Prediction of Nikkei VI increase for reducing investment risk using Yahoo! JAPAN stock BBS

Kentaro Ueda*
ueda.kentaro.ug2@is.naist.jp
Nara Institute of Science and Technology
Nara, Japan

Kodai Sasaki
Nara Institute of Science and Technology
Japan
sasaki.kodai.sb6@is.naist.jp

Hirohiko Suwa
Nara Institute of Science and Technology
Japan
h-suwa@is.naist.jp

Yuki Ogawa
Ritsumeikan University
Japan
y-ogawa@fc.ritsumei.ac.jp

Eiichi Umehara
Niigata University of International and Information Studies
Japan
umehara@nuis.ac.jp

Tatsuo Yamashita
Yahoo Japan Corporation
Japan
tayamash@yahoo-crop.jp

Kota Tsubouchi
Yahoo Japan Corporation
Japan
ktsubouc@yahoo-crop.jp

Keiichi Yasumoto
Nara Institute of Science and Technology
Japan
yasumoto@is.naist.jp

ABSTRACT

In stock investment, it is important to predict future market fluctuations in order to reduce risk. The Nikkei 225 Volatility Index (VI) is a measure of the expectations of the investors of the future of the Japanese market. A rise in this index indicates that investors are concerned about the future of the market, and predicting this rise may be used to reduce investment risk. Social media posts contain the opinions and feelings of the posters. In the present study, we proposed a means of predicting the increase in the Nikkei 225 VI by analyzing the social media of the largest stock trading website in Japan, "Yahoo! Japan Stock Message Board," and capturing changes in the topics of discussion. As a result of evaluation over a long validation period, we developed a prediction model with an F1-measure of 0.26.

KEYWORDS

Stock BBS, Volatility Index, Natural Language Processing, Machine Learning, Social Media Analysis

ACM Reference Format:
Kentaro Ueda, Kodai Sasaki, Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Tatsuo Yamashita, Kota Tsubouchi, and Keiichi Yasumoto. 2021. Prediction of Nikkei VI increase for reducing investment risk using Yahoo! JAPAN stock BBS. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '21 Companion), December 14–17, 2021, ESSENDON, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3498851.3498940

1 INTRODUCTION

In stock investment, predicting future market fluctuations is important for risk reduction. Market price fluctuations are a manifestation of the fears of the investors, which are expressed in a volatility index. Forecasting a volatility index, which represents the fear of the investors in the stock market, can reduce risk in stock investment.

There are statistical methods[12] and machine learning methods[16] to predict the VI. Suwa et al.[19] analyzed the postings on a stock bulletin board to predict the increase in the Nikkei 225 Volatility Index (VI), which is the volatility of the Nikkei 225 stock market over the next month. The VI is calculated based on the Nikkei Stock Average, which is listed on the Osaka Securities Exchange and published daily by Nikkei Inc. However, their study did not evaluate the long-term period. Furthermore, the method used to obtain the distributed representation of the documents is the latent Dirichlet allocation (LDA) method, which may not classify the documents accurately, and the machine learning methods are the logistic regression model and random forests, which may not be accurate enough.

Therefore, in the present study, we increase the validation period to verify the effectiveness of the model in the long term. In addition, to improve the accuracy, we will investigate a more effective method of acquiring distributed representation from stock bulletin board postings using Doc2Vec and BERT to acquire a distributed representation of documents. In addition to random forests and logistic regression as machine learning algorithms, we also use LightGBM to build a VI prediction model. In this way, we propose and compare new methods for predicting the rise of the VI.

*Both authors contributed equally to this research.

WI-IAT '21 Companion, December 14–17, 2021, ESSENDON, VIC, Australia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-9187-0/21/12…$15.00
https://doi.org/10.1145/3498851.3498940

As an experiment, we analyzed the social media "Yahoo! JAPAN Stock BBS" of Japan's largest stock trading website. To extract topics from these messages, we acquired distributed representations of the documents using a language model. We obtained daily topic vectors by summarizing the acquired variance representations by day. Furthermore, in order to predict the increase of the VI, we developed a VI prediction model using logistic regression, random forests and LightGBM with the acquired daily topic vector and financial time series data as features. As a result, the maximum F1-measure of the predictive model was 0.26. We found that downsampling is effective in training the forecasting model and that the forecasting accuracy is lower during periods of downward market trend than during periods of upward market trend.

2 RELATED RESEARCH

2.1 Forecasting financial instruments

In finance, forecasting various market indices is very useful because this leads to profit and risk aversion, and many previous studies have been conducted. For example, Chen et al.[4] predicted the Chinese stock market by modeling the return of Chinese stocks using LSTM. As a result, their model achieved an improvement in accuracy of from 14.3% to 27.2% compared to the random method. In addition, Long et al.[15] proposed a new end-to-end model called the multi-filter neural network (MFNN), which is specialized for the task of feature extraction and price movement prediction of financial time series. They proposed a system that combines CNN and LSTM and reported that the resulting accuracy, profitability, and stability were superior to those of conventional machine learning models and statistical methods. There are also studies that use adversarial learning to predict stock prices. By using adversarial learning, Feng et al.[8] devised a stock price prediction model that takes into account the probability of stock prices and reported a relative improvement of 3.11% on average w.r.t. accuracy over Xu et al[23]. Studies using only financial time series data can only achieve chance rate accuracy due to the efficient market hypothesis[7]. Therefore, many studies have been conducted to show the predictability of the market using not only financial time series data but also press releases, news, social media, and other information.

For example, Antweiler et al.[1] examined the relationship between the number of postings on a stock board and the stock market. They found a positive correlation between 1) the number of postings and trading volume and 2) the predictability for returns and volatility. Preis et al.[17] used Google trends to determine search volume and found that some events had a high search volume for related events before the news occurred. They concluded that search volume not only reflects the current state of the market, but can also predict future trends. Li et al.[13] proposed a method to predict the movement of stock prices during the night between the closing and opening prices of the previous day using nightly news. They proposed the LSTM-RGCN model make use of the relationship among stocks and showed that the introduction of graphs can predict the movements of stocks that are not directly related to the news and the movements of the entire market, which cannot be predicted by conventional methods. Chen et al.[3] extracted a representation of potential stock characteristics from a repository of investment behavior in mutual fund portfolio data. Experiments using real-world stock market data showed the effectiveness of using the extracted stock characteristics for stock forecasting. Many studies have been conducted for the purpose of predicting market indices using various approaches. However, few studies have been conducted to forecast the VI.

2.2 Natural language processing and market forecasting

In a case study of Twitter, Bollen et al.[2] classified tweets into two levels of emotion (positive-negative) and six levels of emotion and investigated the correlation between the sentiment of daily tweets and the Dow Jones Industrial Average. The results show that there is no correlation between positive and negative tweets. On the other hand, "Calm" sentiment is positively correlated with the Dow Jones Industrial Average from 2 to 6 days later, and "Happy" sentiment is positively correlated with the Dow Jones Industrial Average 6 days later. Furthermore, they were able to predict the change of the Dow Jones Industrial Average with 86.7% accuracy, and this report shows the potential of social media applications.

With the development of natural language processing techniques, there are examples of obtaining variance representations for words and sentences from language models and adapting them to forecast financial time series. For example, Feuerriegel et al.[9] extracted 40 topics using the LDA method to analyze the impact of topics found in corporate press releases on stock market returns in the German market. The results show that some topics have no impact on the excess return of stocks, whereas other topics have a significant impact. Yang et al.[24]focused on the Financial and Economic Attitudes Revealed by Search (FEARS) index, which is an index that reflects the attention and sentiment of general investors. They developed a stock return prediction model using BERT and a self-attention deep learning model to test the performance of the FEARS index. The results confirm the effectiveness of the FEARS index. Liu et al.[14] proposed the capsule network based on transformer encoder (CapTE) model, which combines the advantages of the transformer encoder and the capsule network, to predict the upward and downward movements of stock prices using only social media posts. Using the CapTE model, we obtained specific relationships between tweets that can improve the prediction accuracy of stock movements. By following the strategy proposed by Lavrenko for six randomly selected stocks, we obtained a maximum return of 17% over 20 trading days by simulating actual stock trading.

2.3 Research on investment risk reduction

Du et al.[6] obtained a vector representation of stocks, called stock embeddings, using a deep learning framework based on both news articles and price histories. They applied the obtained vector representation to a portfolio optimization problem and performed investment simulations. The results show that the proposed method yields 2.8 times higher capital gains than the baseline method using only stock price data. This suggests that their proposal can be applied to risk control and asset pricing in financial markets.

Suwa et al.[19] used LDA to obtain the topics of Yahoo! JAPAN stock BBS postings and used them to predict the increase of the VI. They divided the postings into morphemes by morphological analysis and obtained the probability of belonging to 100 different topics in each document by using LDA. The acquired vectors are summarized by day and used as input to the machine learning algorithm. Then, we set up the binary classification problems of VI increase above 2σ and others, and created models using logistic regression and random forests. As a result, the model using logistic regression has obtained an accuracy of 0.45 in both precision and recall when evaluated over 395 business days from November 17, 2014 to June 29, 2016.

Furthermore, Sasaki et al.[18] developed a trading simulation of Nikkei 225 options trading based on the price information of intraday data in order to evaluate the effectiveness of the method of Suwa et al.[19] The results suggested the possibility of the effectiveness of the method of Suwa et al.[19] However, since the Japanese market was in an uptrend during this evaluation period, it is necessary to extend the evaluation period. Suwa et al.[19] used LDA as a method to acquire distributed representations of documents, but the prediction accuracy of the model may be improved by using other natural language processing methods that have been developed in recent years. In addition, Suwa et al.[19] used random forests and logistic regression as machine learning algorithms, but there is a possibility that the prediction accuracy of the model can be improved by applying other developed algorithms.

3 PROPOSED METHOD

In this section, we introduce a proposed method for predicting investment risk using the topic of stock BBS.

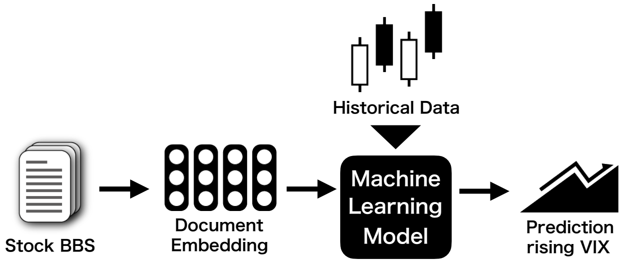

Figure 1: Document embedding, Historical data, Machine learning model, Prediction: rising VIX

3.1 Process of creating an investment risk prediction model

An overview of the investment risk prediction model is shown in the figure1.

In order to extract topics from postings on Yahoo! JAPAN stock BBS, we perform morphological analysis on the posted documents and split them into morphemes. By using a language model, we obtain distributed expressions from the documents divided into morphemes. Each dimension of the acquired distributed representation is defined as a topic. Next, the acquired distributed representations are grouped by day to form a topic vector for the date and time. Financial time series data is added to generate features, which are then used as input to the machine learning model. Finally, a machine learning algorithm is used to predict the rise in the VI, which is an investment risk.

3.2 Formulation of investment risk

We define and formulate the investment risk r as a large increase in the VI. First, we define a large increase in the VI, where v_t is the VI at day t. When T is the number of business days in the experiment, the daily difference x_t in the VI can be expressed as:

$$x_t = v_{t+1} - v_t(t \in 1, 2, 3, \cdots, T)$$

At this time, the standard deviation of the difference between the VI yesterday and that today is expressed by the following formula:

$$\sigma = \sqrt{\frac{1}{T-1}\sum_{k=0}^{T-1}(x_k - \bar{x})^2}$$

The threshold α for determining a significant increase in the VI is defined using the standard deviation σ as follows:

$$\alpha = i\sigma$$

As a condition, we restrict i to be a positive real number. Furthermore, since we consider only ascent in the present study, α is assumed to be a positive number. Next, we define the number of grace days before the threshold is exceeded. The reason for defining the number of grace days is to take into account both situations in which the VI increases rapidly and situations in which the VI increases moderately and significantly. Given a constant d, let $[t+1, t+d]$ be the time window from the day after day t to day d. The maximum value of the VI in the interval of the time window is expressed as follows:

$$v_t^d = \max\left\{v_{[t+1, t+d]}\right\}$$

As a condition, we restrict d to be an integer greater than or equal to 1. For example, when $d = 1$, the time window is one business day. In this case, the maximum difference m_t of the VI in the time window d is given by the following equation:

$$m_t = {v_t}^d - v_t$$

The investment risk r is given by the following equation, given threshold α and the maximum difference m_t of the VI for time window d:

$$r = \begin{cases} 1 & (m_t \geq \alpha) \\ 0 & (m_t < \alpha) \end{cases}$$

In the present study, the time window is set to $d = 5$, since a week is five business days. In addition, the threshold for determining a significant increase in the VI index is set to $i = 2$ to be the same as in previous studies.

3.3 Collection of submitted documents and Nikkei VI

Retrieve the posted documents from Yahoo! JAPAN stock BBS. The BBS used for the analysis was the Nikkei Stock Average. This is because the VI corresponds to the Nikkei 225 options listed on the Osaka Securities Exchange. The data extracted from the posted documents are date, time, and content. The data of the Nikkei 225 and VI indices are collected from JPX Data Cloud[1].

[1] http://db- ec.jpx. co.jp/item/C430509.html

3.4 Topic extraction using language models

Using a language model to extract topics from stock BBS postings, we acquire a distributed representation of the documents. In the previous study, the LDA topic model was used to derive the probability of belonging to 100 different topics for posted sentences. LDA only refers to the frequency of occurrence of words, so it does not take into account the similarity and word order among documents. Therefore, in the present study, in addition to LDA used in previous studies, we use Doc2Vec, which expresses the closeness of meaning between documents and is effective for short documents, and the BERT language model, which can consider the word order of documents and is effective for long documents, to obtain more independent topics.

3.5 Creating features

In existing studies[5, 20, 22], daily forecasting has been shown to perform better than weekly or monthly forecasting. Therefore, we obtain daily topic vectors by summarizing the acquired variance representations by day. As a method to summarize daily forecasting, we follow the previous study[19] and use the simple average of each post by day. Furthermore, we add the number of posts to these to generate the following features.

- Number of posts
- One-day difference of the number of posts, five-day difference
- One-day difference, five-day difference of the number of posts
- Two-day moving average of the number of posts, five-day moving average of the number of posts
- Two-day moving average difference of the number of posts and the number of posts, five-day moving average difference of the number of posts
- Ratio of two-day moving average of posts to posts, ratio of five-day moving average of posts
- Topic Vector
- One-day difference of topic vector, five-day difference
- One-day ratio of topic vector, five-day ratio of topic vector
- Two-day moving average of topic vectors, five-day moving average of topic vectors
- Two-day moving average difference of topic vectors, five-day moving average difference of topic vectors
- Ratio of two-day moving average of topic vectors, five-day moving average ratio of topic vectors

In addition, the financial time series data of the Nikkei 225 and the Nikkei VI are also added to the feature set. From these, we generate the following feature values.

- Opening price of the Nikkei Stock Average
- One-day difference, five-day difference of the opening price of the Nikkei Stock Average
- Nikkei 225 opening price one-day difference, five-day difference
- Two-day moving average of the opening price of the Nikkei 225, five-day moving average of the opening price of the Nikkei 225
- Difference between the opening price of the Nikkei 225 and its two-day moving average, and the difference between the opening price of the Nikkei 225 and its five-day moving average
- Ratio of the opening price of the Nikkei 225 to its two-day moving average, ratio of the two-day moving average of the opening price of the Nikkei 225 to its five-day moving average
- Opening price of the Nikkei VI
- One-day difference of the opening price of the Nikkei VI, five-day difference of the opening price of the Nikkei VI
- One-day difference of the opening price of the Nikkei VI, five-day difference of the opening price of the Nikkei VI
- Two-day moving average of the opening price of the Nikkei VI, five-day moving average of the opening price of the Nikkei VI
- Two-day moving average of the opening price of the Nikkei VI, five-day moving average of the opening price of the Nikkei VI
- Ratio of the two-day moving average of the opening price of the Nikkei VI, five-day moving average of the opening price of the Nikkei VI

In general, forecasts of financial instruments on a daily basis are based only on business days. This is because financial time series data are available only for business days. Therefore, it is difficult to forecast when a major event occurs on a holiday. In spite of the fact that social media and news are posted regardless of holidays or business days, existing studies[6, 19] ignore holiday information. We believe that the prediction accuracy can be improved if holiday information can be incorporated as well. In the present study, when predicting the business day after a holiday, we use the information posted on the previous day and the financial time series of the previous business day as input. Another possible method is to average the information of the holiday, but we do not use this method considering the importance of the information of the previous day and the smoothing of the information by averaging.

3.6 Machine learning algorithms

Previous studies have used logistic regression and random forests as machine learning algorithms to build risk prediction models, and LightGBM may perform better than these algorithms for classification tasks.[10] Therefore, in the present study, we compare the prediction models using logistic regression, random forests and LightGBM as machine learning algorithms. Note that since the positive labels are imbalanced data, the number of positive examples is extremely small, which may bias the learning of the model. Therefore, we will perform downsampling to reduce the number of negative examples in the training data and compare the results with the model without downsampling.

3.7 Training data and test data

We followed the previous study[19] and increased the learning period every day (Figure2). In the experimental period of T days, a constant n was determined in order to guarantee the initial amount of learning, and the learning period was n days from day 1 to day n. For evaluation, we use the first day of $n + d$ days with the time window defined as (proposed method-3.2). Here, n is increased by one, training data is added, and the process is repeated k times.

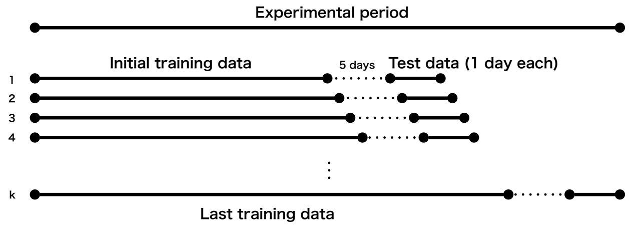

Figure 2: Method of increasing the study period by one day each day

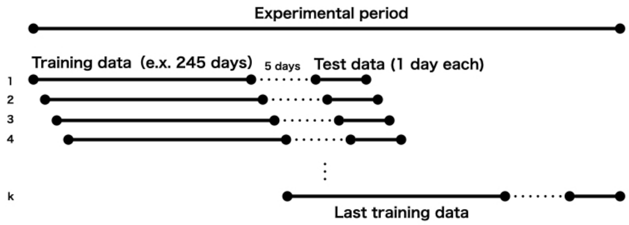

Figure 3: Method to fix the learning period of the model at 1, 2, or 3 years

Then, $n + d + k$ is repeated until T is reached, and the evaluation is performed on a total of k evaluation days. This is done in order to create a prediction model based on data that can actually be used and to determine the increase in the VI for each day. However, in the case of this method of increasing the learning period, there is a possibility that old information may become noise when the learning period becomes longer, resulting in a loss of accuracy.

Therefore, we prepared a model with a fixed learning period of 1, 2, or 3 years to compare the results (Figure3). When downsampling, negative examples are removed so that an arbitrary percentage of the labels are judged to have an increase in the VI over the n-day training period. The training data created by this method is used for training. Note that the ratio of the correct labels at the time of validation is not changed because of this.

4 EXPERIMENT

4.1 Data set

We collected 9,463,597 documents posted on the Nikkei Stock Average BBS from Nov. 26, 2012 to Sept. 30, 2020. The data of the Nikkei 225 and the Nikkei VI indices were collected from JPX Data Cloud. The collection period was from Nov. 26, 2012 to Oct. 07, 2020 (1920 business days).

The reason why there is a slight error between the data collection period of BBS and that of the Nikkei 225 and the Nikkei VI is that the last five business days are used to calculate the correct labels. The experimental period was from Dec. 3, 2012 to Sept. 30, 2020 (1,910 business days), excluding the first five business days. A total of 288 of these days corresponded to the correct answer. The validation period is from Dec. 14, 2016 to Sept. 30, 2020 (925 business days). Among them, the number of days that corresponded to the correct answer was 133, which is 14.37% of the validation period. The Nikkei Stock Average and the VI are based on the opening prices. This is because the prediction model assumes that the investment risk is determined before the start of the trading day.

4.2 Morphological analysis and preprocessing

4.2.1 morphological analysis. As a preprocessing step, the URL, HTML, and line feed codes of the submitted documents were removed. In addition, half-width hiragana and half-width katakana were converted to full-width alphanumeric characters, and full-width alphanumeric characters were converted to half-width alphanumeric characters. Morphological analysis was then performed. A morpheme is the smallest linguistic unit that has linguistic meaning.

4.2.2 Preprocessing according to the language model. Depending on the type of language model to be input, the preprocessing of the documents was changed. The following processes were performed on the submitted documents to be used as input for LDA and Doc2Vec language models. Morphological analysis was used to extract nouns, verbs, and adjectives for which the subtype was not numeric, non-independent, pronoun, or suffix from each document. In addition, the following words were added to the list of stop words[2] published by Kyoto University and removed as stop words: [aru, suru, chau, nai, naru, yaru].

The documents to be input to BERT were documents from which symbols were extracted and removed by morphological analysis. Mecab[11] was used for morphological analysis, and NEologd[21], which is strong in term of new words and unique expressions, was used as a dictionary. The updated version of NEologd was used on May 21, 2020.

4.3 Creating a risk prediction model

The combinations of language models and machine learning models are shown in the table1. The vector sizes of LDA and Doc2Vec are 32, 64, and 128 dimensions, and BERT is a hidden layer of pre-training with 768 dimensions. Here, LDA and Doc2Vec were implemented using Gensim, and BERT was implemented using HuggingFace's transformers. Grid search was used as a parameter setting for the machine learning algorithm. We divided the training data into two parts, determined the parameters, and then trained again on the entire training data. The settings of the train data for this experiment were $n = 980$ for the number of days n shown in (proposed method-3.7) (4 years where 1 year is 245 business days) and $d = 5$ for the time window d. For downsampling, we created two models, one that learns with the ratio of positive and negative examples at 3 : 7, and another that learns with the ratio at 5 : 5.

5 EXPERIMENTAL RESULTS

As evaluation indices for the quantitative investment risk prediction model, we use accuracy, precision, recall, and F1-measure, and compare the values of these indices. For precision, recall, and F1-measure, the values for positive examples are compared. The results of the logistic regression, random forests, and LightGBM models trained with an extended training period are shown in Tables2,3,4.

[2] http://svn.sourceforge.jp

Table 1: Combining language models and machine learning algorithms

	Logistic regression	Random forests	LightGBM
Downsampling	None 3:7 5:5	None 3:7 5:5	None 3:7 5:5
LDA	32 64 128	32 64 128	32 64 128
Doc2Vec	32 64 128	32 64 128	32 64 128
BERT	768	768	768

In logistic regression, the model trained by downsampling the ratio of positive to negative examples to 3:7 in 64 dimensions of LDA had the highest F1-measure and precision, with a precision of 0.19, a recall of 0.44, and an F1-measure of 0.26. The model trained by downsampling the ratio of positive to negative examples to 3:7 in BERT had the highest recall, a precision of 0.15, a recall of 0.59, and an F1-measure of 0.24. The model trained by downsampling the ratio of positive to negative examples to 3:7 in 64 dimensions of Doc2Vec had the same results as the model trained by downsampling the ratio of positive to negative examples to 3:7 in 64 dimensions of LDA, with the highest F1-measure. The precision was 0.17, the recall was 0.48, and the F1-measure was 0.26.

In the random forests, the model trained by downsampling the ratio of positive to negative examples to 3:7 in the 64th dimension of the LDA had the highest precision, with a precision of 0.67, a recall of 0.02, and an F1-measure of 0.03. The model trained by downsampling the ratio of positive to negative examples to 5:5 in 128 dimensions of LDA had the highest recall, precision, recall, and F1-measure of 0.15, 0.60, and 0.24, respectively. The model trained with a 5:5 downsampling of the ratio of positive to negative examples in BERT had the highest F1-measure, with a precision of 0.16, a recall of 0.58, and an F1-measure of 0.25.

In LightGBM, the model trained by downsampling the ratio of positive to negative examples to 3:7 in 64 dimensions of Doc2Vec had the highest precision, with precision of 0.24, a recall of 0.28, and an F1-measure of 0.26. The model trained by downsampling the ratio of positive to negative examples to 5:5 in 128 dimensions of LDA had the highest recall, precision, recall, and F1-measure of 0.17, 0.55, and 0.26, respectively. The F1-measure was the highest for these two models (at 0.26).

For each combination of machine learning model and language model, we fixed the training period at 1, 2, or 3 years for the combination with the best F1-measure value. The results are shown in the following tables5, 6, 7. Window indicates the number of years of fixed learning period.

In the logistic regression, the model trained by downsampling the ratio of positive to negative examples to 3:7 in 64 dimensions of LDA, with the training period fixed at 2 years, had the highest accuracy, with a precision of 0.18, a recall of 0.39, and an F1-measure of 0.24.

The model trained by downsampling the ratio of positive to negative examples to 5:5 in 128 dimensions of the LDA in the random forests, with the training period fixed at 2 years, had the highest accuracy, with a precision of 0.18, a recall of 0.46, and an F1-measure of 0.26.

Table 2: Results of logistic regression with increasing study period

	Dim	Downsampling	Precision	Recall	F1-measure
LDA	32	None	0.16	0.43	0.23
		3:7	0.17	0.49	025
		5:5	0.15	0.51	0.23
	64	None	0.15	0.25	0.19
		3:7	0.19	0.44	0.26
		5:5	0.16	0.53	0.25
	128	None	0.16	0.35	0.22
		3:7	0.16	0.44	0.23
		5:5	0.15	0.58	0.24
Doc2Vec	32	None	0.17	0.26	0.20
		3:7	0.16	0.36	0.22
		5:5	0.14	0.41	0.20
	64	None	0.17	0.38	0.23
		3:7	0.17	0.48	0.26
		5:5	0.15	0.54	0.24
	128	None	0.14	0.29	0.19
		3:7	0.16	0.37	0.22
		5:5	0.14	0.46	0.22
BERT	768	None	0.13	0.29	0.17
		3:7	0.16	0.46	0.23
		5:5	0.15	0.59	0.24

Table 3: Results of random forests with increasing study period

	Dim	Downsampling	Precision	Recall	F1-measure
LDA	32	None	0.29	0.05	0.05
		3:7	0.37	0.05	0.09
		5:5	0.15	0.59	0.24
	64	None	0.25	0.02	0.04
		3:7	0.67	0.02	0.03
		5:5	0.15	0.51	0.23
	128	None	0.26	0.04	0.07
		3:7	0.31	0.07	0.11
		5:5	0.15	0.60	0.24
Doc2Vec	32	None	0.33	0.08	0.12
		3:7	0.27	0.02	0.04
		5:5	0.14	0.44	0.22
	64	None	0.41	0.05	0.09
		3:7	0.27	0.07	0.11
		5:5	0.16	0.51	0.24
	128	None	0.30	0.05	0.08
		3:7	0.21	0.02	0.04
		5:5	0.21	0.02	0.04
BERT	768	None	0.18	0.02	0.03
		3:7	0.36	0.13	0.19
		5:5	0.16	0.58	0.25

In LightGBM, the model trained by downsampling the ratio of positive to negative cases in BERT to 5:5, with the training period fixed at 2 years, had the highest accuracy with a precision of 0.18, a recall of 0.35, and an F1-measure of 0.24. The model trained with a fixed period of two years had the highest accuracy, with a precision of 0.18, a recall of 0.35 and an F1-measure of 0.24.

Table 4: Results of LightGBM with increasing study period

	Dim	Downsampling	Precision	Recall	F1-measure
LDA	32	None	0.17	0.05	0.07
		3:7	0.17	0.22	0.19
		5:5	0.14	0.50	0.22
	64	None	0.22	0.05	0.07
		3:7	0.20	0.18	0.19
		5:5	0.14	0.47	0.22
	128	None	0.20	0.12	0.15
		3:7	0.15	0.23	0.18
		5:5	0.17	0.55	0.26
Doc2Vec	32	None	0.10	0.05	0.06
		3:7	0.16	0.14	0.15
		5:5	0.14	0.40	0.21
	64	None	0.12	0.06	0.08
		3:7	0.24	0.28	0.26
		5:5	0.16	0.52	0.24
	128	None	0.12	0.52	0.24
		3:7	0.12	0.12	0.12
		5:5	0.17	0.14	0.15
BERT	768	None	0.12	0.03	0.05
		3:7	0.20	0.23	0.21
		5:5	0.15	0.50	0.23

Table 5: Results of logistic regression with a fixed learning period

Model	Window	Precision	Recall	F1-measure
LDA_64_3:7	1	0.16	0.34	0.22
	2	0.18	0.39	0.24
	3	0.13	0.3	0.18
Doc2Vec_64_3:7	1	0.15	0.26	0.19
	2	0.16	0.44	0.24
	3	0.13	0.35	0.19
BERT_768_5:5	1	0.12	0.32	0.17
	2	0.16	0.44	0.23
	3	0.14	0.45	0.21

Table 6: Results of random forests with a fixed learning period

Model	Window	Precision	Recall	F1-measure
LDA_128_5:5	1	0.16	0.41	0.23
	2	0.18	0.46	0.26
	3	0.15	0.58	0.24
Doc2Vec_64_5:5	1	0.16	0.32	0.21
	2	0.14	0.28	0.18
	3	0.16	0.44	0.24
BERT_768_5:5	1	0.15	0.4	0.22
	2	0.18	0.45	0.26
	3	0.16	0.52	0.24

Table 7: Results of LightGBM with a fixed learning period

Model	Window	Precision	Recall	F1-measure
LDA_128_5:5	1	0.14	0.26	0.18
	2	0.16	0.37	0.22
	3	0.15	0.5	0.24
Doc2Vec_64_3:7	1	0.19	0.19	0.19
	2	0.12	0.14	0.13
	3	0.15	0.2	0.17
BERT_768_5:5	1	0.13	0.27	0.18
	2	0.18	0.35	0.24
	3	0.16	0.41	0.23

6 DISCUSSION

6.1 Effectiveness of the proposed method

The maximum accuracy obtained in this experiment was obtained by the model trained by downsampling the ratio of positive to negative examples to 3:7 in 64 dimensions of Doc2Vec in LightGBM, with a precision of 0.24, a recall of 0.28, and an F1-measure of 0.26. In the validation period, the number of days that corresponded to the correct answer (a large increase in the VI) was 133. This corresponds to 14.37% (= 133/925) of the validation period. In other words, the expected value of the correct answer in the case of random selection is 14.37%. Therefore, the proposed method may be able to predict the increase in the VI with approximately twice the accuracy of random selection.

6.2 Differences in accuracy due to market trends

There was no improvement in accuracy with different language models. On the other hand, down-sampling improved the accuracy for all language models. This suggests that down-sampling may be effective. However, the maximum F1-measure was 0.26, which was lower than the accuracy during the validation period of the previous study[19]. This may be due to the fact that the validation period includes the period of the Corona shock, which is a bear market with a downward trend.

Therefore, we evaluated the three models with the highest F1-measure among the experimental results in 395 business days from Nov. 17, 2014 to June 29, 2016, matching the validation period with previous studies[19]. The models tested were the logistic regression LDA model with 64 dimensions and 3:7 downsampling (LR_LDA_64_3:7), the LightGBM Doc2Vec model with 64 dimensions and 3:7 downsampling (LG_D2V_64_3:7), and the LightGBM (LG_D2V_64_3:7), and a 128-dimensional, 5:5 downsampled model of LightGBM's LDA (LG_LDA_128_5:5). The results are shown in Table8. The resulting maximum F1-measure is 0.40, and the recall is higher than in previous studies. Therefore, it is necessary to improve the accuracy in bear markets.

6.3 Differences in accuracy due to training period

As in the model in which the training period is increased by one day at a time, there was no improvement in accuracy for different

Table 8: Results of the same validation period as the previous study

Model	Precision	Recall	F1-measure
LR_LDA_64_3:7	0.28	0.51	0.36
LG_D2V_64_3:7	0.29	0.49	0.36
LG_LDA_128_5:5	0.30	0.62	0.40

language models. Since the accuracy is comparable to that of the model in which the training period is increased by one day at a time, it is unlikely that the decrease in accuracy is due to noise from old training data. These results suggest that differences in the market trend have a significant impact on the accuracy. In the future, it is necessary to verify the validation period of the downward trend using a model trained only for the period of the downward trend, and to confirm whether the forecast accuracy improves. Similarly, it is necessary to verify the validation period of the upward trend using a model trained only for the period of the upward trend to confirm the forecast accuracy.

7 CONCLUSION

In the present study, we focused on the VI to reduce investment risk and proposed a new method to predict a significant increase in the VI. We used a language model to obtain distributed representations of Yahoo! JAPAN stock BBS postings, aggregated them by day, and generated topic vectors. We constructed a prediction model by combining the differences in the distribution of these topics and financial time series data as input to a machine learning algorithm. At this time, we tested three different combinations of language models and three different machine learning models to find the most effective combination. We verified the long-term effectiveness of the VI prediction model by testing the model over a longer period than that of suwa et al[19]. As a result, we obtained a maximum F1-measure of 0.26 for the period of Dec. 14, 2016 to Sept. 30, 2020 (925 business days). When the validation period was the same as the validation period of the previous study, the maximum F1-measure obtained was 0.40 and the recall was higher than that of the previous study. The validation period of this experiment includes the period of the downtrend in the corona shock, which suggests that the prediction accuracy in the downtrend is low. Therefore, improving the accuracy during the downward trend is a future issue. It is necessary to confirm the risk aversion when trading according to the output of the prediction model during a downtrend in a trading simulation.

REFERENCES

[1] Werner Antweiler and Murray Z Frank. 2004. Is all that talk just noise? The information content of internet stock message boards. *The Journal of finance* 59, 3 (2004), 1259–1294.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[3] Chi Chen, Li Zhao, Jiang Bian, Chunxiao Xing, and Tie-Yan Liu. 2019. Investment Behaviors Can Tell What Inside: Exploring Stock Intrinsic Properties for Stock Trend Prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2376–2384. https://doi.org/10.1145/3292500.3330663

[4] Kai Chen, Yi Zhou, and Fangyan Dai. 2015. A LSTM-based method for stock returns prediction: A case study of China stock market. In *2015 IEEE International Conference on Big Data (Big Data)*. 2823–2824. https://doi.org/10.1109/BigData.2015.7364089

[5] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep Learning for Event-Driven Stock Prediction. In *IJCAI*.

[6] Xin Du and Kumiko Tanaka-Ishii. 2020. Stock Embeddings Acquired from News Articles and Price History, and an Application to Portfolio Optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3353–3363. https://doi.org/10.18653/v1/2020.acl-main.307

[7] Eugene F. Fama. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance* 25, 2 (1970), 383–417. http://www.jstor.org/stable/2325486

[8] Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019. Enhancing Stock Movement Prediction with Adversarial Training. arXiv:1810.09936 [q-fin.TR]

[9] Stefan Feuerriegel, Antal Ratku, and Dirk Neumann. 2016. Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. 1072–1081. https://doi.org/10.1109/HICSS.2016.137

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[11] T. KUDO. 2006. MeCab : Yet Another Part-of-speech and Morphological Analyzer. *http://mecab.sourceforge.jp* (2006). https://ci.nii.ac.jp/naid/10027284215/

[12] Richard Kyung and Minjun Kye. 2020. Study on the CBOE Volatility Data Forecast Using Statistical and Computational Simulations. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. 1–5. https://doi.org/10.1109/IEMTRONICS51293.2020.9216432

[13] Wei Li, Ruihan Bao, Keiko Harimoto, Deli Chen, Jingjing Xu, and Qi Su. 2020. Modeling the Stock Relation with Graph Network for Overnight Stock Movement Prediction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4541–4547. https://doi.org/10.24963/ijcai.2020/626 Special Track on AI in FinTech.

[14] Jintao Liu, Hongfei Lin, Xikai Liu, Bo Xu, Yuqi Ren, Yufeng Diao, and Liang Yang. 2019. Transformer-Based Capsule Network For Stock Movement Prediction. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. Macao, China, 66–73. https://aclanthology.org/W19-5511

[15] Wen Long, Zhichen Lu, and Lingxiao Cui. 2019. Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems* 164 (2019), 163–173. https://doi.org/10.1016/j.knosys.2018.10.034

[16] Joerg Osterrieder, Daniel Kucharczyk, Silas Rudolf, and Daniel Wittwer. 2020. Neural networks and arbitrage in the VIX. *Digital Finance* 2, 1 (2020), 97–115.

[17] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. 2013. Quantifying trading behavior in financial markets using Google Trends. *Scientific reports* 3, 1 (2013), 1–6.

[18] Kodai Sasaki, Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Tatsuo Yamashita, and Kota Tsubouchi. 2020. Evaluation of VI Index Forecasting Model by Machine Learning for Yahoo! Stock BBS Using Volatility Trading Simulation. https://doi.org/10.24251/HICSS.2020.305

[19] Hirohiko Suwa, Yuki Ogawa, Eiichi Umehara, Kento Kakigi, Keiichi Yasumoto, Tatsuo Yamashita, and Kota Tsubouchi. 2017. Develop method to predict the increase in the Nikkei VI index. In *2017 IEEE International Conference on Big Data (Big Data)*. 3133–3138. https://doi.org/10.1109/BigData.2017.8258289

[20] PAUL TETLOCK, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance* 63 (02 2008), 1437–1467. https://doi.org/10.2139/ssrn.923911

[21] Sato Toshinori. 2015. Neologism dictionary based on the language resources on the Web for Mecab. https://github.com/neologd/mecab-ipadic-neologd

[22] Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer. 2013. Semantic Frames to Predict Stock Price Movement. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 873–883. https://aclanthology.org/P13-1086

[23] Yumo Xu and Shay B. Cohen. 2018. Stock Movement Prediction from Tweets and Historical Prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1970–1979. https://doi.org/10.18653/v1/P18-1183

[24] Linyi Yang, Ruihai Dong, Tin Lok James Ng, and Yang Xu. 2019. Leveraging BERT to Improve the FEARS Index for Stock Forecasting. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. Macao, China, 54–60. https://aclanthology.org/W19-5509