# Evaluation of Similarity Analysis of Newspaper Article Using Natural Language Processing

**Ayako Ohshiro [1†], Takeo Okazaki [2††], Takashi Kano[3†††], and Shinichiro Ueda[4††††]**
[1†]Department of Business Administration, Okinawa International University,
Ginowan, 0kinawa 901-2701, Japan
[2††]Faculty of Engineering, University of the Ryukyus, Okinawa 903-0213, Japan
[3†††]Graduate School of Economics Hitotsubashi University, Naka, Kunitachi Tokyo, Japan
[4††††]Department of Clinical Research and Quality Management Graduate School of Medicine University of the Ryukyus
Nishihara, 0kinawa 903-0215, Japan

**Abstract**
Comparing text features involves evaluating the ”similarity” between texts. It is crucial to use appropriate similarity measures when comparing similarities. This study utilized various techniques to assess the similarities between newspaper articles, including deep learning and a previously proposed method: a combination of Pointwise Mutual Information (PMI) and Word Pair Matching (WPM), denoted as PMI+WPM. For performance comparison, law data from medical research in Japan were utilized as validation data in evaluating the PMI+WPM method. The distribution of similarities in text data varies depending on the evaluation technique and genre, as revealed by the comparative analysis. For newspaper data, non-deep learning methods demonstrated better similarity evaluation accuracy than deep learning methods. Additionally, evaluating similarities in law data is more challenging than in newspaper articles. Despite deep learning being the prevalent method for evaluating textual similarities, this study demonstrates that non-deep learning methods can be effective regarding Japanese-based texts.

***Keywords:***
*Pointwise Mutual Information, Simpson coefficient, Doc2vec, BERT, Newspaper.*

## 1. Introduction

The development of computer functions has facilitated access to previously inaccessible data, thereby expanding the possibilities of data analysis and enabling diverse perspectives in interpreting data that differ from traditional approaches. For instance, by quantitatively analyzing time-series text data from newspapers that document changes in societal conditions over time, researchers have visualized the daily evolution of topics [1] and compared article characteristics across different newspapers [2]. Comparing text features involves evaluating the “similarity” between texts. It is crucial to use appropriate similarity measures when comparing similarities. In natural language processing, the extraction of structured or similar documents [3] has been crucial in comprehending and managing them more efficiently. For example, local government ordinances have similar content, but their differences across regions are compared manually by professionals [4]. Adopting a computer-based language processing approach is expected to reduce working costs for these tasks [5][6][7][8]. A study has been conducted that applies text mining methods directly to law documents. The study aims to interpret law documents using word frequency as an index [9]. Recently, comparative studies for law documents using deep learning algorithms such as a document vector feature from BERT [10] and Doc2vec [11] have increased. Our study has thus far concentrated on clinical research laws including, three guidelines, ministerial ordinances, and the law regarding clinical research. Visualizing relationships can help clinical researchers busy with clinical practice to understand them. Our research has successfully predicted similarities between these laws and visualized them. First, the study of the possibility of interpreting the relationship between laws related to clinical research using word2vec and topic model has been considered [12][13]. In addition, each law was analyzed by generating a co-occurrence network of included words. The network was used to interpret the law at the word level, by identifying pairs of words with strong co-occurrence relationships, commonly occurring words, and comparing their co-occurring words [14]. Furthermore, the study predicted the similarity between three laws using ”word-matching” under the hypothesis that sentences with more common words are more similar [15]. The verification indicated that while the method correctly identified the ”high similarity/match” combination, it also incorrectly evaluated some pairs as ”low similarity/mismatch,” presenting issues to be addressed. Subsequently, a new scale for predicting similarity was proposed by [16], which combines ”Word Group Matching” with ”Pointwise Mutual Information” to evaluate the similarity of laws based on the probability of co-occurrence of multiple words. The evaluation experiment result suggests that this approach has the potential to improve accuracy compared to conventional methods depending on

Manuscript received June 5, 2024
Manuscript revised June 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.6.1

the combination used. In cases where a similarity evaluation method is constructed based on the text data used in the experiment, it becomes necessary to validate the method for generalization by applying it to different text data and methods. For instance, comparing the accuracy of similarity evaluations using text data from different genres can further assess the property of the proposed method and lead to measures suited to specific text genres. Therefore, this study aims to evaluate the similarity of newspaper articles using multiple similarity evaluation methods in natural language processing, including the previously proposed PMI+WPM. The structure of this study is as follows: Section 2 describes the similarity evaluation methods used, and Section 3 outlines the experimental setup. Section 4 presents the experimental data, the results, and discussion, and Section 5 concludes the study.

## 2. Similarity Evaluation Method

The objective of this study is to evaluate the similarities between newspaper articles, as depicted in Figure 1.

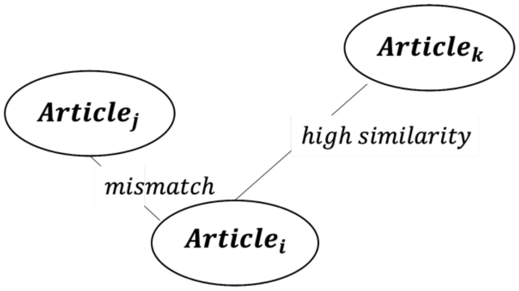


Fig 1: Similarity between articles (example)

This section describes the natural language processing methods and evaluation index used for assessing similarity and the experimental procedures.

### 2-1. Similarity Evaluation Methods without Deep Learning

This section describes the similarity evaluation methods that utilize set and information theories rather than deep learning. If a specific combination of words appears frequently, it infers a high similarity between the documents. When two words x and y appear within a single text, their Pointwise Mutual Information (PMI), PMI(x,y), can be expressed as indicated in Equation (1).

$$PMI(x,y) = \log\frac{P(x,y)}{P(x)P(y)} \qquad (1)$$

Furthermore, if the combinations of a specific word group have a high degree of matching, the documents are considered highly similar. Considering the proportional relationship between the occurrence frequency of word groups and the length of the document, it is essential to consider the ratio of the total number of occurrences of word groups represented as $F\{x,y\}$ when determining the similarity between documents based on the degree of match of word groups. In this study, the degree of word group matching represented as $w_{x,y}$ in texts $A$ and $B$ is denoted as $\mathrm{WGM}\{w_a(x,y), w_b(x,y)\}$, and the similarity between the two documents using the total number of occurrences of word groups in the compared texts, $F\{x,y\}$, is defined as PMI + WGM according to Equation (2).

$$\frac{P(x,y) + \mathrm{WGM}\{w_a(x,y), w_b(x,y)\}}{F(x,y)} \qquad (2)$$

When the words in documents A and B are considered sets A and B, the similarity between documents A and B is calculated as in (3) using the Simpson coefficient.

$$\frac{|A \cap B|}{min(|A|,|B|)} \qquad (3)$$

Despite the availability of related coefficients such as Dice and Jaccard, the Simpson coefficient was selected for comparison in this study due to its superior performance in preliminary experiments.

### 2-2. Similarity Evaluation Methods with Neural Network

In the Doc2vec algorithm, text features are converted vector and calculated similarities with unsupervised learning. It was developed after Word2Vec with a neural network that generates hundreds of dimensional vectors for each word based on surrounding words and numerical relationships between words. By extending word-distributed representation to Paragraph Vector [17], document similarity could be calculated using a document-distributed representation.

The BERT algorithm can accurately infer from the contextual information of sentences. The accuracy of word prediction based on context, which was previously challenging to improve, has been enhanced by combining bidirectional learning with conventional unidirectional learning and completing fill-in-the-blank questions. Consequently, the study aims to correlate similar law documents [10][11].

## 3. Experiment

This section outlines the procedure for determining article similarity in newspapers using the previously mentioned evaluation methods.

The rank correlation coefficient [0,1] is utilized as preprepared benchmark data to evaluate the performance of the calculated similarities. The benchmark data were collected by soliciting evaluations from economics experts on the similarity between newspaper articles, using a four-level ordinal scale of "match/high similarity/low similarity/mismatch," as presented in Table 1.

Table 1: Benchmark data for similarity between articles (EXAMPLE)

| $\boldsymbol{Article_i}$ | $\boldsymbol{Article_j}$ | $\boldsymbol{Similarity}$ |
|---|---|---|
| $\boldsymbol{Article_1}$ | $\boldsymbol{Article_2}$ | mismatch |
| $\boldsymbol{Article_1}$ | $\boldsymbol{Article_3}$ | match |
| : | : | : |
| $\boldsymbol{Article_{10}}$ | $\boldsymbol{Article_5}$ | low similarity |

Furthermore, to calculate the rank correlation coefficient with the benchmark data, the similarity scores obtained using the similarity evaluation methods must be converted into ordinal scales. In this study, the range of each similarity score was evenly divided into four levels, and, similar to the benchmark data, the scores were converted into "identical/high similarity/low similarity/no similarity" based on their rank of similarity.

The procedure for evaluating the similarity of newspaper articles is outlined below and depicted in Figure 2.

Step1 Obtain benchmark data by requesting an expert in economics to rate each combination of newspaper articles on a four-level ordinal scale of "match/high similarity/low similarity/mismatch."

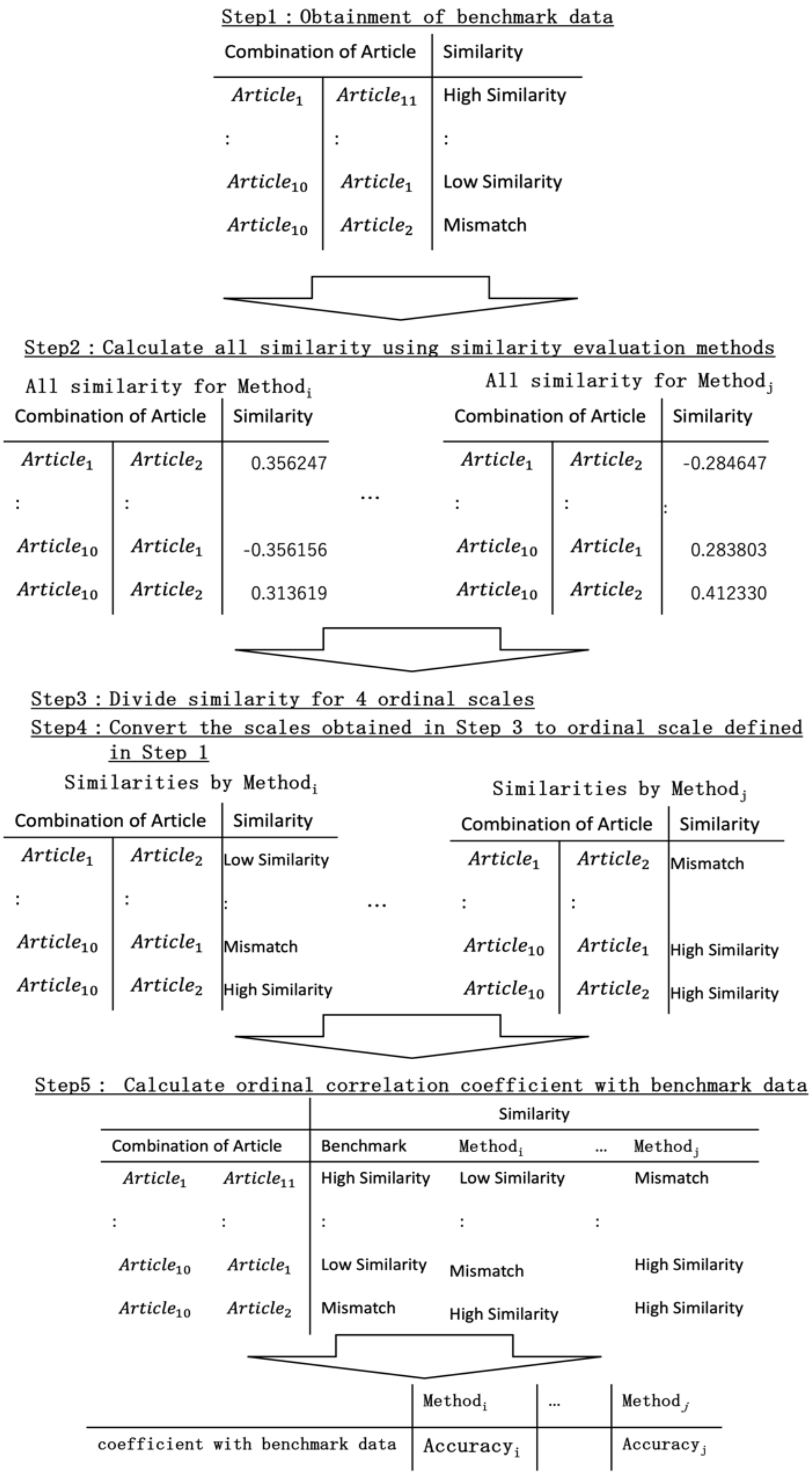
Step2 Calculate the similarity scores for all possible combinations of the target text data.

Step3 Divide the distribution of similarity scores obtained in Step 2 into quarters, with the range defined as (maximum value - minimum value) / 4.

Step4 Convert the quartile measures derived in Step 3 into the four-tiered ordinal scale defined in Step 1.

Step5 Calculate the rank correlation coefficients between each quartile measure and the benchmark data to determine the estimation accuracy.

Step 1 corresponds to the procedure for generating benchmark data, Steps 2 to 4 correspond to the procedure for generating test data using the similarity assessment scale, and Step 5 corresponds to the procedure for evaluating the performance of the test data.



Step1：Obtainment of benchmark data

| Combination of Article | | Similarity |
|---|---|---|
| $Article_1$ | $Article_{11}$ | High Similarity |
| : | : | : |
| $Article_{10}$ | $Article_1$ | Low Similarity |
| $Article_{10}$ | $Article_2$ | Mismatch |

Step2：Calculate all similarity using similarity evaluation methods

All similarity for $\mathrm{Method}_i$

| Combination of Article | | Similarity |
|---|---|---|
| $Article_1$ | $Article_2$ | 0.356247 |
| : | : | |
| $Article_{10}$ | $Article_1$ | -0.356156 |
| $Article_{10}$ | $Article_2$ | 0.313619 |

…

All similarity for $\mathrm{Method}_j$

| Combination of Article | | Similarity |
|---|---|---|
| $Article_1$ | $Article_2$ | -0.284647 |
| : | : | : |
| $Article_{10}$ | $Article_1$ | 0.283803 |
| $Article_{10}$ | $Article_2$ | 0.412330 |

Step3：Divide similarity for 4 ordinal scales

Step4：Convert the scales obtained in Step 3 to ordinal scale defined in Step 1

Similarities by $\mathrm{Method}_i$

| Combination of Article | | Similarity |
|---|---|---|
| $Article_1$ | $Article_2$ | Low Similarity |
| : | : | : |
| $Article_{10}$ | $Article_1$ | Mismatch |
| $Article_{10}$ | $Article_2$ | High Similarity |

…

Similarities by $\mathrm{Method}_j$

| Combination of Article | | Similarity |
|---|---|---|
| $Article_1$ | $Article_2$ | Mismatch |
| : | : | |
| $Article_{10}$ | $Article_1$ | High Similarity |
| $Article_{10}$ | $Article_2$ | High Similarity |

Step5： Calculate ordinal correlation coefficient with benchmark data

| | | Similarity | | | |
|---|---|---|---|---|---|
| Combination of Article | | Benchmark | $\mathrm{Method}_i$ | … | $\mathrm{Method}_j$ |
| $Article_1$ | $Article_{11}$ | High Similarity | Low Similarity | | Mismatch |
| : | : | : | : | : | |
| $Article_{10}$ | $Article_1$ | Low Similarity | Mismatch | | High Similarity |
| $Article_{10}$ | $Article_2$ | Mismatch | High Similarity | | High Similarity |

| | $\mathrm{Method}_i$ | … | $\mathrm{Method}_j$ |
|---|---|---|---|
| coefficient with benchmark data | $\mathrm{Accuracy}_i$ | | $\mathrm{Accuracy}_j$ |

Fig 2: **Procedure of investigation (Steps 1- Step4)**

Table 2: Distribution of similarity with each method applied to each data

| | Law | | | | | | | | | Newspaper | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1・L2 | | | L2・L3 | | | L1・L3 | | | | | |
| **Method** | **Max** | **Min** | **SD** | **Max** | **Min** | **SD** | **Max** | **Min** | **SD** | **Max** | **Min** | **SD** |
| WGM+PMI | 0.707 | 0 | 0.009 | 0.638 | 0 | 0.150 | 0.544 | 0 | 0.016 | 0.825 | 0.002 | 0.077 |
| Simpson's Coefficient | 0.783 | 0 | 0.124 | 0.270 | 0 | 0.034 | 0.199 | 0 | 0.039 | 0.973 | 0.014 | 0.150 |
| Doc2vec | 0.522 | -0.356 | 0.621 | 0.450 | -0.436 | 0.627 | 0.476 | -0.466 | 0.666 | 1.000 | -0.355 | 0.216 |
| BERT | 0.728 | 0.561 | 0.032 | 0.755 | 0.556 | 0.032 | 0.769 | 0.538 | 0.033 | 0.829 | 0.556 | 0.033 |

Table 3: Ordinal correlation coefficient between similarity calculated by each method for each data and benchmark data

| **Method** | Law | | | Newspaper |
|---|---|---|---|---|
| | L1・L2 | L2・L3 | L1・L3 | |
| WGM+PMI | 0.454 | 0.320 | 0.205 | **0.627** |
| Simpson's Coefficient | 0.311 | 0.355 | 0.218 | **0.662** |
| Doc2vec | -0.018 | -0.019 | 0.012 | 0.368 |
| BERT | 0.205 | **0.405** | 0.397 | 0.262 |

## 4. Results and Discussion

This section describes the data used in the experiment and the results obtained following the procedure outlined in the previous section.

### 4-1. Experimental Data

Newspaper articles widely reflect the social landscape and are read by a broad spectrum of society. In the field of language processing, these are frequently processed. They must be written in a manner that is easily comprehensible to anyone who has completed compulsory education. The aim is to convey content briefly to readers, with the main points presented at the beginning and details following later. This structure is a characteristic feature [20]. In addition to the content [21], the sentence structures and their organization have been observed to change over time. In this study, we analyzed articles related to ”economics” from the ”Ryukyu Shimpo,” a local newspaper, to ensure that our research was regionally relevant. The data covers two weeks from September 2023.

### 4-2. Discussion

In this section, accuracy is defined as the ordinal correlation coefficient between benchmark data and the proposed similarity evaluation scale. Additionally, the differences between methods and text data were explored. The similarity evaluation performance was also compared with the traditional method. To compare the differences in results depending on the text data, the results of similarity evaluation between each article of the three types of law data that is validation, data for the previously proposed method as WGM+PMI, was utilized. First, to show the similarity distribution of each method used for each text data, the maximum value (Max), the minimum value (Min), and the standard deviation (SD) of similarity are presented in Table 2.

The minimum value is 0 when using WGM + PMI or Simpson coefficient, indicating that ”the degree of similarity at the word level is 0” from the result of using law data. However, the minimum value of the BERT model exceeds 0.5, indicating that it is significantly higher than the minimum value of other methods. Additionally, the standard deviation of Doc2vec exceeds 0.6 and the variation is significantly higher than other methods, owing to the similarity of WGM+PMI, Simpson coefficient, and BERT, which all fall in the range of [0,1]. This aspect may be because Doc2vec is [-1,1]. Subsequently, when using newspaper data, the Simpson coefficient and Doc2vec exhibit maximum values exceeding 0.95, while WGM+PMI and BERT demonstrate values below 0.85. Additionally similar to the law data, the minimum values of WGM+PMI, Simpson coefficient, and Doc2vec are below 0.015, indicating a very low degree of similarity. However, the minimum value of BERT is above 0.5, which suggests a higher degree of similarity. Consequently, BERT exhibits a lower standard deviation of similarity degree compared to other methods Regardless of the method used, the maximum similarity value is higher when using newspaper data. In addition, the standard deviation is particularly low

Manuscript received June 5, 2024
Manuscript revised June 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.6.1

when using Doc2vec with newspaper data. From the result of BERT, no significant changes were observed in any of the indicators, including maximum value, minimum value, or standard deviation, depending on the data genre. Applying WGM + PMI, Simpson coefficient, and Doc2vec to newspaper data resulted in a higher maximum value. Additionally, the distribution of similarity measures was not significantly affected by differences in data genre by the case using BERT.

Subsequently, the ordinal correlation coefficient between the similarities calculated by each method and the benchmark data is presented in Table 3. This coefficient represents the accuracy of each method.

First, the similarities of the data used were compared. When analyzing law data, negative values were observed in some cases, indicating that the accuracy of Doc2Vec was significantly lower than that of the proposed method WGM + PMI, Simpson coefficient, and BERT. WGM+PMI or BERT were the preferred methods, but the best results achieved were less than 0.5. Consequently, identifying an appropriate method was challenging. However, when using newspaper data, the proposed method and the Simpson coefficient achieved an accuracy of over 0.6, while deep learning methods such as Doc2Vec and BERT resulted in an accuracy below 0.4. Classical similarity measure may be more effective than deep learning in indicating the similarity of newspaper data, as demonstrated by the results. Next, we compare the similarities among different methods. When using WGM + PMI, Simpson coefficient, and Doc2vec, the accuracy is higher with newspaper data. When using BERT, the accuracy varies depending on the combination, but it is still higher with law data. When focusing on the characteristics of law data and newspaper article data, due to the limited vocabulary and semantics of law data, BERT, which is bidirectional is more suitable for similarity evaluation than WGM+PMI and Doc2vec, which predict specific words from surrounding words. Newspaper article data are expected to exhibit a higher frequency of relevant words in proximity to the target word than law data. Thus, the use of similarity measures such as WGM+PMI and the Simpson coefficient, which are based on word co-occurrence probabilities and sets, is more appropriate.

Consequently, this demonstrates the potential for selecting appropriate mining methods tailored to the characteristics of each genre within textual data.

## 5. Conclusion

In this study, the similarities between newspaper articles were evaluated using various similarity evaluation methods in natural language processing. These methods included approaches based on deep learning, incorporating set theory and information theory. For performance comparison, in addition to newspaper article data, law document data that had been used to verify the previously proposed WGM+PMI were also utilized. Results of the similarity distribution comparison for all data combinations indicate that the similarity distribution varies with the text data and the methods used. Furthermore, even when using the same method, the accuracy of similarity evaluation performance varies depending on the genre of text data handled.

Many similarity evaluation methods have been proposed in the field of language processing. However, this study demonstrates the potential for better similarity evaluation by selecting methods appropriate to the genre of text data being analyzed.

## References

[1] James Allan, Rahul Gupta, and Vikas Khandelwa, ” Temporal Summaries of News Topics.”, In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001.

[2] Yusuke Hoshino, “A Preliminary Analysis of Newspaper Editorials on COVID-19 Using Natural Language Processing Technologies : Differences among Newspapers and Further Research”, Musashino University Management Journal (5), 113-148, 2022

[3] Larry M. Manevitz, Malik Yousef, ”One-Class SVMs for Document Classification”, Journal of Machine Learning Research 2 ,139-154, 2001

[4] Fujii, Machiko, “The Present Condition and Issues of Municipal Ordinances in Merger of Nuncipalities : Case of the Ordinance Making of Koka City ”, The bulletin of the Graduate School of Law, Ryukoku University, 181-214, 2007

[5] KAKUTA Tokuyasu, “An analysis of regulations of local governments using a supercomputer and the application to a regulation database”, Nagoya University Journal of Law and Politics, Vol. 246, 69-91, 2012

[6] TAKENAKA YOICHI, WAKAO TAKESHI, “Automatic Generation of Article Correspondence Tables for the Comparison of Local Government Statutes”, Journal of natural language processing, Vol. 19 No. 3, pp.194- 212, 2012.

[7] Gaitake.K, Tomoya.S, Youiti.T, “Meizi minpou seitei zi ni okeru nitihutu minpou zyoubun no sansyou kankei sai suitei ” (in Japanese) The 25th Annual Meeting of the Association for Natural Language Processing, pp.398-401, 2019.

[8] Gaitake.K, Tomoya.S, Youiti.T, ”Meizi minpou to kakukokumin hou to no zyoubun ruizi kankei ni motozuku rikkyakuten no kaiseki” (in Japanese), The 26th Annual Meeting of the Association for Natural Language Processing, pp93-96, 2020

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,” Bidirectional Encoder Representations from Transformers”, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186, 2019

[10] Quoc V. Le, Tomas Mikolov, ”Distributed Representations of Sentences and Documents”, proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196, 2014

[11] Ayako OHSHIRO, Shinichiro UEDA. “ Feature extraction of each laws for clinical research and their relation“ , Institute of Electronics, Information and Communication Engineers. 2019; D-5-4

[12] Ayako OHSHIRO, Shinichiro UEDA. ”Interpretability of laws related to clinical research with text mining.”, THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS. 2019 ; NLC2019-33 (320)NLC2019-33:35-40

[13] Ayako OHSHIRO, Takeo OKAZAKI, Shinichiro UEDA. “Visualization of clinical research-related laws using co-occurrence network”, The Japanese Society of Clinical Pharmacology and Therapeutics (JSCPT).2024; 55(1): 57-62.

[14] Ayako OHSHIRO, Takeo OKAZAKI, Shinichiro UEDA. “ Study on relationship visualization of clinical research-related laws using word- matching“ , The Japanese Society of Clinical Pharmacology and Therapeutics (JSCPT). 2023 ; 54(1): 43–48.

[15] Ayako OHSHIRO, Takeo OKAZAKI, Shinichiro UEDA. ziko sougo zyouhou ryou to tango gun itti do wo ku mi a wase ta rinsyou kenkyuu kanren hourei no ruizisei hyouka no kentou (in Japanese). The 29th Annual Meeting of the Association for Natural Language Processing, pp1216-1219, 2023

[16] Mikolov, T.; Le, Q. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning, ICML 2014, 2014, p.1188-1196.

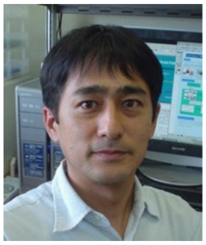
[17] Hiroki T, Makoto N, ”BERT wo moti i ta hikakuhou kenkyuu ni okeru ruizi zyoukou no taiou zu ke” (in Japanese), The 28th Annual Meeting of the Association for Natural Language Processing, pp948-951, 2022

[18] Kaito Koyama, Tomoya Sano, Yoichi, ”The legislative study on Meiji civil code by machine learning”, Fifteenth International Workshop on Juris-informatics (JURISIN 2021)

[19] Sinryou.H, ”「wakariyasusa」 wo mezasi te ka ka re ta sinbun kizi no buntai teki tokutyou syakai gengogaku (in Japanese)”, The Japanese journal of language in society, pp.43-54, vol.15(2015).

[20] Yuta Ichikawa, “PRELIMINARY STUDY ON DETECTION OF NEWSPAPER TREND AMONG THEIR PUBLISHERS USING TEXT-MINING APPROACH”, Bulletin of graduate studies. Engineering Hosei University, Vol.57(2016)

**Ayako Ohshiro** received the B.S. and M.S. degrees from University of the Ryukyus in 2009 and 2011, respectively. She took Ph.D. from University of Ryukyus in 2017 and she has been an associate Professor at Okinawa International University. Her research interests are data analysis at the field of Intelligent information engineering.



**Takeo Okazaki** received B.Sc. and M.Sc. degrees from Kyushu University in 1987 and 1989, respectively. He earned his Ph.D. from University of the Ryukyus in 2014. He is currently a professor at the University of the Ryukyus. His research interests are statistical data normalization for analysis, statistical analysis, data analysis, genome informatics, tourism informatics, geographic information systems, and data science. He is a member of JSCS, IEICE, JSS, GISA, and BSJ Japan.

**Takashi Kano** received his Ph.D. from the University of British Columbia in 2003 and has been a professor at Hitotsubashi University since 2016. His research interests include macroeconomics, international finance, and applied econometrics.

**Shinichiro Ueda** received the M.D in 1985 and took Ph.D. from Yokohama City University in 1997 and he has been professor at University of the Ryukyus. His research interests are Clinical Pharmacology and Therapeutics, and Clinical Research education.