

Exploring Social Media's Role in Predicting Stock Market Trends

Masatoshi Beppu¹, Masatomo Ide¹, Seita Nagashima², Satoshi Ikeda^{1*},
Amane Takei¹, Makoto Sakamoto¹, Tsutomu Ito³, Takao Ito⁴

(¹Graduate School of Engineering, University of Miyazaki, Japan),

(²MEITEC CORPORATION, Japan),

(³National Institute of Technology, Ube College, Japan),

(⁴Hiroshima University, Japan)

*Corresponding Author

E-mail: bisu@cs.miyazaki-u.ac.jp

Abstract

This study analyzes tweets from the official Twitter accounts of NHK News and Nikkei to incorporate sentiment data into a predictive model for the Nikkei Stock Average. Adding sentiment data improved the R^2 score from 45.1% to a maximum of 70.5%, indicating the potential of SNS data in forecasting social indicators. However, no strong correlation between sentiment data and stock prices was observed. Challenges include the short data collection period and the difficulty of sentiment analysis in Japanese. Future work should focus on employing more effective methods for extracting sentiment.

Keywords: Sentiment Analysis, Nikkei Stock Average, Social Indicators, Natural Language Processing

1. Introduction

1.1. Research background

The rapid proliferation of social media in recent years has heightened interest in text mining using SNS. Among them, Twitter, one of the largest SNS platforms, is believed to hold substantial potential value. A study by Bollen et al. [1] utilized Twitter data to conduct sentiment analysis based on the psychological index POMS and successfully predicted the Dow Jones Industrial Average three days ahead with an accuracy of 86.7%. This suggests a correlation between Twitter data and social indicators.

1.2 Research Objective

This study aims to analyze tweets from the official Twitter accounts of NHK News and Nikkei using two sentiment analysis methods. By incorporating sentiment data into Nikkei 225 stock price data, the study seeks to predict closing prices and verify the extent to which SNS tweet data can contribute to forecasting social indicators.

2. Methodology

2.1 Acquisition of Tweets from Twitter

TwitterAPI [2] was employed to collect tweet data. StreamingAPI was used to retrieve real-time tweets, including tweet IDs, timestamps, and contents, and store them in a database. By executing the API, tweets from the past 15 minutes were collected, and this process was repeated every 15 minutes.

2.2 Sentiment Analysis and Time-Series Conversion

Two methods were used for sentiment analysis: MeCab+PN Table and Google Natural Language API (GNL). Both methods return sentiment scores where positive sentiment approaches 1 and negative sentiment approaches -1. The average sentiment score per day was calculated based on tweet timestamps, creating time-series data referred to as sentiment data (Fig. 2.1).

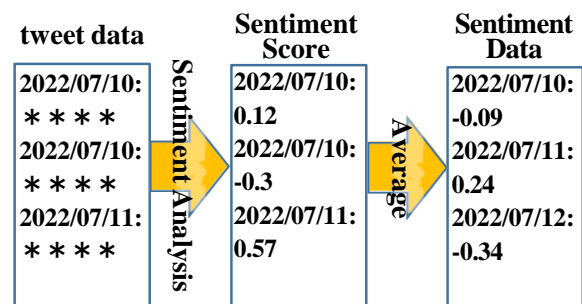


Fig. 2.1 Creation of Sentiment Data

2.2.1 Sentiment Analysis Using MeCab and PN Table

MeCab [3] was used for morphological analysis to decompose text into words. Then, corresponding sentiment scores were assigned using the PN Table [4] developed by Takamura et al. The sentiment value for each tweet was calculated by dividing the total sentiment score by the number of words.

2.2.2 Sentiment Analysis Using Google Natural Language API

The GNL [5], a text analysis tool by Google, was utilized for sentiment analysis. This API provides sentiment scores (-1 to +1) and magnitudes using a pre-trained model. The score was adopted as the evaluation metric in this study.

2.3 Building the Learning Model Using Sentiment Data

A machine learning model was developed using the LSTM algorithm, originally proposed by Hochreiter et al. [6], to learn from sentiment data and the closing prices of the Nikkei 225 index. Since the sentiment data and stock price data have different units, the data were normalized (Eq. (1)) to scale them between 0 and 1. The explanatory variables consisted of sentiment data and the closing prices of the Nikkei 225 for the past 3 days (look_back), while the target variable was the closing price of the Nikkei 225.

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

The structure of the LSTM was based on the approach of Mhamed et al. [7]. The number of nodes, activation functions (ReLU, Softmax, Tanh), and optimization algorithms (Adam, RMSprop) were selected, and approximately 100 tuning tests were performed to finalize the parameter settings.

The performance of the model was evaluated using two key metrics: the coefficient of determination (R² score) and Root Mean Squared Error (RMSE).

2.4 Evaluation of the Learning Model

2.4.1 Coefficient of Determination (R² score)

The R² score [8] measures how well the model explains the data, with values closer to 1 indicating better performance. The R² score is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_{\text{true},i} - \overline{y_{\text{true}}})^2}{\sum_{i=1}^N (y_{\text{true},i} - y_{\text{pred},i})^2}$$

where $\overline{y_{\text{true}}}$ is the mean of the actual values.

2.4.2 Root Mean Squared Error (RMSE)

RMSE [9] quantifies the average prediction error, with smaller values indicating better accuracy. It is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (y_{\text{true},i} - y_{\text{pred},i})^2}$$

In this study, the training data was divided into 60% for the early part of the time series and 40% for the later part, and model evaluation was performed using this split.

3. Experiment and Discussion

From July 10, 2022, to December 10, 2022, a total of 22,850 tweets from the official NHK News account (@nhk_news) and 3,688 tweets from the official Nikkei account (@nikkei) were collected using the Twitter API. Sentiment analysis was conducted on the collected tweets using MeCab+PN Table and GNL, assigning evaluation scores ranging from -1 to +1 to each tweet and creating time-series sentiment data (Fig.3.1). Sentiment data evaluated using MeCab+PN Table for NHK News is referred to as "NHK_PN," and data evaluated using GNL is referred to as "NHK_GNL." Similarly, the data for Nikkei are termed "Nikkei_PN" and "Nikkei_GNL."

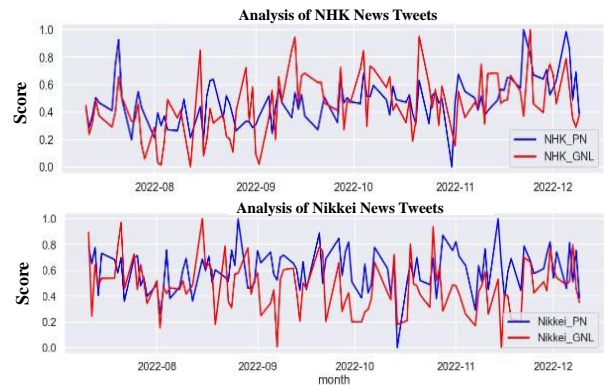


Fig. 3.1 Sentiment Data for @nhk_news and @nikkei

Table 3.1, which presents the correlation coefficients, lists the strength and direction of linear relationships between sentiment data and the Nikkei Stock Average, as well as among the sentiment datasets themselves. A correlation coefficient measures the linear association between two variables, where values range from -1 to 1. Positive values indicate a direct relationship, negative values indicate an inverse relationship, and values close to 0 indicate little to no linear relationship.

Table 3.1 Correlation of Sentiment Data with Nikkei 225 Closing Prices.

	NHK_PN	Nikkei_PN	NHK_GNL	Nikkei_GNL	Nikkei_close
NHK_PN	1.000	0.088	<u>0.398</u>	0.028	0.024
Nikkei_PN	0.088	1.000	0.024	<u>0.349</u>	0.002
NHK_GNL	0.398	0.024	1.000	0.142	-0.049
Nikkei_GNL	0.028	0.349	0.142	1.000	0.191
Nikkei_close	0.024	0.002	-0.049	0.191	1.000

As shown in Table 3.1, a weak positive correlation was found within the same account, but there was almost no correlation between different accounts. The sentiment data most correlated with the Nikkei Stock Average was "Nikkei_GNL," though the correlation was very weak.

Nikkei_PN and Nikkei_GNL were combined to create 16 types of learning models. These models were labeled with indices: 0 for the Nikkei Stock Average closing price (Nikkei_close), 1 for NHK_PN, 2 for NHK_GNL, 3 for Nikkei_PN, and 4 for Nikkei_GNL. For instance, when the explanatory variable was limited to the Nikkei closing

price, the model was labeled as F0. The learning models were applied to the last 40% of the dataset as validation data to predict the Nikkei closing price. The predictive performance of each model was evaluated using the R² score and RMSE.

Table 3.2 Results for R² Score and RMSE.

Explanatory variable	R2_score	RMSE
F0	0.4514	0.0143
F01	0.5824	0.0139
F02	0.6110	0.0101
F03	0.6366	0.0095
F04	0.6460	0.0092
F012	0.6374	0.0094
F013	0.6252	0.0098
F014	0.6574	0.0089
F023	0.6370	0.0095
F024	0.6914	0.0080
F034	0.6233	0.0089
F0123	0.6127	0.0101
F0124	0.7047	0.0077
F0134	0.6229	0.0098
F0234	0.6709	0.0086
F01234	0.5954	0.0105

As shown in Table 3.2, the F0 model achieved an R² score of only 45.1%, indicating poor performance as it fell below 50%. However, by adding sentiment data as explanatory variables, the performance improved significantly, with the F0124 model achieving an R² score of 70.5%. Fig. 3.2 and 3.3 illustrate the training and prediction results for model F0124.

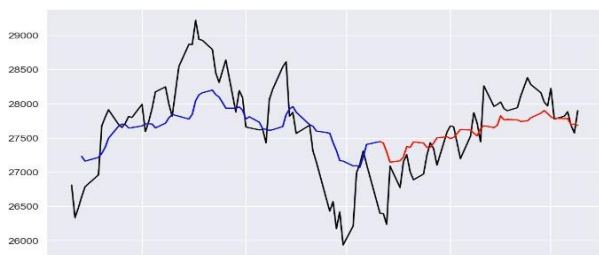


Fig. 3.2 Training and Prediction Results for F0.

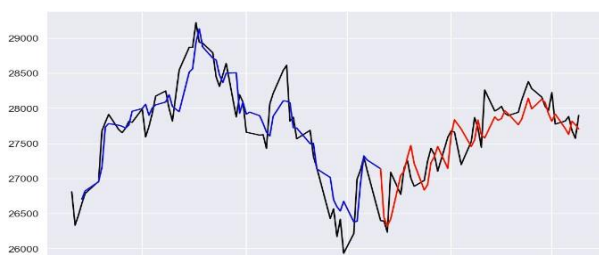


Fig. 3.3 Training and Prediction Results for F0124

In Fig. 3.2 and 3.3, the horizontal axis represents dates, while the vertical axis indicates the Nikkei closing price. The black line denotes the actual closing price, the blue line represents the predicted values from the training data, and the red line corresponds to the predicted values from the validation data. In the F0 model, the blue line fails to capture the black line adequately. In contrast, the F0124 model shows a strong resemblance between the blue and black lines, with the red line also capturing the characteristics more effectively than the F0 model.



Fig. 3.4 Errors associated with the F0 and F0124 models.

Fig. 3.4 illustrates the prediction errors for the validation data. Each point represents normalized data, with the x-coordinate indicating the actual data and the y-coordinate indicating the predicted data. The black line represents $y=x$, and points closer to this line indicate lower errors. In the F0124 model, points are more closely aligned with the $y=x$ line compared to the F0 model. Notably, points highlighted in the red box (2022/10/12 and 2022/10/13) represent cases where the F0124 model performed significantly better than the F0 model, while points in the blue box (2022/10/14) indicate instances of degraded performance.

Additionally, the "Crimean Bridge Explosion" on October 8, 2022, a critical event in the Ukraine War, likely influenced NHK News articles, leading to a drop in three predicted values.

4. Conclusion

This study demonstrated that incorporating sentiment analysis data from the official Twitter accounts of NHK News and Nikkei into machine learning models significantly improved the prediction accuracy of the Nikkei Stock Average. Using MeCab+PN Table and GNL for sentiment analysis, the R² score of the F0 model (which used only the Nikkei closing price as an explanatory variable) was 45.1%, whereas the F0124 model, which included sentiment data, achieved an R² score of 70.5%. This highlights the substantial predictive value of sentiment data.

However, unlike the findings of Bollen et al., no strong correlation between tweet data and stock prices was observed. This may be due to the nature of the collected tweets, which primarily consisted of news articles and might not adequately reflect the emotions of the Japanese populace. NHK News articles, in particular, tended to

focus on events and were often biased toward negative content. Additionally, the news articles themselves were not highly emotive in their language.

Future research should explore more efficient methods for extracting emotional information from Twitter data. Extending the analysis period and increasing the data volume would also likely enable the construction of more accurate predictive models.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP24K0792901 and JP24K15516.

References

1. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 18.
2. Toriumi, F. (2015). Twitter-based big data collection and analysis (in Japanese). *Organizational Science*, 48(4), 47-59.
3. MeCab Official. (2022, December 6). MeCab: What is MeCab. Retrieved from <https://taku910.github.io/mecab>
4. Takamura, D. (2022, December 6). Word sentiment polarity correspondence table. Tokyo Institute of Technology. Retrieved from http://www.lr.pi.titech.ac.jp/~takamura/pndic_en.html
5. Google Cloud. (2022, December 6). Natural Language AI. Retrieved from <https://cloud.google.com/natural-language?hl=ja>
6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
7. Moghar, A., & Hamiche, M. (2020). Stock market prediction using LSTM recurrent neural network. *Procedia Computer Science*, 170, 1168-1173.
8. Kvalseth, T. O. (1985). Cautionary note about R^2 . *The American Statistician*, 39(4), Part 1.
9. Hyndman, R. J. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688.

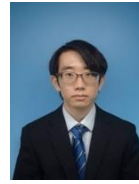
Authors Introduction

Mr. Masatoshi Beppu



He is a master student at Department of Computer Science and System Engineering, University of Miyazaki. His current research topic is an educational support using VR for children with physical disabilities.

Mr. Masatomo Ide



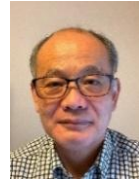
He is a graduate student at the Graduate School of Engineering, University of Miyazaki. His recent main research activity is to utilize various digital technologies such as web development to contribute to the local community.

Mr. Seita Nagashima



He graduated from the Faculty of Engineering at Miyazaki University in 2023 and is currently employed at MEITEC CORPORATION.

Dr. Satoshi Ikeda



He received Ph.D. from Hiroshima University and is currently an associate professor in the Faculty of Engineering at the University of Miyazaki. His research interests include time series data forecasting, graph theory, probabilistic algorithms, combinatorial optimization, and measure theory.

Dr. Amane Takei



He is working as Professor for Department of Electrical and systems Engineering, University of Miyazaki, Japan. His research interest includes high performance computing for computational electromagnetism, iterative methods for the solution of sparse linear systems, etc. Prof. Takei is a member of IEEE, an expert advisor of IEICE, a delegate of the Kyushu branch of IEEEJ, a director of JSST.

Dr. Makoto Sakamoto



He is a professor in the Faculty of Engineering, University of Miyazaki. His current main research interests are computer science and information processing. In particular, he deals with automata, language theory, computation, CG, image processing, virtual technologies, data science, artificial intelligence, and so on.

Dr. Tsutomu Ito



He is Associate Professor of the Department of Business Administration at National Institute of Technology, Ube College, Japan. His current research interests include internet of things (IoT), mechanical engineering, artificial intelligence (AI), automata theory, quantitative analysis of Japanese Keiretsu. Dr. Ito earned his doctor degree of Engineering from Hiroshima University, Japan in 2018.

Dr. Takao Ito



He (Ph.D., Kyoto University) is a professor of Management of Technology, Graduate School of Engineering, Hiroshima University.
