情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-NL-199 No.10
2010/11/19

ロコミ分析による日経２２５先物の相場変動予測

セーヨー サンティ[†]　榊剛史[†]　内山幸樹[†]

ソーシャルメディアを介した情報伝搬が活発になる中で、多くの人々の個人単位での行動や心理状態が把握しやすくなりつつある。本研究は大量のロコミ分析によって得られた市場心理と日経２２５先物の推移を機械学習させ、売買予測モデルを確立した。取引シミュレーションの検証結果は日経２２５先物の動向予測において、ロコミ分析が有効であることを示している。

Prediction of Nikkei 225 Index Futures based on Word of Mouth Analysis

Santi Saeyor[†]　Takeshi Sakaki[†] and Koki Uchiyama[†]

It is rather simple to investigate personal activities and mental states of people as the use of social media tools grows rapidly. This paper explains the research on prediction of Nikkei225 Index Futures based on word of mouth analysis and machine learning. A prediction model was proposed to generate trade signal. The model was tested against real price movement and the result has shown that the proposed algorithm constitutes an effectively profitable prediction model.

1. Introduction

The wide spread use of social communication tools like blog, Twitter, Facebook, etc. has exposed a huge communication data to the public. With the application of various natural language processing and data mining, we are more capable of investigating public opinions and personal activities. In this research, we try to analyze these kinds of word of mouth and apply to a real world application like predicting the stock market. Since the analysis of word of mouth represents the whole image of the crowd, we target on an investment instrument called Nikkei 225 Index Futures which is an index derived from stock price of major enterprises rather than some particular enterprises' stock.

So far, investors speculate the stock market with technical, fundamental and sentimental analysis. The technical analysis and fundamental analysis deal mostly with numbers such as prices and economic figures while sentimental analysis refers to a broad area of natural language processing, computational linguistics and text mining. However, investment in the stock market is a zero sum game. It is not always true that if the crowd expects the market to go upward then the market goes upward. There are always deceptive moves and wrong expectations. Instead of those conventional analyses, we tried to learn to map the crowd's attitude to the daily price movement with machine learning.

Some early studies show that stock market prices do not follow a random walk and can indeed to some degree be predicted [3],[4],[5],[6]. In this paper, we studied how to extract some useful features from the word of mouth. The main source of the data is Japanese blogs. We applied simple word extraction techniques and proposed a feature selection that related to price fluctuation of the index futures price. Then, we proposed a training algorithm based on Naive Bayes classifier. The derived model has been tested against the real price data for about one year after the training period. The results have shown that the prediction of price movement based on the word of mouth analysis is profitable and quite promising for investment decision in the world where we can access public communication like today.

We describe the trading system and some basis on the investment instrument in section 2. Section 3 explains how to extract features from the word of mouth. We proposed a training algorithm in section 4. The proposed model was tested and evaluated in section 5. We discussed and then conclude in Section 6.

[†] Hottolink Inc.
4F Hitotsubashi SI Bldg. 3-26 Kandanishikicho Chiyoda-ku Tokyo 101-0054　http://www.hottolink.co.jp/

2. Nikkei 225 Index Futures Trading System

The main purpose of our study is to predict the price movement of the Nikkei 225 Index Futures so to make profit from the capital gain. The Nikkei 225 Index Futures was selected because we can virtually think of it as a result of overall investors' opinion over Japanese economy, and besides, it is popular and heavily traded intraday. Our system aims to analyze the huge data of word of mouth daily and make sure to get the daily trade decision before the stock market opens.

In this section, we introduce the basis of Nikkei 225 Index Futures and how to trade this investment instrument. Then we incorporate those bases to build our proposed trading system.

2.1 Nikkei 225 Index Futures

Nikkei 225 Index Futures is a derivative of Nikkei 225 stock index. The base index is a price-weighted average which is designed to reflect the overall stock market and its components are reviewed once a year. It is traded in unit of lot. The index moves in unit of 10 yen for Nikkei 225 Large and 5 yen for Nikkei 225 mini.The Nikkei 225 Large has a leverage of 1,000 times while the mini one has a leverage of 100 times. Unlike trading a real stock, the derivative trading allows an investor to enter the market either by a long (buy) or a short (sell) position. A long position is profitable if the price goes up. In contrast, a short position is profitable when the price goes down. Trading the index futures has chances to make profit both when the stock market is going up and down. The investor can take profit or admit loss by issuing a counter trade: selling for a long position or buying for a short position.

Due to the high leverage by its nature, an investor has chances of making big profit and suffering from margin call when the stock price moves rapidly. The margin call occurs when the base index price moves in the opposite direction of an open position and the free margin is running short. In this case, an investor is urged to deposit more money or otherwise the broker will force a counter trade and the investor would loss most or all of the money deposited.

Basically, the Nikkei 225 Index Futures is used as a hedge instrument to the ordinary stock trading. However, trading the index itself is also very attractive. The professional index futures traders are called CTA (Commodity Trading Advisor). In their career, it is their job to implement any kinds of strategies to make their clients profit from trading the index futures including making deceptive moves from time to time.

The Nikkei 225 Index Futures has expired date on the second Friday, every 3 months: March, June, September, and December. On the expired date any open position will be closed. Once opening a position, the investor has right to close the position at any time, or leave the position opened until the expired date and let it close automatically. However, most of the long term investors do rollover their positions to the next term of the index future before they expire.

2.2 Trading the Index Futures

As we described the nature of the index futures and some uncertainties upon trading it in previous topic, our task is to predict the direction of the price movement as precisely as possible. Before all, we have to choose trading time. The index futures are traded around the world as shown Table 1:

Table 1 Markets that trade Nikkei 225 Index Futures and trading sessions.

Market	Trading sessions (except holidays)
Osaka Securities Exchange (OSE)	09:00 – 11:00 12:30 – 15:10 16:30 – 23:30
Singapore Exchange	08:45 – 15:30 16:30 – 20:00
Chicago Mercantile Exchange (CME)	20:00 – 06:15(next day) (until 05:15 during summer time)

The index futures are traded almost 24-hour a day on three major markets. There many sessions that trade at the same time on different markets. Most of the time, the prices are almost the same and moving in the same direction. In our work, we locally trade the index futures at Osaka Securities Exchange (OSE). However, keeping an open position while the market is closed, we are exposed to a great risk because if any unpredictable disasters suddenly occurred, the index futures price might move rapidly on other markets while we cannot make any move on OSE.

In order to eliminate any risks while the market is closed, we prefer day-trading to keeping open a position overnight. That means we buy or sell the index futures at 09:00 then close the position on 15:10 no matter it is making profit or loss. In Fig. 1, we plotted a histogram of open and close price different for almost 4 years of data during 2006-12-18 to 2010-06-30. The distribution looks like a normal distribution. The cumulative distribution function in Fig. 2 is almost symmetry around 0 which means we have almost the same chance to make a profit (or loss) both on the up and down days.

When randomly buy or sell the index futures in a long term, the base line of win ratio should be around 50 percents. This is also the same when only buy or sell the index futures every day. The prediction problem seems to be easy but it is not that easy in practice. The point is that if we take a look at Fig. 3, we will find that the open/close price width has its own specific distribution. Moreover, in Fig. 4, we notice that almost 80% of the time, the open and close price width is not greater than 30 yen. This fact implies that even we have a prediction model that yields more than 50% in accuracy; it is not guaranteed that we can make some profits. This is true when the model always wins small price width days, while losses on several big price width days. In contrast, a model of less than 50% in accuracy has chances to make some profits if it always wins on big price width days and losses on small price width days.

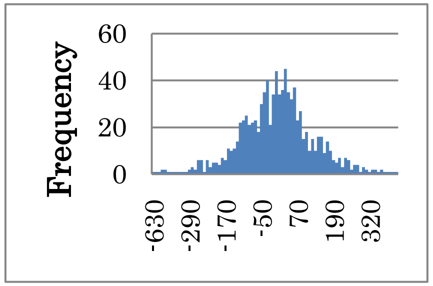

Fig. 1 Histogram of Open/Close price different during 2006-12-18 to 2010-10-19.

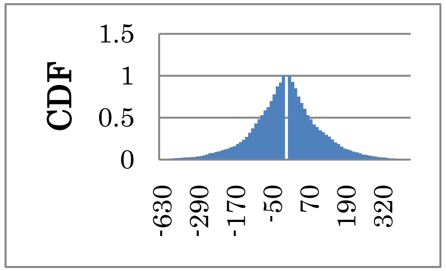

Fig. 2 Cumulative Distribution Function of price different on both minus and plus sides during 2006-12-18 to 2010-10-19.

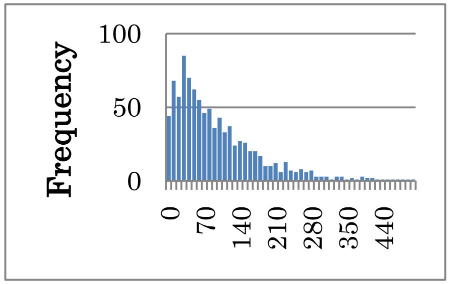

Fig. 3 Open/Close Price Width Histogram during 2006-12-18 to 2010-10-19

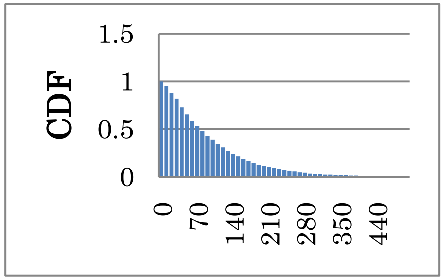

Fig. 4 Cumulative Distribution Function (CDF) of price width during 2006-12-18 to 2010-10-19

From the fact above, our challenges are not only to find a high accuracy model but to find a model that makes right trade decision on the days of big price width. And the model should win more than losses including the transaction fees in order to make some profits. We will explain our trading system in the following topic.

2.3 System Architecture

The proposed trading system consists of word of mouth processing and machine learning as shown in Fig. 5. The algorithm of feature word extraction will be explained later in section 3. The system extracts a set of feature words from a pre-selected set of blogs. The set of feature words serves as a set of features for machine learning. A training set is then fed, in accordance with Nikkei 225 index futures price database, to the Machine Learning module. The Machine learning here is Naive Bayes classifier.

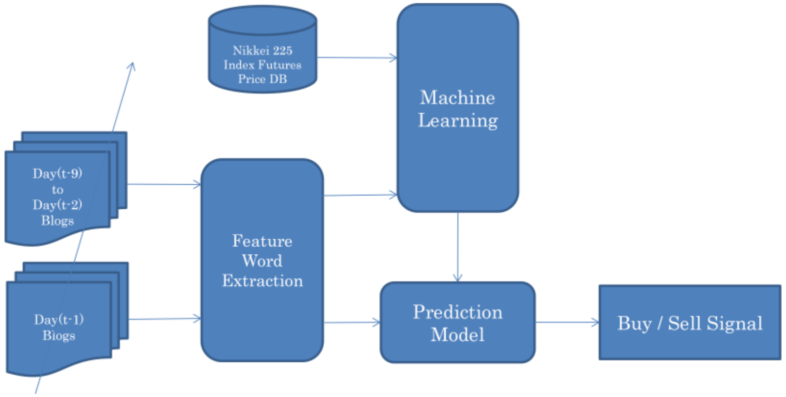

Fig. 5 Architecture of the Nikkei 255 index futures day-trading system based on word of mouth analysis.

The detail of training algorithm will be explained deliberately in the section 4. Once the training is done, we derived a prediction model for the current day. The next process is to predict whether the index is going up or down by the end of afternoon session on the current day. The signal of Buy or Sell is then applied for the investment decision and the order will be done before 09:00 and the position will be closed on 15:10. The feature word extraction module is used in common for both the training and the prediction to make sure that the extracted feature words are shared and the extraction algorithm is exactly the same.

The system can adapt itself to the recent scenario of the market by using only a recent set of feature words and investment blogs. In this fashion, the feature words and the prediction model are updated every day to track any scenario shift in the stock market. The philosophy behind this system is to overcome the uncertainty of stock price fluctuation with the uncertainty of word of mouth written in the set of investment blogs.

3. Feature Extraction from Word of Mouth

The word of mouth is widely exposed on the internet as the social media services and tools are growing rapidly. It is almost impossible for a human to read, scrutinize, and interpret all the personal messages every day before the market open. We propose here an algorithm for extracting useful features from word of mouth on the internet and explain how to implement them with machine learning that predict daily price movement of the Nikkei 225 index futures. All the processes dealt with Japanese text processing which is more complex than English because Japanese words are not separated by spaces like English. In this section we describe the main source of word of mouth, desirable features and the feature extraction criteria.

3.1 Data Source

The main data source in the research is word of mouth selected from blogs. We rely on social filtering when selecting a set of investment blogs for our prediction model. We believe that blogs which were ranked by people are likely to have been reviewed and meaningful. In this research we selected about 3,000 blogs under the category of investment from the “BLOG RANKING” [1]. We gather the latest updates and have it ready for analysis by 08:30 every day. Top-n blogs seems to have high quality of writing as well as the comments provided. The blogs are written by experienced brokers, day-traders and some are written by housewives. Rather than focusing only on expert comments, in this research, we tried to catch the overall picture from the set of investment blogs. By this reason, we also treat all the blogs by the same weight regardless of their ranks.

3.2 Feature Words

In order to train the machine learning we need to select a set of attributes that are predictive to the Nikkei 225 index futures. First of all, we gather words from the set of investment blogs using MeCab, a part-of-speech and morphological analyzer for Japanese text. MeCab outputs words and part-of-speech classified by its engine. In order to take balance of words from various parts of speech, we select only 80 words from each kind of part of speech: Noun, Adjective, Verb, and others as shown in Table 2. When picking up the feature words, we only take into account the blog written within the Last N days of investment blogs. Suppose the day *t* that we want to predict the price movement is denoted by D_t, We prepare data for feature extraction by applying all the text on D_{t-N} to D_{t-2} to MeCab, gathering the words list with their part-of-speech labeled, and count up each word according to the price movement direction (up or down) on the following day. In the process we also take into account the negation of words stated in each sentence. It is obvious that blog writers express their feelings and attitudes using natural language and there are both affirmative and denial states upon the subjects. For example, “I think the GDP is growing.” and “I don’t think the GDP is growing” are completely opposite to each other. So we carefully pick up all the words in both affirmative and denial manners. In this case, the word GDP should be treated as “GDP” and “GDPneg” separately

Table 2 Parts of speech and corresponding number taken for training set.

Parts of Speech	Number
Noun	80
Adjective	80
Verb	80
Other	80

Thank to the capability of MeCab which is able to classify auxiliary verbs that indicate negation in various forms, we are able to point out quickly the negative points in each sentence by detecting “-NAI”, “-MASEN”, “-ZU”, “-NU”, etc., and their corresponding tenses. According to the fact that there can be multiple negations in one sentence, we also consider multiple negations because a negation of negation gives an affirmative sense.

Table 3 the preprocessing algorithm for feature words extraction.

```
$W_n$\{up\} = 0
$W_n$\{down\} = 0
For i = 2 to N
    For each blog message on day $D_{t-i}$
        Wordset = Analyze the messages with MeCab and gather words list marked by their part-of-speech and mark the word as negated if found negation in the sentences.
        For each $W_n$ in Wordset
            IF the price movement on $D_{t-i+1}$ is up then $W_n$\{up\}++
            Else IF the price movement on $D_{t-i+1}$ is down then $W_n$\{down\}++
        End
    End
End
```

However, there are some grammar patterns in Japanese that some particles always force the sentences to end with negative form but doesn't make denial sense, for instance, “SHIKA…NAI”. We neglect the later form in our preprocessing as our main purpose is to find word pattern rather than analyze positive or negative opinions. The summary of data preprocessing algorithm for feature words extraction is shown in Table 3.

3.3 Extraction Criteria

Once we derived a word set labeled with price movement counts on the next day, both in up and down directions, we select only top 80 words for each part-of-speech shown in Table 2 that make the most sense of prediction to the Nikkei 225 index futures' movement. In order to so, we introduced a scoring system for calculating predictive degree on each word. The score of W_n is given by the equation (1), where DF{W_n} is the document frequency of word W_n (the number of blog messages that contain word W_n) and W_n {up}, W_n {down} are the data derived in the preprocessing step.

$$W_n\{score\} = \frac{|W_n\{up\} - W_n\{down\}| \log(1 + DF\{W_n\})}{W_n\{up\} + W_n\{down\}} \qquad (1)$$

The reason behind the scoring system is to give a higher score to the word that has a higher rate of up/down direction different based on its term frequency (TF) and document frequency (DF). The extraction algorithm is shown in Table 4.

Table 4 the feature words extraction algorithm.

```
For each $W_n$
    Calculate $W_n$ {score} as defined in the equation (1)
End
Sort $W_n$ {score}
Pick up Top 80 $W_n$ {score} of type Noun
Pick up Top 80 $W_n$ {score} of type Adjective
Pick up Top 80 $W_n$ {score} of type Verb
Pick up Top 80 $W_n$ {score} of type Other
```

In this manner, we derive a set of 320 influential words to be used as attributes in our training set every day.

4. Training

Since we have no explicit mathematical model for predicting the direction of Nikkei 225 index futures price movement, we rely on machine learning based on Naive Bayes classifier. This section explains how to train the classifier so that the prediction system can be driven by input data and be able to adapt to the changes in the stock market environment.

4.1 Training System

The training system of the Nikkei 225 index futures prediction model is shown in Fig. 6. We analyze a set of blog messages on the last N days and pick up a set of influential words as explained in previous topics. Then build a training set for our machine learning system. The training set consists of count of feature words on each training day and corresponding UP/DOWN label that indicate the direction of the index price movement on the next day.

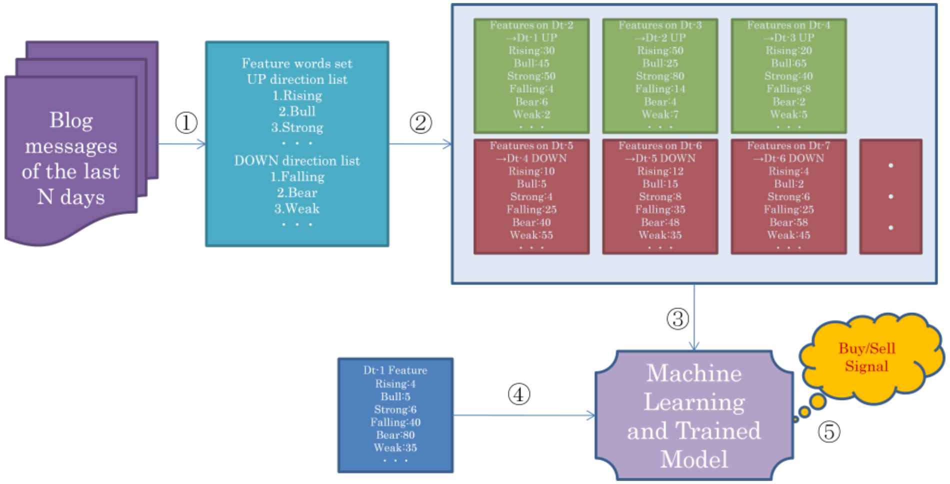
Blog messages of the last N days
①
Feature words set
UP direction list
1.Rising
2.Bull
3.Strong
・・・
DOWN direction list
1.Falling
2.Bear
3.Weak
・・・
②
Features on Dt-2 →Dt-1 UP Rising:30 Bull:45 Strong:50 Falling:4 Bear:6 Weak:2 ・・・
Features on Dt-3 →Dt-2 UP Rising:50 Bull:25 Strong:80 Falling:14 Bear:4 Weak:7 ・・・
Features on Dt-4 →Dt-3 UP Rising:20 Bull:65 Strong:40 Falling:8 Bear:2 Weak:5 ・・・
Features on Dt-5 →Dt-4 DOWN Rising:10 Bull:5 Strong:4 Falling:25 Bear:40 Weak:55 ・・・
Features on Dt-6 →Dt-5 DOWN Rising:12 Bull:15 Strong:8 Falling:35 Bear:48 Weak:35 ・・・
Features on Dt-7 →Dt-6 DOWN Rising:4 Bull:2 Strong:6 Falling:25 Bear:58 Weak:45 ・・・
・
・
・
③
Dt-1 Feature Rising:4 Bull:5 Strong:6 Falling:40 Bear:80 Weak:35 ・・・
④
Machine Learning and Trained Model
Buy/Sell Signal
⑤

Fig. 6 the training system of the Nikkei 225 index futures prediction model.

The resulted model is then use to predict the direction of the index price movement on current day. We train the model with only the last N days in order to keep tracking on recent events that have impact on the movement of the index price. The reason behind this training policy is based on the fact that investors' attitudes toward the stock market always change.

Normally, the index price moves in upward direction when a lot of companies announce good performance or increasing revenues. However, that is not always true, for example when the government declared monetary relaxation policies. In such a case, worse companies' performance or decreasing revenues trend to drive the index price up because of the expectation on monetary relaxation. There are many causes and situations like this so we hope the model adapt itself to the timely situations.

4.2 Training and Predicting Processes

The training phrase and prediction phrase are done in sequence in order to get the prediction result on each day. The detailed of training and prediction processes are described in Table 5.

Table 5 the training and predicting processes.

①Analyze blog messages on the day D_{t-N} to D_{t-2} and pick up a set of influential feature words.
②Prepare training set that consists of term frequency of each feature word and the direction of price movement on the following day.
③Feed the training set to the machine learner and train it to derive a prediction model.
④Analyze the blog messages on the day D_{t-1} and predict with the model derived in step ③.
⑤Interpret the Buy or Sell trading signal and take the action correspondingly.

The processes are straightforward and done quickly. We always start the training process around 08:30 and the prediction result is ready by 08:45 which long enough to take any action before the market opens on 09:00.

5. Experiment

In this paper, our goal is to study if the analysis of word or mouth could be useful in predicting the direction of Nikkei 225 index futures price movement. We decided to carry out the experiment on Nikkei 225 mini because of its more precise tick of 5 yen. First of all, we have done the experiment described in this section and evaluated the derived model in various points of view. The proposed prediction model can output 2 types of trading signal: Buying and Selling. We further study whether the proposed model could be improved when combined with an anomaly about stock trading. There is an anomaly says: buying a stock at higher price or selling at a lower price is likely to make some profits. This is quite obvious true when the stock price breaks its trading channel as the demands and supplies change rapidly. We applied a simple rule that refrain from entering the market on the day that the prediction model gives a selling signal while the open price seems to start higher than 20-day Exponential Moving Average and vice versa on the buying signal. So the combination of word of mouth based prediction model and the anomaly rule can output 3 types of signals: Buying, Selling, and Skip.

5.1 Determine an appropriate set of training points

We used the Nikkei 225 mini price records and the investment blogs data from 2006-12-01 to 2010-06-30 for the experiment. The statistics of the data is shown in Table 6.

Table 6 Statistics of Nikkei 225 index futures price movement during 2007-01-01 to 2010-06-30

Statistics	Learning Period (2007-01-01 to 2009-08-31)	Evaluation Period (2009-09-01 to 2010-06-30)
Up days	314	95
Down days	328	103
Unchanged days	11	3
Total tradable days	653	201
Possible max. profit (yen/lot)	7,258,500	1,238,000

We simulated trading 1 lot of the index futures every working day according to the prediction of the proposed model. First of all, in order to find an appropriate N day of training, we ran the proposed training model with the N varied from 3 to 20 on the data from 2007-01-01 to 2009-08-31. While running through the process for each N, we record the number that the prediction hits and the profit derived from trading with the model as two indicators for choosing an optimal N. In the process, we treated the day that the index futures opened and closed at the same price as a loss day regardless of the prediction. The reason is that, in the real trading, we have to pay transaction fees when opening and closing a position. The testing period of about 2.5 years with 653 day-trades is the longest period we can afford during the work on this model.

Around the end of August 2009, we have found that training with N=9 gives high profit while the accuracy is around the base line of 50% as shown in Table 7. Since then we keep using the training algorithm with N=9. Then we evaluated the proposed model on the period from 2009-09-01 to 2010-06-30. In this period, we may state that the result is purely the performance of the model itself.

5.2 Performance

The comparison of the proposed model, the model with skip, and various kinds of random trading models are shown in Table 7 and Table 8.

Table 7 the comparison of the proposed model and some random models in the learning period from 2007-01-01 to 2009-08-31 assuming no transaction fee.

Models	Hit	Miss	Total	Accuracy	Net Profit (yen)	Profit per Trade (yen)
Proposed Model	324	329	653	49.62%	294,500	451
Model with Skip	309	308	617	50.08%	417,500	677
Buy only	314	339	653	48.09%	-97,500	-149
Sell only	328	325	653	50.23%	97,500	149
Random1	323	330	653	49.46%	-187,500	-287
Random2	342	311	653	52.37%	283,500	434
Random3	318	335	653	48.70%	63,500	97
Random4	307	346	653	47.01%	28,500	44
Random5	327	326	653	50.08%	-330,500	-506

Table 8 the comparison of the proposed model and some random models in the evaluation period from 2009-09-01 to 2010-06-30 assuming no transaction fee.

Models	Hit	Miss	Total	Accuracy	Net Profit (yen)	Profit per Trade (yen)
Proposed Model	110	91	201	54.73%	136,000	677
Model with Skip	110	89	199	55.28%	140,500	706
Buy only	95	106	201	47.26%	-12,000	-60
Sell only	103	98	201	51.24%	12,000	60
Random1	98	103	201	48.76%	-126,000	-627
Random2	95	106	201	47.26%	49,000	244
Random3	106	95	201	52.74%	37,000	184
Random4	108	93	201	53.73%	13,000	65
Random5	94	107	201	46.77%	-127,000	-632

According to the results, we have learned that the Nikkei 225 index futures price seems to be a random walk pattern and most of the prediction and random models have accuracy of about 50%. In short term, selecting “Buy only” strategy or “Sell only” strategy every day may result in higher than 50% in accuracy due to the market trend. However in long term, “Buy only” or “Sell only” strategy has the accuracy of 50%± α and does not result in much gain or loss. The random models are also have the accuracy of 50%± α. We notice that even the accuracy of a model is higher than 50%, it has possibility to be a loss model.

However, our proposed model turns out to be more accurate in the evaluation period than the training period as the accuracy reached 54.73% and in case of using the model with anomaly, its accuracy reached 55.28%. The proposed model seems to be able to predict correctly on the days that it should be correct. That means to be correct on the days that the closing prices are far away from the opening prices.

During the training period, the skip strategy seems to reduce loss and unnecessary trades. Thus the model with skip strategy significantly increases the profit. Combining the proposed model with some technical analysis may result in a better model.

6. Conclusion

We proposed a model that predicts the Nikkei 225 index futures price movement based on word of mouth analysis. Though, the accuracy of the model is not much higher than trading in random, in long term, the accuracy of the model is on the plus side of 50%. The model seems to predict correctly on the days that have relatively big open-close price widths. As a result, the model can make profit both in training and evaluation periods. We believe that the randomness of the word of mouth is predominant over uniform random trades when apply appropriately to a prediction model. The experiment results further show that the model can be improved when using with a basic anomaly. This gave us a hint that, instead of building a prediction model based purely on word of mouth analysis, we may find a better model if taken into account the combination of technical and fundamental analysis as well.

7. References

[1] BLOG RANKING by @with http://blog.with2.net/rank1530-0.html
[2] MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.sourceforge.net/
[3] Qian, Bo, Rasheed, & Khaled. (2007) Applied Intelligence 26, 25-33.
[4] Gallagher, L. A & Taylor, M.P. (2002) Southern Economic Journal 69, 345-362.
[5] Kavussanos, M & Dockery, E. (2001) Applied Financial Economics 11, 573-79.
[6] Butler, K. C & Malaikah, S.J. (1992) Journal of Banking and Finance 16, 197-210.