy and non-business related news.

We introduce some notations to clarify our SIs on the basis of such three categories. If a pair of headline and article belongs to one of three categories, denoted by $\mathcal{O} \in \{econ, all, non\ econ\}$, then every word that appears either in the headline or in the article is also categorized as $\mathcal{O}$. In the newspaper issued at day t, we have $n_t^{\mathcal{O}}$ headlines and articles that are categorized as $\mathcal{O}$. $W_{ij,t}^{\mathcal{O}}$ denotes the j-th word ($j = 1, \cdots, n_{i,t}^{\mathcal{O}}$) that comprises the i-th article and headline ($i = 1, \cdots, n_t^{\mathcal{O}}$) that belong to the category $\mathcal{O}$.

In our analysis, we use the *Tango Kanjo Kyokusei Taio Hyo* (Semantic Orientations Dictionary) of Takamura et al. (2005) to create three SIs. The dictionary is denoted by $\mathcal{D} := \{(D_l, S(D_l)) | l = 1, \cdots, L\}$. The dictionary comprises pairs of a semantic expression D_l and its positive or negative polarity score $S(D_l)$, which ranges from -1 to $+1$.

We define an indicator function to identify whether a word matches with the dictionary:

$$I_{ij,t}^{\mathcal{O}}(l) := \begin{cases} 1 & (if\ W_{ij,t}^{\mathcal{O}}\ matches\ D_l) \\ 0 & (otherwise) \end{cases} \quad (6)$$

Using this indicator function, we define the number of positive and negative words, respectively.

$$P_t^{(\mathcal{O})} := \sum_{i=1}^{n_t^{\mathcal{O}}} \sum_{j=1}^{n_{i,t}^{\mathcal{O}}} \sum_{\{l:S_l>0\}}^{L} I_{ij,t}^{\mathcal{O}}(l), \quad (7)$$

$$N_t^{(\mathcal{O})} := \sum_{i=1}^{n_t^{\mathcal{O}}} \sum_{j=1}^{n_{i,t}^{\mathcal{O}}} \sum_{\{l:S_l<0\}}^{L} I_{ij,t}^{\mathcal{O}}(l) \quad (8)$$

We finally define the SI at time t as follows:

$$SI_t^{(\mathcal{O})} := \frac{-N_t^{(\mathcal{O})}}{P_t^{(\mathcal{O})} + N_t^{(\mathcal{O})} + 1} \quad (9)$$

Remark that we add 1 to the denominator in case the SI becomes undefined when neither positive nor negative words appear.

3. Empirical Analysis

We examine the relationship between news and stock prices along the seminal discussions by Tetlock (2007) and Okimoto and Hirasawa (2014). They considered a trivariate VAR (vector autoregression) model comprising three variables: SIs, log returns of the Nikkei 225, and log trading volumes for all the constituents of the Nikkei 225. This section elaborates on three impacts: the impact of SIs on log returns and vice versa, and also the impact of SIs on log trading volumes[b]. We remark that these variables are normalized so that they have zero means and unit standard deviations.

3.1 Impact of SIs on log returns

First, we regress log returns with a VAR model to explore how SI causes log returns by controlling log trading volumes.

$$NK225_t = \alpha_1 + \sum_{i=1}^{p} \{\beta_{1i} NK225_{t-i} + \gamma_{1i} SI_{t-i}^{(\mathcal{O})} + \delta_{1i} Vol_{t-i}\} + \varepsilon_{1t} \quad (10)$$

where $\mathcal{O} = \{econ, all, non\ econ\}$ and p represents the lag order. Changing the lags from one to five, we choose $p = 5$ that displays the lowest AIC. Table 3 summarizes the estimation results.

Table 3: Estimated log returns by Eq. (10). We report the estimated coefficients of SIs in three cases: Econ SI $(SI^{(econ)})$, All SI $(SI^{(all)})$ and Non Econ SI $(SI^{(non\ econ)})$. Also Newey-West standard errors robust to heteroscedasticity and autocorrelation are also shown in parentheses. *, **, and *** indicate 10%, 5%, and 1% significance, respectively (We use the same notation in the following Table 4).

Ind. Var.	Econ SI	All SI	Non Econ SI
SI_{t-1}	0.05391*	0.05238*	0.00488
	(0.03062)	(0.03071)	(0.02472)
SI_{t-2}	0.01739	0.00306	-0.02081
	(0.02928)	(0.02988)	(0.02581)
SI_{t-3}	0.04910*	0.05415*	0.01489
	(0.02964)	(0.02983)	(0.02495)
SI_{t-4}	0.04347	-0.04824*	-0.02746
	(0.02893)	(0.02925)	(0.02796)
SI_{t-5}	0.01147	-0.00987	-0.00293
	(0.02815)	(0.02819)	(0.02947)
Adj. R	0.00531	0.00490	0.00107

We focus on the coefficients of $SI_{t-i}^{(\mathcal{O})}$. When estimated with $SI^{(econ)}$ created from headlines and articles that can be related to the economy and business news, the coefficients of SI_{t-1} and SI_{t-3} are significantly positive at 10%, whereas other coefficients are not significant. Although this result is consistent with Okimoto and Hirasawa (2014), it differs from Tetlock (2007). As Okimoto and Hirasawa (2014) cited *Information Theory*, the Nikkei news related to the economy and business might possess intrinsic information on Japanese stock prices.

When estimated with $SI^{(all)}$ constructed from all the headlines and articles, the coefficients of SI_{t-1} and SI_{t-3}, are significantly positive, while the coefficient of SI_{t-4} is significantly negative. This is consistent with the result of Tetlock (2007), who pointed out a *return reversal*, i.e. a change in the direction of a price trend reflected in these two opposite consecutive coefficients.

With $SI^{(non\ econ)}$ constructed from headlines and articles not related to the economy and business, no coefficients are

b Though we explored the impact of SIs on log trading volumes in detail, we omit elaborating on the result due to limitations of space.

significant. This result indicates that $SI^{(non\ econ)}$ has no information about the stock price.

In summary, we observe no return reversal in case of $SI^{(econ)}$, a return reversal in case of $SI^{(all)}$, and no significant estimate in case of $SI^{(non\ econ)}$.

Further, as for the goodness of fit in terms of adjusted R-squares, the model with $SI^{(econ)}$ slightly surpasses other two models with $SI^{(all)}$ and $SI^{(non\ econ)}$.

3.2 Impact of log returns on SIs

Second, we focus on a VAR model with SI as a dependent variable to examine the influence of log returns on SI:

$$SI_t^{(\mathcal{O})} = \alpha_2 + \sum_{i=1}^{p} \{\beta_{2i} NK225_{t-i} + \gamma_{2i} SI_{t-i}^{(\mathcal{O})} + \delta_{2i} Vol_{t-i}\} + \varepsilon_{2t}, \quad (11)$$

where $\mathcal{O} = \{econ, all, non\ econ\}$ and p represents the lag order. Changing the lags from one to five, we choose $p = 5$ that displays the lowest AIC. Table 4 summarizes the estimation results.

Table 4: Estimated SIs by Eq. (11) in three case.

Ind. Var.	Econ SI	All SI	Non Econ SI
$NK225_{t-1}$	0.02644	0.02419	-0.01600
	(0.01934)	(0.01912)	(0.02438)
$NK225_{t-2}$	0.04194**	0.04392**	0.03635
	(0.01973)	(0.01982)	(0.02577)
$NK225_{t-3}$	0.03721**	0.04395***	0.02305
	(0.01696)	(0.01644)	(0.02604)
$NK225_{t-4}$	0.00796	0.01618	0.02786
	(0.01887)	(0.01891)	(0.02692)
$NK225_{t-5}$	0.01061	0.00451	-0.03154
	(0.01840)	(0.01785)	(0.02427)
Adj. R	0.36680	0.36570	0.06126

If SI describes some aspect of stock prices, the coefficients of $NK225_{t-1}$ should be significantly positive or negative. Accordingly, we interpret the estimation results as shown in Table 4. The results with $SI^{(econ)}$ or $SI^{(all)}$, the coefficients of $NK225_{t-2}$ and $NK225_{t-3}$ are significantly positive. This shows that stock prices will affect the sentiment reflected in two SIs. On the other hand, no coefficients of $SI^{(non\ econ)}$ are significant. In addition, regarding the goodness of fit in terms of adjusted R-squares, the model with $SI^{(econ)}$ is superior to the other two models with $SI^{(all)}$ and $SI^{(non\ econ)}$.

4. Conclusion

We show that an SI constructed only from the economy-and-business-related news better predicts log returns of the Nikkei 225 than one constructed from the general or non-economy-and-business news. More precisely, an SI created from the economy-and-business-related news significantly explained log returns and trading volume during the next business day. We did not observe the reversal in returns pointed out by Tetlock (2007) when we use an SI created from the economy-and-business-related news. Moreover, we found that our SIs created from the economy-and-business or general news followed the movement of log returns.

Reference

[1] Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003), "Latent Dirichlet allocation," Journal of Machine Learning Research, 3, 993–1022.
[2] Hoffman, M., Bach, F. R., and Blei, D. M. (2010), "Online learning for latent Dirichlet allocation," Advances in Neural Information Processing Systems 23, 856–864.
[3] Ishijima, H., Kazumi, T. and Maeda, A. (2014), "Quantifying Sentiment for the Japanese Economy and Stock Price Prediction," Proceedings of PDPTA 2014, 255–259.
[4] Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004), "Applying Conditional Random Fields to Japanese Morphologiaical Analysis," IPSJ SIG Notes, 2004, 89–96.
[5] Mao, H., Counts, S., and Bollen, J. (2011). "Predicting financial markets: Comparing survey, news, twitter and search engine data," https://arxiv.org arXiv:1112.1051v1.
[6] Okimoto, T. and Hirasawa, H. (2014), "Nyusu Shihyo niyoru Kabushiki Shijo no Yosoku Kanosei (Stock Market Predictability Using News Indexes)," Security Analysts Journal, 52(4), 67–75.
[7] Takamura, H., Inui, T, and Okumura, M. (2005), "Extracting semantic orientations of words using spin model," Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 133–140.
[8] Tetlock, P.C. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," Journal of Finance, 62(3), 1139–1168.

Acknowledgments We are grateful to Akira Maeda (University of Tokyo), Masamitsu Ohnishi (Osaka University), Wataru Ohta (Osaka University), Naohiro Matsumura (Osaka University), and Katsuhiko Okada (Kwansei Gakuin University) for the fruitful discussions with them. We would also like to thank anonymous referees and organizers of MPS 114, a satellite workshop of PDPTA 2017, for their useful comments. This work was supported by JSPS KAKENHI Grant Numbers JP15K12465 and JP16H03127.