Financial Text Data Analytics Framework for Business Confidence Indices and Inter-Industry Relations

Hiroki Sakaji[1] , **Ryota Kuramoto**[1] , **Hiroyasu Matsushima**[1] , **Kiyoshi Izumi**[1] ,
Takashi Shimada[1] and **Keita Sunakawa**[3]
[1]School of Engineering, The University of Tokyo
[2]The Bank of Okinawa
{sakaji, matsushima, izumi}@sys.t.u-tokyo.ac.jp, m2017rkuramoto@socsim.org,
houjin-kikaku3@okinawa-bank.co.jp

Abstract

In this paper, we propose a novel framework for analyzing inter-industry relations using the contact histories of local banks. Contact histories are data recorded when employees communicate with customers. By analyzing contact histories, we can determine business confidence levels in the local region and analyze inter-industry relations using industrial data that is attached to the contact history. However, it is often difficult for bankers to create analysis programs. Therefore, we propose a banker-friendly inter-industry relations analysis framework. In this study, we generated regional business confidence indices and used them to analyze inter-industry relations.

1 Introduction

In economics and finance, various data are used to forecast and analyze future market trends. The analysis methodology for financial forecasting is generally divided into technical analysis and fundamental analysis, depending on the type of data used. The former analyzes historical transaction prices and volume, while the latter involves a wider range of information, including forecasts of a company's business performance in addition to the data used in technical analysis. Although investors and analysts use the above-mentioned methodologies in conjunction with one another, they focus more on numerical data than on textual information, as the former is simpler to handle. However, the latter can also be useful for market analysis. For example, economic analysis reports written by financial experts provide rich data, while newspaper articles report important information concerning past events and their impact. In addition, comments on social networking sites reflect people's impressions of the economy. The demand for using textual information to forecast future trends is increasing, and a number of studies using machine learning have been conducted.

Local banks are also an important source of financial text data. For example, banks generate demand histories, financial summaries, and contact histories. In this study, we focus on contact histories, which document customers' authentic views concerning businesses and local economic conditions. These data can provide important information about the industry sector. In this study, we aim to analyze relations between industry sectors in a region by using contact histories to create business confidence indices conveying industry sector information. These indices can serve as valuable tools for visualizing local economic conditions and evaluating local industries, which can be useful for the revitalization of local communities. Furthermore, business confidence indices are useful as reference indices for investment or policy decisions, as they capture economic trends.

In this paper, we propose a novel framework for visualizing local economic and local inter-industry relations using contact histories. Local business confidence indices currently exist[1]; however, they often lack immediacy, as they reflect conditions from several months prior to their publication. With our proposed framework, however, it is possible to use text data to create an immediacy index to analyze local inter-industry relations. Furthermore, our framework is the first to generate business confidence indices using the contact histories of local banks.

The contribution of this study is to provide a framework for generating business confidence indices and analyzing inter-industry relations based on text mining techniques. This study makes the following contributions: 1) a means of generating business confidence indices that enables the frequent presentation of data while revealing correlations with existing indicators; 2) the visualization of business conditions based on generated indices in a manner that permits the analysis of inter-industry relations; and 3) a clarification of the effectiveness of the data owned by local banks, which have not been utilized up to now. This study is the first to analyze inter-industry relations using text data, and the usefulness of this approach is demonstrated by applying it to real data.

1.1 Related work

Bollen et al. have demonstrated that Twitter moods are useful for forecasting the Dow Jones Industrial Average [Bollen *et al.*, 2011]. The researchers used a self-organizing fuzzy neural network for forecasting with which they were able to predict rise and fall with an accuracy of over 80%. Schumaker et al. proposed a machine learning approach for predicting stock prices using financial news article analysis

[1] http://www.okigin-ei.co.jp/report_DI.html

Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2019), pages 40-46, Macao, China, August 12, 2019.

[Schumaker and Chen, 2009]. Their method predicted indicators and stock prices; however, it did not analyze inter-industry relations.

With regard to financial text mining, Sakai et al. proposed a method for extracting causal information from Japanese financial articles concerning business performance [Sakai and Masuyama, 2007]. Their method used clues for extracting causal information, and it was able to automatically gather clues using the bootstrapping method. Sakaji et al. proposed a method for automatically extracting basis expressions indicating economic trends from newspaper articles using a statistical approach [Sakaji *et al.*, 2008]. In addition, Koppel et al. proposed a method for classifying a company's news stories on the basis of their apparent impact on the company's stock performance [Koppel and Shtrimberg, 2006]. Ito et al. proposed a neural network model for visualizing online financial textual data [Ito *et al.*, 2018]; this model determined the sentiment of words and their categories. Lastly, Milea et al. predicted the MSCI euro index (upwards, downwards, or constant) based on fuzzy grammar fragments extracted from a report published by the European Central Bank [Milea *et al.*, 2010].

The above-mentioned studies all extract information for investors or predict stock prices using information extracted from text data. In the present study, however, our objective is to analyze inter-industry relations using the contact histories of a local bank.

2 Proposed framework

In this section, we describe our proposed framework, which generates local business confidence indices and analyzes local inter-industry relations using the contact histories of local banks. Our framework uses a learned bidirectional long short-term memory (BiLSTM) model [Graves and Schmidhuber, 2005] for generating business confidence indices from banks' contact histories. It then analyzes inter-industry relations using the generated business confidence indices.

The procedure of our proposed framework is as follows.

Step 1: As raw input data, text related to finance and the economy are selected (e.g., newspaper articles, economic trend surveys) that reflects local economic conditions. Then, a word embedding model is created from the input. Next, our method assigns monthly sentiment scores to the inputs using the created word embedding model and learned BiLSTM. Finally, business confidence indices are generated by summing each month's sentiment score.

Step 2: Our method analyzes inter-industry relations using Granger causality analysis, impulse response analysis, and forecast error variance decomposition (FEVD) of the generated business confidence indices.

In Step 1, we generated word embeddings with 200 dimensions and default settings using *gensim*. In addition, we assumed that each raw input data entry corresponded to an industry sector. In this step, by changing the scale, our framework was also able to generate weekly indices. An overview of our framework is presented in Figure 1. In our framework, by inputting text data, users can obtain four types of generated data: business confidence indices, graphs of inter-industry relations, graphs of inter-industry impacts, and graphs of FEVD.

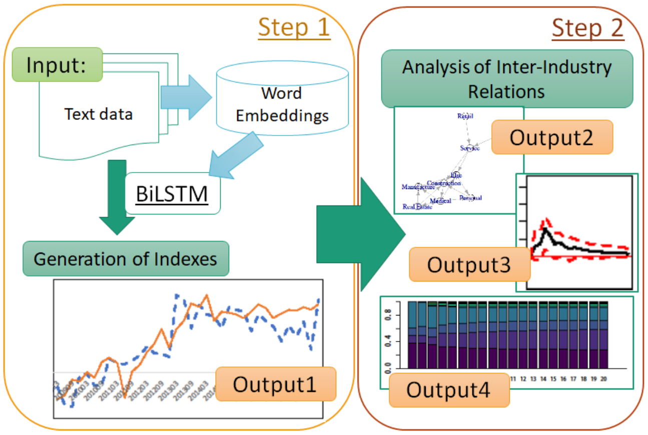


Figure 1: Proposed framework.

Although other machine learning methods exist, our framework uses BiLSTM, as it is a model in which long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] is bidirectional. A detailed explanation of the machine learning method selection is provided in Section 2.1.

2.1 Machine learning method selection

To select a machine learning method, we experimented with a classification test using the Economy Watchers Survey[2], which provides a timely and accurate overview of regional economic trends. The Economy Watchers Survey is scored on a 1–5 point Likert scale according to the level of business confidence, and each response is accompanied by comments. Using the survey responses, we constructed a machine learning method that predicts points using comments as input. To the best of our knowledge, the Economic Watcher Survey was the only available source of tagged text data on economics with a sufficient amount of information.

In this study, we used logistic regression (LR), random forest (RF), multi-layer perceptron (MLP), LSTM, and BiLSTM for classification testing. We used 20,000, 5,000, and 5,000 pairs of points and comments as training data, validation data, and test data, respectively. The pairs of points and comments were randomly extracted from the Economy Watchers Survey from 2010 to 2017. The results are presented in Table 1.

Table 1 reveals that BiLSTM is the optimal machine learning method for classifying the Economy Watchers Survey; as a result, it was adopted in our framework.

2.2 Bidirectional LSTM

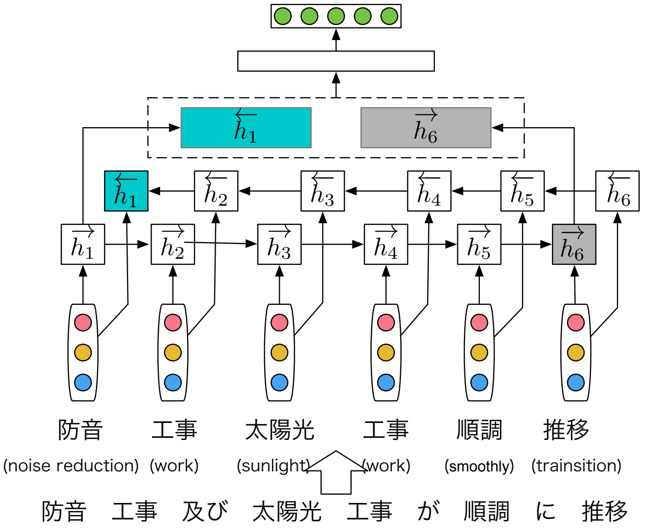
Figure 2 illustrates our BiLSTM model.

As input, we used word embeddings of content words (nouns, verbs, and adjectives) selected using morphological

[2] https://www5.cao.go.jp/keizai3/watcher-e/index-e.html

Table 1: Classification test results

	Accuracy
LR	0.612
RF	0.525
MLP	0.617
LSTM	0.636
BiLSTM	**0.642**



$\overleftarrow{h}_1$
$\overrightarrow{h}_6$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$ $\overleftarrow{h}_5$ $\overleftarrow{h}_6$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_4$ $\overrightarrow{h}_5$ $\overrightarrow{h}_6$
防音 工事 太陽光 工事 順調 推移
(noise reduction) (work) (sunlight) (work) (smoothly) (trainsition)
防音 工事 及び 太陽光 工事 が 順調 に 推移

Figure 2: BiLSTM model.

analysis. Here, we define LSTM processing from the beginning of a sentence as $\overrightarrow{LSTM}$ and from the end of the sentence as $\overleftarrow{LSTM}$. For each input, our method obtains $\{\overrightarrow{h_i}\}_i^n$ and $\{\overleftarrow{h_i}\}_i^n$ through LSTM($\overrightarrow{LSTM}, \overleftarrow{LSTM}$).

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(e_i), \overleftarrow{h_i} = \overleftarrow{LSTM}(e_i) \quad (1)$$

Here, n is the number of input words, and e_i is the word embedding entered ith words.

Then, $\overleftarrow{h_1}$ and $\overrightarrow{h_n}$ are concatenated and entered into the output layer as follows:

$$s = [\overleftarrow{h_1}; \overrightarrow{h_n}] \quad (2)$$

$$t = tanh(W_s \cdot s + b_s) \quad (3)$$

$$Y = W_t \cdot t + b_t \quad (4)$$

Here, $h_1 \in \mathbb{R}^m$, $h_n \in \mathbb{R}^m$, $s \in \mathbb{R}^{2m}$, and $t \in \mathbb{R}^l$.

Here, W_s and W_t are weighted matrices, b_s and b_t are bias vectors, m is the number of units in the hidden layer, l is the number of units in the middle layer, and Y is an output layer comprising $Y = (y_1, y_2, y_3, y_4, y_5)$. In this study, the output layer is activated by formula 5.

$$\beta_i = \log\left(\frac{\exp(y_i)}{\sum_j \exp(y_j)}\right) \quad (5)$$

Here, β_i is the activated output layer, and y_i is the output layer passed to the activation function. Finally, our method selects y_n as the maximum value from the output layer as output.

2.3 Granger causality analysis

We determined the causal relations between industry types using a Granger causality test [Granger, 1969]. Specifically, the null hypothesis is that there is no Granger causality from Industry A to B, and whether to accept or reject this hypothesis is determined by comparing its p value with the significance level. In this study, we used partial Granger causal analysis [Guo *et al.*, 2008], which extends Granger causality to multivariate data. Partial Granger causality analysis reduces the influence of exogenous factors and latent variables, and is useful in multivariate data analysis.

2.4 Impulse response analysis

The Granger causality test is a method for determining the causality between time series data; however, it cannot measure the strength of a relationship. The impulse response function is used as a means to quantitatively capture a relationship. By analyzing the degree to which changes in one variable contribute to changes in other variables, it is possible to perform a quantitative analysis. In the proposed framework, the orthogonalized impulse response function [Sims, 1980] is used and analyzed to quantitatively evaluate how changes in one industry affect other industries based on the business confidence index for each industry category.

2.5 Forecast error variance decomposition

FEVD is a method used to quantitatively analyze relationships between variables, and it is often used for the same purpose as impulse response analysis. Impulse response analysis is a method used to measure the influence of the fluctuation of a variable on other variables. In contrast, FEVD is used to analyze the business cycle of industries. It analyzes the relationships between variables by clarifying the degree to which each variable contributes to the forecast error.

In this study, we assumed that economic flow can be understood by quantitatively evaluating the contribution of each type of industry to changes in the business condition index of other types of industries.

3 Application

In this study, we focus on Okinawa Prefecture in Japan as an example of utilizing financial or economic text information and analyzing inter-industry relations within a region.

3.1 Data

The text data used in this study is as follows.

News texts published by Ryukyu Shimpo[3], a local newspaper in Okinawa Prefecture in Japan. In this study, newspaper articles from Ryukyu Shimpo from January 2014 to December 2017 were used as data. Due to its geographical specificity and historical background, Okinawa Prefecture has the lowest penetration of national newspapers in Japan. The primary reason for targeting this local newspaper for analysis is its large regional influence.

[3] https://ryukyushimpo.jp

Contact history owned by Okinawa Bank, a local bank in Okinawa Prefecture in Japan. Contact history is the text data recorded when a bank employee communicates with customers. Thus, the unit of a contact history is a transaction. Contact histories are categorized into industry sectors by local banks; examples of contact histories are provided in Table 2. Economic circulation is supported by the flow of money, most of which occurs via financial institutions. In other words, data owned by a bank representing a local area is assumed to reflect the local economic entity. A contact history records not only actual transactions, but also customer backgrounds and their views on economic and business conditions. Therefore, it comprises useful data to evaluate the levels of business confidence. In this study, we used approximately 8 million contact histories from April 2011 to July 2017.

Table 2: Examples of contact histories

他行住宅ローン借換推進で訪問するが不在。名刺、チラシ投函する。(I visited for a mortgage loan promotion, but the client was not present. So, I put cards and flyers.)
現況確認　台風の影響で前期は売上減少 (Confirming the present condition, sales decreased in the previous year owing to typhoons.)
キャッシュカード暗証番号変更でご来店。定期商品や保険を少しご案内しました。老後のことが不安との事なので時間があるときに検討する。(Customer visited us to change the cash card security code. We introduced a few regular products and insurance. The customer said, “I will consider these products when I have time because I am anxious about old age.”)

3.2 Generation of business confidence index

News text data

Using text data from newspaper articles, in this experiment, the generated business confidence index was evaluated as two viewpoints. To evaluate the effectiveness of the generated business confidence index by unit change of text data, business confidence was supplied for various units of text data, such as news articles, as follows: (a) sentences, (b) paragraphs, (c) economic articles, and (d) all articles. Therefore, using the method proposed in [Sakaji *et al.*, 2008], text data was divided into roughly four types of units, and a business confidence index was generated using each unit.

Contact histories

he business confidence index was derived from the contact history on the basis of the sentiment classification model learned from the Economic Watcher Survey data. As described in Section 2.1, by performing sentiment classification using the Economic Watcher Survey, an optimum sentiment classification model was produced, and the adopted model calculated the sentiment value in the contact history. Five sentiment values were provided: good, slightly good, neutral, somewhat bad, and bad. The average sentiment value of a certain period was then calculated as the business confidence index. An analysis of the inter-industry relationship was then performed on the basis of the generated business confidence index.

4 Result and discussion

4.1 Evaluation of generated business confidence index

The generated business confidence index was evaluated through a comparison with an existing index, such as the Okigin cooperation trend survey in this study. Thus, the correlation coefficient r calculated in (6) was adopted as the evaluation criteria.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2} \qquad (6)$$

Here, x is the generated index value, and y is the existing index value. Additionally, $\overline{x}$ is the average value of the generated index, while $\overline{y}$ is the average value of the existing index.

Table 3 presents the correlations between the generated business confidence indices and the existing index. Additionally, Figures 3 and 4 present the generated indices.

Table 3: Correlations between generated business confidence indices and existing index

	Correlation
(a)Per sentence	0.301
(b)Per paragraph	0.275
(c)Per article (economic)	0.300
(d)Per article (all)	0.175
Contact histories	0.856

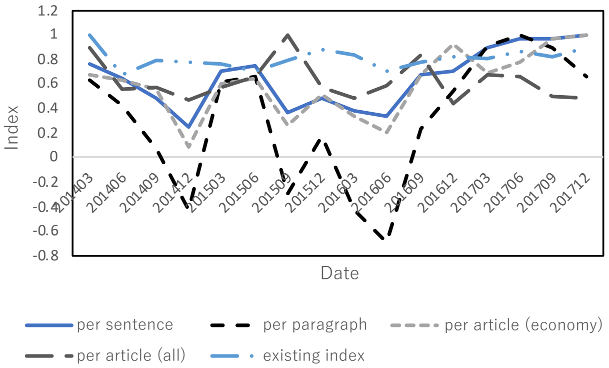


Figure 3: Generated index based on news text data.

Table 3 indicates that the index generated from the contact histories outperforms the generated indices from the news text. The reason for this is most likely the high affinity of the contact histories. Contact histories occasionally include customers’ levels of business confidence; therefore, a business confidence index can be generated effectively using contact histories.

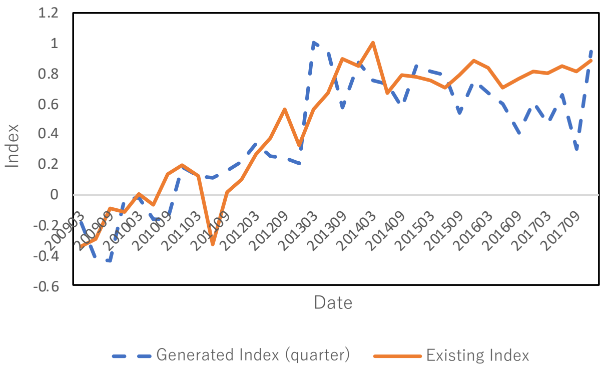


Figure 4: Generated index based on the contact histories.

4.2 Analysis of inter-industry relations

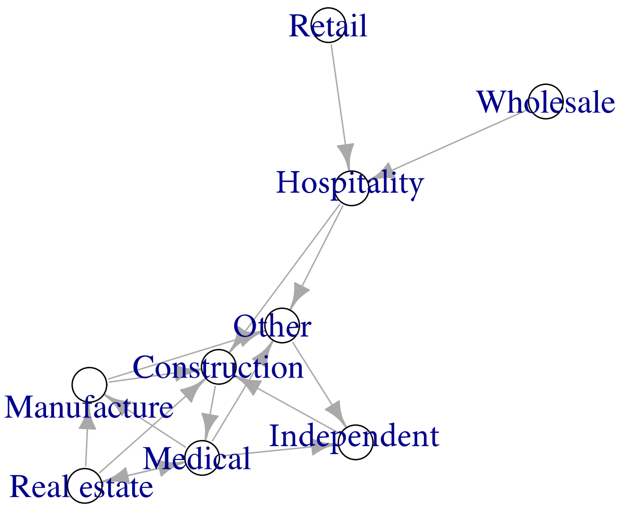


Figure 5: Visualization of partial Granger causality analysis among industries

function of the hospitality industry sector on the other industries, while dashed red lines represent the two-sided 95% confidence interval. In Figure 6, the horizontal and vertical axes represent the monthly and standard deviation, respectively. This figure indicates that the hospitality industry has affected business confidence in other industries for a long period of time. For example, at the 20th term in Figure 6, the manufacturing and wholesale industries have a high score (0.004).

Figure 7 presents the FEVD results for business confidence by industry sector. In this Figure 7, the vertical axis indicates the contribution of the influence by other industries, while the horizontal axis indicates monthly. In Figure7, it can be seen that the retail, wholesale, and hospitality industries have

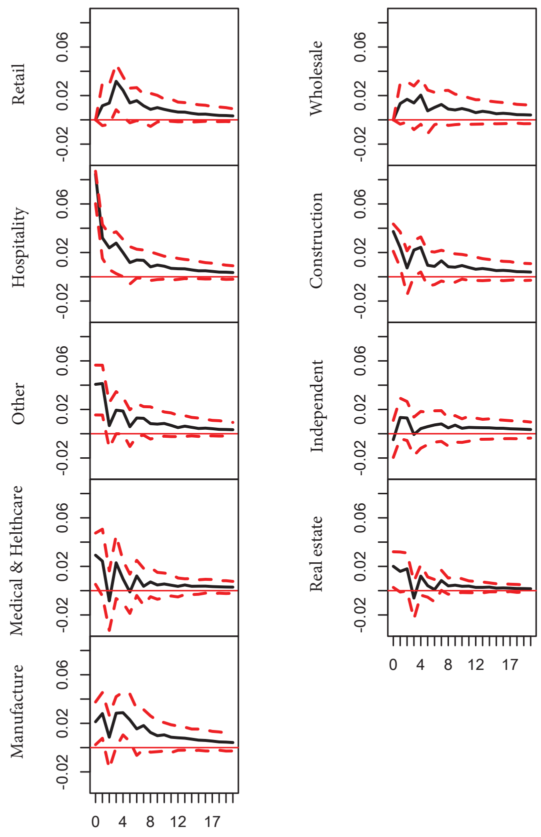


Figure 6: Impulse responses (hospitality industry).

a large impact on other industries.

In this study, we validated these results by comparing the production ripple effects between industries. As a result, the production ripple effect to be compared is related to the power of dispersion and the production inducement coefficient in Okinawa Prefecture. The power of dispersion is an index that quantifies the production ripple effect on an entire sector when one unit of final demand occurs in one sector. In contrast, the production inducement coefficient is an index that quantifies the production ripple effect on each sector when one unit of final demand occurs in all sectors.

In Figure 5, the production inducement coefficients of industries in Okinawa Prefecture are listed in descending order of magnitude as follows: commercial (retail and wholesale businesses) > construction industry > hospitality > manufacturing > medical care, health care, social security, nursing care > real estate. We analyzed this order in conjunction with the analysis results of the changes in business confidence. The top three industries (commercial, construction, and hospitality) correspond to the industries located upstream in Figure 5, and they appear as major fluctuation factors for other industries in the impulse response analysis. This was also confirmed by the FEVD results. A large power of dispersion has the effect of promoting production for other industries. The improvement in business confidence in these industries stimulates production and enhances business con-

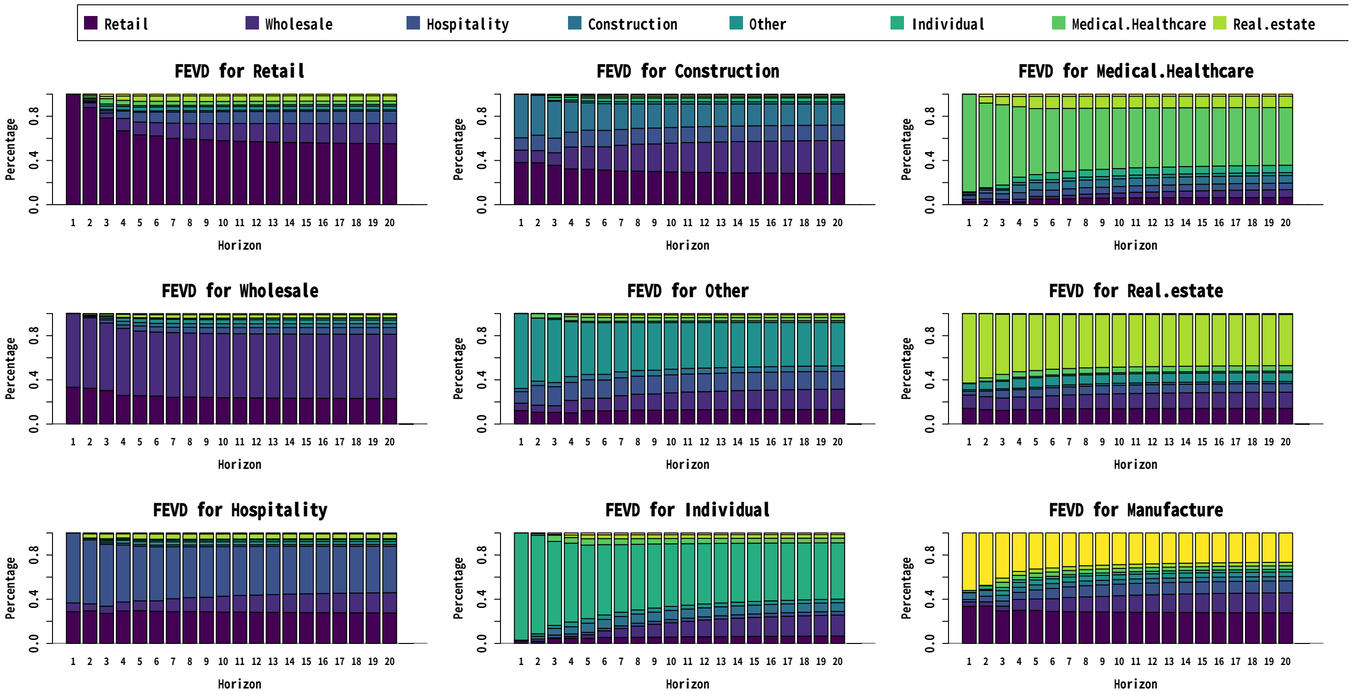


Figure 7: Results of forecast error variance decomposition (FEVD).

fidence in other industries. This propagation is reasonable, as industries that cause changes in business confidence in other industries correspond to industries with a high production ripple effect.

Next, we analyzed relationships using the production inducement coefficient. Six industries are listed in descending order of the production inducement coefficient as follows: hospitality > manufacturing > commercial (retail and wholesale) > real estate > construction industry > medical care, health care, social security, nursing care. The results of the partial Granger causal test demonstrate that the top two industries are influenced by a large number of industries. Figure 5 reveals that manufacturing is significantly affected by three industries, while hospitality is significantly affected by two industries. When the analysis is correlated with the results of FEVD presented in Figure 7, it is difficult to explain any unpredictable changes in the hospitality and manufacturing industries owing to the fact that both the sensitivity coefficient and contribution from other industries are high. In other words, the hospitality and manufacturing industries are strongly affected by production activities in other industries and are highly sensitive to economic fluctuations in those industries.

5 Conclusion

In this study, we proposed a framework for generating a business confidence index and analyzing the inter-industry structure of a local area, using text data representing the characteristics of the economy of the local area. We demonstrated that the business confidence index generated using the contact history owned by a local bank can reproduce the existing index with high accuracy. In addition, unlike the existing index, the business confidence index in a local area can be obtained more frequently than other text sources. Furthermore, because category classification was performed for the contact history owned by the local bank in question, it was possible to obtain the business confidence for each industry category.

In this study, we used Granger causal analysis, impulse response function analysis, and variance decomposition methods to analyze the causality of different time series data from the obtained business confidence index for each industry category. The results revealed the changes in business confidence among industries, the effect of the business confidence index on each industry category, and the contribution of each industry category to other industries.

In future work, we plan to use other forms of data to generate business confidence indices. In addition, we intend to adapt our framework to data from another local bank as a means to better evaluate our framework and expand our understanding of business confidence indices.

References

[Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[Granger, 1969] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[Graves and Schmidhuber, 2005] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[Guo *et al.*, 2008] S. Guo, A.K. Seth, K.M. Kendrick, C. Zhou, and J. Feng. Partial granger causality—eliminating exogenous inputs and latent variables. *Journal of Neuroscience Methods*, 172(1):79–93, 2008.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Ito *et al.*, 2018] Tomoki Ito, Hiroki Sakaji, Kiyoshi Izumi, Kota Tsubouchi, and Tatsuo Yamashita. Ginn: gradient interpretable neural networks for visualizing financial texts. *International Journal of Data Science and Analytics*, Dec 2018.

[Koppel and Shtrimberg, 2006] Moshe Koppel and Itai Shtrimberg. *Good News or Bad News? Let the Market Decide*, pages 297–301. Springer Netherlands, Dordrecht, 2006.

[Milea *et al.*, 2010] Viorel Milea, Nurfadhlina Mohd Sharef, Rui Jorge Almeida, Uzay Kaymak, and Flavius Frasincar. Prediction of the msci euro index based on fuzzy grammar fragments extracted from european central bank statements. In *2010 International Conference of Soft Computing and Pattern Recognition*, pages 231–236, Dec 2010.

[Sakai and Masuyama, 2007] Hiroyuki Sakai and Shigeru Masuyama. Extraction of cause information from newspaper articles concerning business performance. In *Proc. of the 4th IFIP Conference on Artificial Intelligence Applications & Innovations*, pages 205–212, 2007.

[Sakaji *et al.*, 2008] Hiroki Sakaji, Hiroyuki Sakai, and Shigeru Masuyama. Automatic extraction of basis expressions that indicate economic trends. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 977–984, 2008.

[Schumaker and Chen, 2009] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2):12:1–12:19, March 2009.

[Sims, 1980] Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980.