

# Text Mining for Factor Modeling of Japanese Stock Performance

K. ISHIZUKA

Department of Industrial Engineering and Economics  
Tokyo Institute of Technology  
Tokyo, Japan  
e-mail: ishizuka.k.ad@m.titech.ac.jp

K. NAKATA

Department of Industrial Engineering and Economics  
Tokyo Institute of Technology  
Tokyo, Japan  
e-mail: nakata.k.ac@m.titech.ac.jp

**Abstract**—Recent advances in natural language processing (NLP) have made a significant breakthrough in various NLP tasks. Word embeddings and pre-trained language models can convert unstructured textual data to computable numerical vectors, containing meanings and contextual information. However, most public language models are only for English, and therefore, financial applications are also in English. Few Japanese studies a direct approach, which predicts not attribute of a text but real-world information. This study aims to build a factor model based on Corporate Annual Securities Reports' textual data for the Japanese stock market and investigate its effectiveness. We found that textual data complement financial numerical data and improve the p-value of the Gibbons-Ross-Shanken Test, reducing the factor model's anomaly. Furthermore, BERT [1] is much better than word2vec [2] and LDA [3] for this factor modeling.

**Keywords**—component; text mining, factor model, natural language processing, stock market

## I. INTRODUCTION

In recent years, with the rapid progress in natural language processing (NLP), word embeddings and pre-trained language models allow converting unstructured textual data to computable numerical vectors. An example of former is word2vec [2] and that of the latter is BERT [1]. These vectors contain word meanings and contextual information and can be used or fine-tuned for downstream tasks. In financial applications using annual reports, a method for predicting bankruptcy from the text data of Form 10-K (annual reports) has been proposed in the U.S. [4], a method for discriminating causal statements from Corporate Annual Securities Reports in Japan [5]. Although the former study predicts real events, the latter discriminates an attribute of a sentence from the text. Few Japanese financial studies adopt a direct approach, which predicts not attribute of a text but real-world information. Reference [6] analyzes Japanese stocks using classical NLP methods built a factor model from postings on stock bulletin boards. Portfolio optimization requires the estimation of returns and risks. The former can be expressed using a real-valued vector, the mean vector, and the latter can be expressed using a positive definite matrix, the covariance matrix. In portfolio optimization, the ratio of investments in each stock is calculated by solving an optimization problem with the additional risk aversion input. The factors are generated using characteristics from market information and financial

information, such as traded volume, return on equity, momentum, price to earnings, and market capitalization [7]. A Corporate Annual Securities Report describes, for each business year, the status of the accounts, essential matters concerning the business, and other matters necessary for the protection of investors, as required by the Financial Instruments and Exchange Law in Japan. These reports include not only quantitative information, such as "Overview of business results," but also more qualitative information, such as "Issues to be Addressed" and "Risks of the Business." In contrast, textual information from other sources, such as news and stock bulletin boards, is limited to only a few companies and has a bias in information frequency. At the same time, Corporate Annual Securities Reports are disclosed by all listed companies for each fiscal year, which enables data analysis without sampling bias. To make the most of annual Corporate Annual Securities Reports for portfolio optimization, we developed a factor model based on the "Business and Other Risks" in Corporate Annual Securities Reports as a factor complementary to financial market data and verified its effectiveness. The factor is generated for each company using their 2014 Annual Securities Reports. The model with the proposed factors was then verified by the Gibbons-Ross-Shanken Test (hereafter referred to as the GRS test).

## II. RELATED WORK

In the U.S., a method for predicting bankruptcy from text data of Form 10-K (an annual report) has been proposed [4], and in Japan, a method for determining causal relationships from Corporate Annual Securities Reports has been proposed [5]. The former predicts bankruptcy from textual, financial and accounting data, while the latter discriminates text attributes (causality) from the text. Reference [6] analyzes Japanese stocks using classical NLP methods built a factor model from postings on stock bulletin boards. This study generates ten factors by summing the term frequency-inverse document frequency values of each post, embeddings created, and principal component analysis performed.

## III. BACKGROUND

### A. Formulation of the Factor Model

A factor model is a linear model where the common cross-sectional variables expressed in the form of (1) explain the returns:

$$\begin{pmatrix} y_{1,t} \\ \vdots \\ y_{N,t} \end{pmatrix} = \begin{pmatrix} b_{1,1} & \cdots & b_{1,K} \\ \vdots & \ddots & \vdots \\ b_{N,1} & \cdots & b_{N,K} \end{pmatrix} \begin{pmatrix} f_{1,t} \\ \vdots \\ f_{K,t} \end{pmatrix} + \begin{pmatrix} u_{1,t} \\ \vdots \\ u_{N,t} \end{pmatrix} \quad (1)$$

where  $y_{i,t}$  is the excess rate of return of stock  $i$  in period  $t$ ,  $f_{k,t}$  is the excess rate of return of factor  $k$  in period  $t$ ,  $\beta_{i,k}$  is the coefficient of stock  $i$  on factor  $k$  (factor loading), and  $u_{i,t}$  is the disturbance term of stock  $i$  in period  $t$  (speculative return).

### B. Types of Factor Models

Factor models of the securities market are classified into macroeconomic models, fundamental models, and statistical models [8]. All models are linear models, but the explanatory variables and the type of regression used (e.g., time-series or cross-section) are different. Table 1 gives an overview of the factor models.

The macroeconomic model uses macroeconomic variables, such as long-term interest rates and inflation rates, as explanatory variables, and regressions with time series are used. The output of the model is each factor loading (factor-beta). To estimate the model, we regress the macroeconomic variables for each security, with the rate of return on the security collected over a period of time as the objective variable. However, the coefficients (betas) are assumed to be constant over the period, and long-term data are needed for stable estimation.

In the fundamental model, factor loadings such as market capitalization, pay-out ratio, and price book-value ratio (PBR) are considered to be attributes of each firm. They are estimated by cross-sectional regression. In other words, we conduct cross-sectional regression for each characteristic with a given return period as the objective variable. For example, if the return rate on a given day is higher for firms with lower PBRs, the corresponding value factor is estimated to have been positive if there are no other confounders. Because this model does not estimate coefficients but gives them exogenously, it is possible to use more factors than those in the factor model with time-series regression. In practice, features are not used as coefficients as they are, but the so-called Gauss Rank transformation is often applied in the cross-section direction to give robustness to outliers by transforming the empirical distribution into a normal distribution and maintaining the order of the features.

In statistical models, statistical factors are extracted by statistical methods, such as maximum likelihood and principal component analysis, on the return series and regressed as the objective variable. As in the macroeconomic model, the coefficients are assumed to be constant over the period, and long-term data are required to perform regressions on each firm with a time series.

### C. Factor-Simplified Estimation of the Fundamental Factors Model

The factor's value in a fundamental factor model is not included in the data and must be determined by some method. In the present method, the standard approach is to find the factor from the factor loading as follows. First, the stocks are divided into groups based on the factor loading size. Then,

the return of one unit long in the upper group and one unit short in the lower group are considered as the factor value. This method has been widely used, including in the verification of the Fama-French three-factor model [9].

### D. Factor-Simplified Estimation of the Fundamental Factors Model

The three-factor model of Fama and French [9] consists of excess asset returns on three factors: SMB (Small market capital Minus Big), HML (High book to market ratio Minus Low), and MKT (market factor). MKT is calculated by the market portfolio less the risk-free rate, SMB by the returns on small market capitalization companies minus that of big companies, and HML by high book-to-market ratio less low companies.

### E. SWEM

SWEM (simple word-embedding-based model) [10] employs a method for calculating sentence vectors using only word vectors. Similar methods include those using doc2vec and bidirectional encoder representations from transformers (e.g., BERT). However, these methods require neural networks to be trained from an extensive data set to obtain sentence vectors in addition to word vectors, which poses problems of tuning and securing data. Among the four proposed SWEMs, SWEM-aver applies average pooling to the word vectors contained in a sentence.

### F. GRS test

GRS test is a statistical F-test for the hypothesis that all the alphas are zero of the multi-factor asset pricing model. The multi-factor model with  $K$  factors for  $N$  assets is expressed as

$$R_{i,t} = \alpha_i + \beta_{i,1}F_{1,t} + \cdots + \beta_{i,K}F_{K,t} + \epsilon_{i,t}, t = 1, \dots, T. \quad (2)$$

where  $R_{i,t}$  is the excess return of stock  $i$  in period  $t$  and  $F_{k,t}$  is the value of factor  $k$  in period  $t$ . In addition,  $\alpha_i$  is a constant term, which in theory is expected to have the value  $\alpha_i = 0$  for all  $i = 1, \dots, N$ . A higher  $p$ -value indicates that a model is theoretically more suitable as an asset pricing model since GRS test tests the hypothesis  $H_0: \alpha_i = 0$  for all  $i = 1, \dots, N$ .

F-test statistics of GRS test

$$\frac{T - N - K}{N} (1 + \bar{F}'\hat{\Omega}^{-1}\bar{F})^{-1} \hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha}$$

follows the F-distribution with degree of freedom  $N$  and  $T - N - K$ , where  $\bar{F} = \frac{1}{T} \sum_{t=1}^T F_t$ ,  $F_t = (F_{1,t}, \dots, F_{K,t})'$ ,  $\hat{\Omega} = \frac{1}{T} \sum_{t=1}^T (F_t - \bar{F})(F_t - \bar{F})'$ ,  $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_N)'$ ,  $\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^T \hat{\epsilon}_t \hat{\epsilon}_t'$ , and  $\hat{\epsilon}_t = (\hat{\epsilon}_{1,t}, \dots, \hat{\epsilon}_{N,t})'$ . We obtain the estimation values  $\hat{\alpha}_i$  by performing a least-square regression on the above regression equation for all assets  $i = 1, \dots, N$ .

TABLE I. TYPES OF FACTOR MODELS (ADAPTED FROM [8])

Model Type	Inputs	Estimation Method	Output
Macroeconomic	Returns, macro variables	Time series	Beta
Fundamental	Returns, features (beta)	Cross section	Fundamental factor
Statistical	Return	Cross section	Statistical factor, Beta

#### IV. PROPOSED METHOD

This study aims to improve the accuracy of the factor model by using factors extracted from Corporate Annual Securities Reports. Because returns are considered to be determined by risks in a rational market, if risks can be correctly read from the "Business and Other Risks" section of Corporate Annual Securities Reports, they are expected to explain the returns. We test the above hypothesis with a simple method for estimating fundamental factors. The factor loading required for the model is described below.

##### A. Generation of Factor Loading from Corporate Annual Securities Reports

This section describes how to generate features based on the "Business and Other Risks" section of a Corporate Annual Securities Report. Business and other risks extracted from Corporate Annual Securities Reports are text and cannot be converted directly into numbers. For this reason, factor loading is generated based on the following framework. An arbitrary method of vectorization of words and sentences can be used for generating the loading.

###### 1) Generation of sentence vectors

Because a sentence consists of multiple words, if word vectors exist, we can generate sentence vectors by aggregating them. Although there are various methods for this aggregation, in our experiments, we adopted the more straightforward and more validated method of obtaining sentence vectors by averaging word vectors (SWEM-aver [10]). Because we could not obtain a significant difference with other methods compared with SWEM-aver, we limited ourselves to this method to aggregate sentence vectors from word vectors.

###### 2) Generation of word vectors

Although there are various methods for generating word vectors, we used the typical pre-trained word2vec [2] and the Japanese version of BERT [11]. The tokenizer used in each of the pre-trained models was the same as that used in the previous study. In the word2vec technique, each word is definitively given a vector. In contrast, BERT [1] does not definitively determine a vector by looking at only individual words but instead outputs a vector for each word in the final layer, considering the whole sentence's context. The length of words is as long as possible. If the sentence's length exceeds the maximum length, we divided it into multiple vectors.

##### B. Framework for Feature Generation Based on Corporate Annual Securities Reports

The framework for feature generation from "Business and Other Risks" in Corporate Annual Securities Reports consists of the following steps:

- 1) Extract and cleanse the text data of "Business and Other Risks" from financial statements.
- 2) Use the extracted text data to find the word vectors contained in each sentence.
- 3) Use SWEM-aver to convert the word vectors of the sentences into sentence vectors.
- 4) If necessary, perform dimensional compression of sentence vectors.

##### C. Other Methods of Sentence Vector Generation

Some methods for generating text vectors, such as doc2vec and linear discriminant analysis (LDA)[3], do not aggregate word vectors. We did not disregard such methods but used LDA as a backup comparison method. As a method to generate a sentence vector from LDA, we considered all the companies in the "Business and Other Risks" section of the Corporate Annual Securities Reports as a corpus. Then, by setting the probability as a stack for each topic to which each document belongs, we generated a sentence vector.

#### V. EVALUATION AND RESULTS

We examined whether features obtained from Corporate Annual Securities Reports improve factor models' explanatory power using word2vec, BERT, and LDA. BERT recognizes the entire sentence context and performs vectorization, while word2vec and LDA look at the words only and cannot consider the global context. The measure is the p-value calculated by the GRS test. GRS test tests the validity of the multi-factor asset pricing model, which is the p-value for testing the null hypothesis that  $\alpha$  goes to zero for N stocks simultaneously.

##### A. Data and Evaluation Methods

We used a dataset [12] containing textual, financial, and stock price data from Corporate Annual Securities Reports. Table II shows the amount of data for each year. We selected the 1,298 firms with no missing data for all periods. Daily stock price data for 1,202 days from December 1-st, 2014, to October 31-th, 2019, were added to form the final dataset. We did not separate the data into training, testing, and validation datasets in this verification because there are no hyperparameters. In other words, we applied the model to the entire period and performed statistical tests. B. Results Table III shows the GRS test results for 16 portfolios with four

market capitalization levels and four levels of book-to-market ratios for factors of 5, 10, and 30. The baseline for comparison is the capital asset pricing model with only the market factor and the Fama-French model with the three factors of the market, size, and value. The proposed factors were added to the baseline factors, and the model was constructed by increasing the number of factors in the baseline. The results for the cases of word2vec, BERT, and LDA are shown in Fig. 1, 2, and 3, respectively. The number of factors was adjusted by performing dimensionality compression on the text vectors using principal component analysis. The results show that the Fama-French model (ffm) plus the factors extracted by BERT has the highest p-value. This means that it is a better asset pricing model, and the significance of the anomaly is low compared with the other models. In contrast, the word2vec (w2v) and LDA (lda) methods did not show much improvement from ffm. Each model's trend can also be read from Fig. 1–3, and it can be seen that only BERT shows bimodality. For word2vec, the p-value is unimodal for the number of factors, and for LDA, there is no relationship between the number of factors and the p-value.

TABLE II. DATA DESCRIPTION

Fiscal Year	#Reports	#Financial Data	#Stock Price Data
2014	3,724	3,583	3,595
2015	3,870	3,725	3,751
2016	4,066	3,924	3,941
2017	3,578	3,441	3,472
2018	3,513	2,893	3,413

TABLE III. COMPARISON OF P-VALUE AND ADJUSTED R<sup>2</sup> OF THE GRS TESTS

model	#factors	p-value(↑)	adj-R <sup>2</sup> (↑)
Market(mkt)	1	6.66e-16	0.803
Fama-French (ffm)	3	4.24e-06	0.915
ffm + word2vec	3 + 5	2.57e-07	0.923
ffm + BERT	3 + 5	7.99e-05	0.924
ffm + lda	3 + 5	2.84e-06	0.924
ffm + word2vec	3 + 10	6.61e-06	0.929
ffm + BERT	3 + 10	8.25e-04	0.929
ffm + lda	3 + 10	4.17e-06	0.927
ffm + word2vec	3 + 30	3.46e-05	0.941
ffm + BERT	3 + 30	2.92e-04	0.940
ffm + lda	3 + 30	3.65e-05	0.937

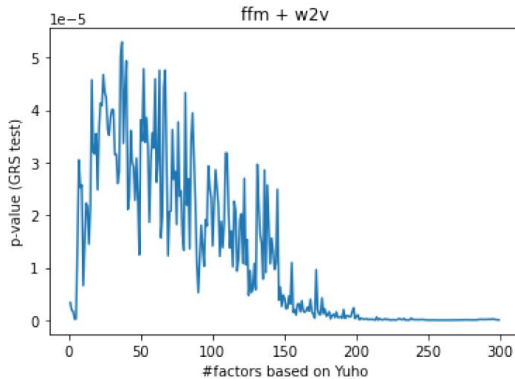


Figure 1. p-value of ffm + word2vec (w2v) and #factors.

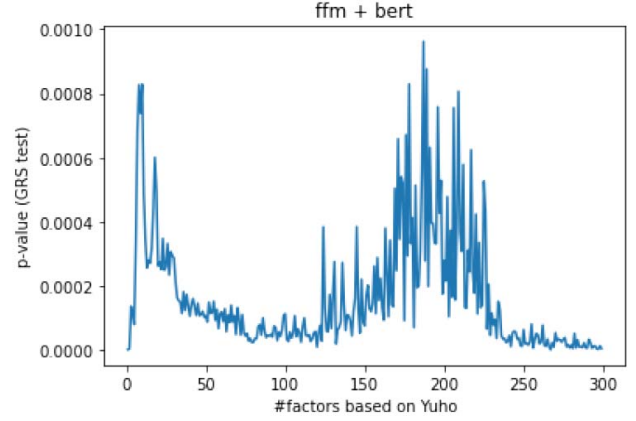


Figure 2. p-value of ffm + BERT and #factors.

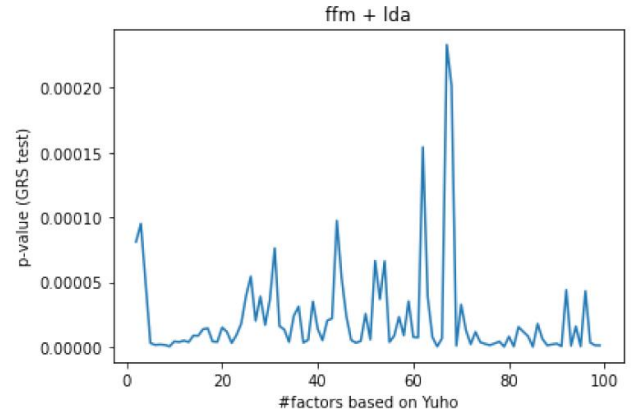


Figure 3. p-value of ffm + lda and #factors.

## VI. DISCUSSION

We used BERT to generate factors from Corporate Annual Securities Reports to improve the p-value in the GRS test over the Fama–French three-factor model. We believe that the p-value improved because the information we extracted from Corporate Annual Securities Reports using BERT included risk information. To confirm this, we replaced the word2vec vectors with random vectors and examined the relationship between the number of factors and the p-value, as shown in Fig. 4. The random vectors were sampled from the standard normal distribution with 768 dimensions. We can see that BERT and word2vec are very different from those of the random case. This suggests that information that is not included in the existing factors is described in natural language in Corporate Annual Securities Reports and can be extracted by natural language processing methods.

Also, it was found that the information extracted differed depending on the method (Fig. 5). There was a more significant difference in BERT's p-value, which considers the context, than in word2vec, which extracts the words contained in the text and ignores their order. In contrast, there was almost no difference in the decision coefficients between the models. Thus, the information in a Corporate Annual Securities Report can potentially be extracted for return prediction through machine learning with text vectors

as features. Therefore, this can be examined by training machine learning models, such as gradient boosting decision trees and deep neural networks. This point is a future problem.

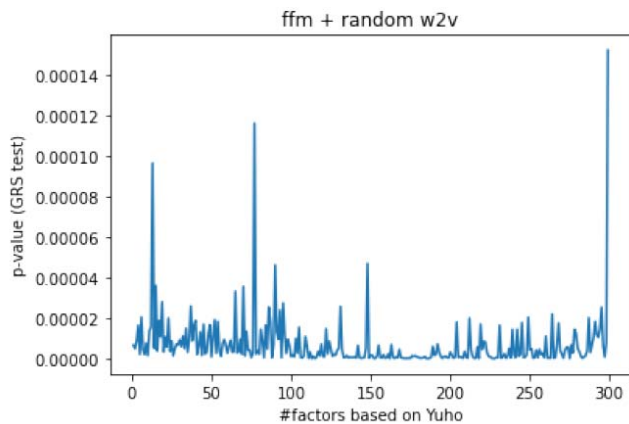


Figure 4. p-Value of ffm + random w2v and # factors.

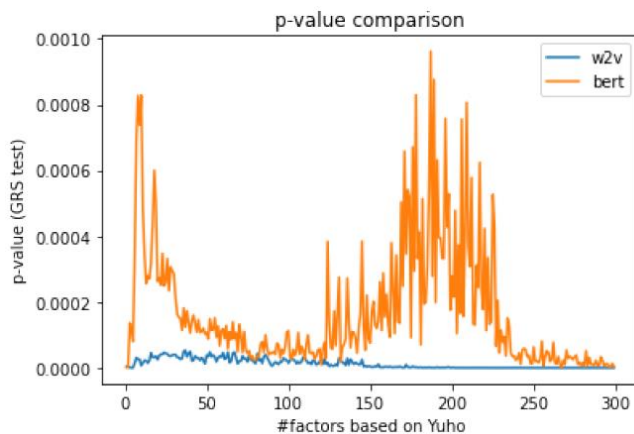


Figure 5. p-Value comparison of w2v versus BERT

## VII. CONCLUSION

This study proposed a method to create factors based on Corporate Annual Securities Reports. Based on the fact that Corporate Annual Securities Reports are readily available as information that can be used for investment decisions, we constructed factors by obtaining text embeddings from the "Business and Other Risks" section of these reports. To evaluate the extracted factors, we performed GRS tests to evaluate the asset pricing model's appropriateness. As a result, we confirmed that the p-value was improved by adding the Corporate Annual Securities Report's factors. The results suggest that the factors extracted from Corporate Annual Securities Reports have information not available in the existing factors. We also found that the results differed greatly depending on the method used to construct the factors from textual data. The improvement using BERT was

more extensive than that using word2vec or LDA. This was probably since only BERT was able to take the context into account. Furthermore, the results were much worse when the word vectors were replaced by random vectors, suggesting that natural language processing methods can play an important role in extracting information from Corporate Annual Securities Reports. However, there was no difference in the decision coefficients between models. The use of text embeddings in Corporate Annual Securities Reports in predicting the return rate is an issue for the future.

## REFERENCES

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their Figure 5. p-Value comparison of w2v versus BERT compositionality," in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 3111–3119.
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, Mar. 2003.
- [4] F. Mai, S. Tian, C. Lee, and L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures," *European Journal of Operational Research*, vol. 274, no. 2, pp. 743–758, 2019.
- [5] F. Sato, H. Sakuma, S. Koderia, Y. Tanaka, H. Sakaji, and K. Izumi, "Extraction of causal knowledge from annual securities report (Japanese)," *Proceedings of the Annual Conference of JSAI*, vol. JSAI2018, pp. 20404–20404, 2018.
- [6] T. O. Hirohiko Suwa, Eiichi Umehara, "Factor model based on contents analysis of stock bbs postings (Japanese)," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 27, no. 6, pp. 376–383, 2012.
- [7] K. N. Masaya Abe, "Deep learning for multifactor models in Japanese stock markets," *The 33rd Annual Conference of the Japanese Society for Artificial Intelligence*, vol. JSAI2019, pp. 4Rin135–4Rin135, 2019.
- [8] G. Connor, "The three types of factor models: A comparison of their explanatory power," *Financial Analysts Journal*, vol. 51, pp. 42–46, 05 1995.
- [9] H. T. Keiichi Kubota, "Fama-french re-validation of the validity of the factor model (Japanese)," *Modern Finance*, vol. 22, pp. 3–23, 2007.
- [10] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 440–450.
- [11] M. Suzuki and R. Takahashi, *cl-tohoku/bert-japanese*. GitHub, 3 2020. [Online]. Available: <https://github.com/cl-tohoku/bert-japanese>
- [12] chakki works, "Coarj: Corpus of annual reports in japan," 2019. [Online]. Available: <https://github.com/chakki-works/CoARj>