Asia-Pacific Financial Markets (2020) 27:621–632
https://doi.org/10.1007/s10690-020-09309-1

ORIGINAL RESEARCH

A Text Mining Model to Evaluate Firms' ESG Activities: An Application for Japanese Firms

Takuya Kiriu[1,2] · Masatoshi Nozaki[1]

Published online: 17 June 2020

Abstract
Environmental, social, and corporate governance (ESG) refers to the three important contributors to the sustainable growth of firms. Firms publish corporate social responsibility (CSR) reports that include quantitative and qualitative information concerning ESG activities. Although these reports are easily accessed, their qualitative information is hard to apply because manual analyses are difficult. We develop a text mining model that visualizes ESG activities from the structure of ESG-related words in CSR reports. This model quickly, effectively, and objectively facilitates processing CSR reports and comparing them with reports from peer firms. We analyze Japanese CSR reports and present an example. Further, we propose scores to evaluate the quantity and specificity of ESG activities. From the result, we obtain the following findings. First, large quantity and high specificity of ESG activities indicate a higher current ESG quantitative performance. Second, the high specificity of E-related and the large quantity of S and G-related activities portend subsequent improvement of ESG quantitative performance.

Keywords CSR report · ESG · Text mining · Visualization

JEL Classification C81

✉ Takuya Kiriu
kiriu@mtec-institute.co.jp

Masatoshi Nozaki
mnozaki@mtec-institute.co.jp

1 Mitsubishi UFJ Trust Investment Technology Institute Co., Ltd. (MTEC), Tokyo, Japan

2 Graduate School of Science and Technology, Keio University, Yokohama, Japan

1 Introduction

ESG (environmental, social, and corporate governance) refers to the three important contributors to the sustainable growth of firms. Globally, assets under management that incorporate ESG factors when making investment decisions increase by 34 percent from 2016 to 2018 and increase by 304 percent in Japan (GSIA 2018). This suggests that the importance of ESG information has been increasing.

According to KPMG (2017), 72 percent of large and mid-cap firms globally and 99 percent of those in Japan disclose ESG activities in corporate social responsibility (CSR) reports that contain both quantitative and qualitative information. Quantitative information is numerical and captures actual ESG-related performance, whereas qualitative information comprises non-numerical information such as strategies and activities for improved ESG performance. Quantitative information is more commonly used for ESG scoring because broadly quantitative datasets offered by data vendors facilitate objective comparisons. However, the evaluation of qualitative information is limited for two reasons: manually analyzing CSR reports is difficult because formats differ among firms, and manual comparisons of qualitative information tend to be subjective.

A text mining model resolves those difficulties because it processes CSR reports in a consistent manner that minimizes human subjectivity. Several studies investigate the text mining of CSR reports.[1] Some studies analyze topics presented in CSR reports by using machine learning approaches (Neural Network, Memory Based Reasoning, and Decision Tree (Tremblay et al. 2015), Naïve Bayes (Das et al. 2016), and Latent Dirichlet Allocation (Szekély and Brocke 2017)). Their goal is to identify topics in the CSR reports rather than to evaluate the reports themselves. Shahi et al. (2014) developed a scoring system for analyzing completeness in meeting requirements under the Global Reporting Initiative. The most relevant study is Liew et al. (2014). They developed a text mining model to analyze ESG activities from a qualitative perspective. To organize various topics, they specify a hierarchical structure of ESG-related words from 112 CSR reports of global processing companies. However, their method for constructing the structure is based on human qualitative judgment. There are only 18 branch nodes in their hierarchical structure. Therefore, it is challenging to apply their method for thousands of topics to make the hierarchical structure without using human judgment.

We develop a text mining model to evaluate the quality of ESG activities that visually displays ESG activities. Our model specifies a hierarchical structure of ESG-related words similar to Liew et al. (2014). Our optimization-based approach allows us to analyze and compare a large number of CSR reports quickly, objectively, and in detail. We analyze 8729 Japanese CSR reports between 1999 and 2016 and show an example of the visualized result. Although our application is in Japanese, our model serves other languages as well.

[1] As a similar study, Azhar et al. (2019) developed a text mining model for news articles to extract ESG-related information.

Further, we propose scores to evaluate qualitative information based on the quantity and specificity of ESG activities. We conduct regression analyses to investigate the relationships between the quality of ESG activities and ESG quantitative performance. Although the relationships between ESG information and firm performance have been widely studied,[2] to the best of our knowledge, our study is the only study to date that investigates the relationships between ESG qualitative information and ESG quantitative performance. From our study, we obtain the following findings. First, large quantity and high specificity of ESG activities indicate a higher current ESG quantitative performance. Second, the high specificity of E-related and the large quantity of S and G-related activities portend subsequent improvement of ESG quantitative performance.

This paper proceeds as follows. Section 2 explains our visualization model and presents an example. Section 3 proposes a score to evaluate the quality of ESG activities and examines the relationships between the score and ESG quantitative performance. The final section concludes the research and suggests directions for future studies.

2 Visualization Model

Developing the model entails five steps: preprocessing, word embedding, word classification, word structuring, and visualization. This section explains each step.

2.1 Step 1: Preprocessing

First, we gathered 8729 CSR reports spanning 1999-2016 from websites of unlisted or publicly traded Japanese firms. The reports were formatted as PDFs, which we converted to text files. Second, we extracted nouns from the text files through a morphological analysis by using MeCab (Kudo 2006), an open-source morphological analysis tool, for text in Japanese. We removed proper nouns, numbers, and words appearing fewer than five times. This process resulted in 87,084 nouns for consideration. Removing other parts of speech (e.g., verbs and adjectives) did not impair analysis because messages in the CSR reports are nearly fixed based on each word. For example, in most cases, the word "CO2" is used in the context of "CO2 reduction." This step may require slight adjustments when applying our model to languages other than Japanese. For example, nouns in Japanese do not have inflectional forms, while nouns in English have plural form such as "-s" or "-es." Therefore, stemming step should be introduced, and inflectional forms should be converted into the base form.

[2] For example, there are studies that investigate the effects of ESG disclosure on firm performance (Sampong et al. 2018; Wang et al. 2017), of ESG quantitative performance on firm performance (Friede et al. 2015; Velte 2017; Busch and Friede 2018), and of a combination of ESG disclosure and quantitative performance on firm performance (Gutsche et al. 2017; Xie et al. 2018; Fatemi et al. 2018).

Table 1 Examples of word classification

Group	Example of words	Num. of words
Environmental	Global warming, water quality	2450
Social	Overtime work, employee, human rights	1895
Governance	Committee, succession plan, decision-making	1316
Others	Strategy, operating profit, data, activity	81,423
Total		87,084

2.2 Step 2: Word Embedding

We mapped the extracted words to 50-dimensional vectors of real numbers, a procedure known as word embedding, by using a skip-gram model of word2vec, which employs neural network mapping. Refer to Mikolov et al. (2013) for detailed information on the model. Words with similar meanings tend to appear closer together in the mapped vector space. Therefore, we used word embedding vectors to classify words and specify their hierarchical structure in the following steps.

2.3 Step 3: Word Classification

To capture activities related to E, S, and G separately, we categorized the extracted words into four groups of embedded word vectors labeled "E," "S," "G," and "Other." Words labeled "Other" showed no direct relation to ESG activities. We manually labeled frequently used words as training data and classified the remaining words by using a neural network. Consequently, we obtained 2450 E-related words, 1895 S-related words, and 1316 G-related words. Because it would be cumbersome to display all the classification results, Table 1 shows some examples of the translated results.

2.4 Step 4: Word Structuring

To aid interpretation, we captured a hierarchical word structure based on the frequency and divergence of words in a tree model. An example appears in Fig. 1. Words in the upper hierarchies appear frequently and express general concepts or topics. Words in the lower hierarchies appear less frequently and express specific measures or issues. Lines in the figure tie highly related (low-divergence) words. We can interpret the position of each word intuitively. For example, "Air pollution" is an instance of "Pollution" under the "Environment" topic, and "Particulate matter" and "Exhaust gas" are more specific topics in "Air pollution."

The concrete method for specifying word structure is as follows. We define $\boldsymbol{v}_k$ as an embedding vector for word $k\,(k = 1, \dots, K)$, and $\|\boldsymbol{v}\|$ as a norm of vector $\boldsymbol{v}$. We specify the structures for E-related, S-related, and G-related words but omit suffixes indicating E, S, and G for simplicity in the following expressions. First, we sorted words in descending order by the number of occurrences in all 8729

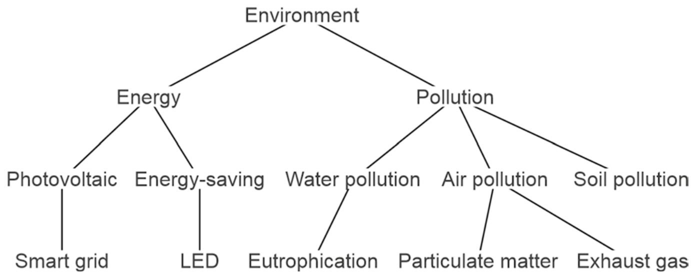

Fig. 1 An example of word structure

reports. However, the first word ($k = 1$) is fixed to "environment," "society," or "governance." Second, we calculated the divergence $d_{i,j}$ between a word embedding vector $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$. We define divergence $d_{i,j}$ as

$$d_{i,j} = (1 - \cos(\boldsymbol{v}_i, \boldsymbol{v}_j))^2, \tag{1}$$

where $\cos(\boldsymbol{v}_i, \boldsymbol{v}_j)$ denotes cosine similarity between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$. The value of $\cos(\boldsymbol{v}_i, \boldsymbol{v}_j)$ ranges from -1 to 1 and that of $d_{i,j}$ from 0 to 4. Therefore, if divergence diminishes (expands), words i and j have a similar (opposite) meaning. This definition of the divergence is heuristic with no theoretical foundation, but we find it generates comparatively intuitive results.

Third, we estimated tree structure by minimizing the sum of divergences of words under the constraint that words appearing more frequently in CSR reports must occupy a higher hierarchy than words appearing less frequently. This is formulated as the following binary combinatorial optimization problem:

$$\underset{x_{i,j} \in \{0,1\}, i<j}{\text{minimize}} \quad \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} x_{i,j} d_{i,j}, \tag{2}$$

$$\text{subject to} \quad \sum_{k=1}^{j-1} x_{k,j} = 1 \quad (j = 2, \dots, K), \tag{3}$$

where $x_{i,j}$ ($i < j$) are binary variables. Each variable takes one if there is a line (edge) between a word i and a word j and zero otherwise. The constraints (3) mean that each word must have one line connecting to an upper word. This problem can be solved by the greedy algorithm. Because upper hierarchies of the tree are important for interpreting results, we adjusted some nodes in the several top hierarchies manually from the perspective of qualitative plausibility by introducing additional constraints in the optimization problem.

The structures specified by our optimization-based approach are very large. Each tree consists of more than a thousand words. Liew et al. (2014) construct a similar hierarchical word structure, but it is considerably smaller than ours because their approach is based on human qualitative judgment.

Table 2 An example of word counting

Word	Count
Environment	53
Energy	25
Energy-saving	14
Photovoltaic	13
Smart grid	6
Pollution	4
Soil pollution	3
LED	2
Air pollution	2
Particulate matter	1
Water pollution	0
Eutrophication	0
Exhaust gas	0

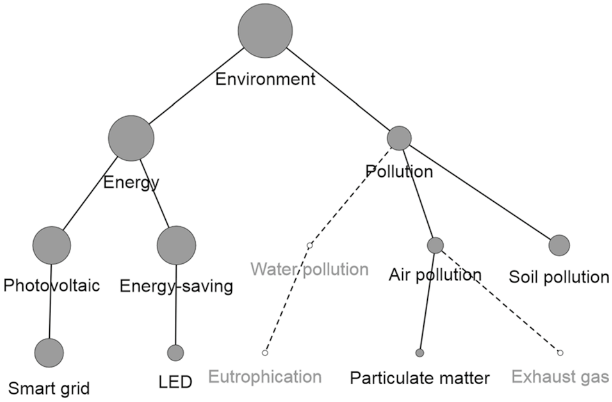

Fig. 2 An example of visualization

2.5 Step 5: Visualization

We describe our visualization procedure for a simple hypothetical example in Table 2 and Fig. 2. Table 2 presents the number of E-related words in the CSR report of a hypothetical firm. Figure 2 is a visualized result. Words used (not used) in the report are indicated as black (white) dots, and their diameter suggests the number of words. Therefore, we can discern that the firm works harder to resolve energy issues than pollution issues, and a smart grid is a key measure for this issue. We easily obtain a picture of ESG activities from the figure. If we seek more details about a word in the figure, we could see the relevant part of the original CSR report.

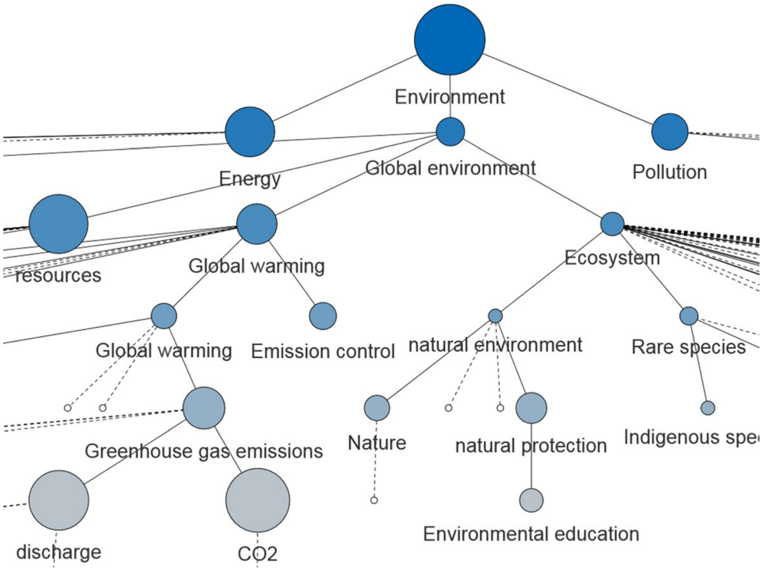

Fig. 3 Visualization of a CSR report

2.6 Visualization Example

Figure 3 presents a visualized result from the CSR report of a Japanese manufacturer of electric appliances. We show partial results to conserve space and translate the Japanese words. This example displays the connection from "environment" to "environmental education" via "global environment," "ecosystem," "natural environment," and "natural protection." This firm engages with ecosystem issues under the "global environment" topic and protects the "natural environment" through "environmental education." This model quickly, effectively, and objectively facilitates the processing of CSR reports and comparisons of such reports to reports from peer firms.

3 Scoring of ESG Activities

Scoring is essential in evaluating ESG activities because it provides foundations for comparisons. Therefore, we develop a method to score ESG activities from a qualitative perspective. We hypothesize that firms that are vigorously engaged in ESG deliver a higher ESG performance; thus, we checked the relationships between qualitative scoring and ESG quantitative performance.

3.1 Scores to Evaluate Qualitative Information

The CSR reports of firms engaged in extensive ESG activities likely contain numerous words concerning such activities. Therefore, the number of ESG-related words can act as a proxy for the quantity of ESG activities. Moreover, firms that make particular efforts to improve ESG performance likely use specific words in their CSR reports. Therefore, verbal specificity can act as a proxy for the quality of ESG activities. We define quantity and specificity scores. Although we calculate scores for E, S, and G, we again omit suffixes for simplicity. The quantity score $s_{i,t}^{qnt}$ is the sum of the logarithm of $n_{k,i,t}$, the appearance frequency of the word k in the report of firm i issued in year t. To prevent antilogarithms from being 0, we add 1 to $n_{k,i,t}$.

$$s_{i,t}^{qnt} = \sum_{k=1}^{K} ln(n_{k,i,t} + 1) \tag{4}$$

The specificity score $s_{i,t}^{spc}$ is the average of the divergence of words in the report from the top word ("environment," "social," or "governance"). $\mathbb{1}_{(n_{k,i,t} \geq 1)}$ represents the indicator function that takes one if the frequency of appearance is greater than or equal to one and takes zero otherwise.

$$s_{i,t}^{spc} = \frac{\sum_{k=1}^{K} \mathbb{1}_{(n_{k,i,t} \geq 1)} d_{1,k}}{\sum_{k=1}^{K} \mathbb{1}_{(n_{k,i,t} \geq 1)}}. \tag{5}$$

3.2 Relationship with ESG Performance

We suppose that large quantity and high specificity of ESG activities improve ESG quantitative performance, and some of them are already reflected in the current ESG quantitative performance, and others will be reflected in the future performance. Therefore, we expect that there are positive relationships between the scores and both the current ESG quantitative performance and subsequent improvement of the performance. To verify our expectation, we use the Thomson Reuters Asset4 score $s_{i,t}^{A4}$ as a proxy of ESG quantitative performance. Asset4 score evaluates quantitative outcomes of ESG activities. ESG analysts at Asset4 assess key performance indicators (KPIs) from public sources such as CSR reports or web pages. The number and examples of KPIs to calculate the Asset4 score are shown in Table 3. KPIs are evaluated by relative performance and summarized as Asset4 E-score, S-score, and G-score normalized between 0 and 100. Therefore, the scores show how well the company's ESG performance compared to other companies.[3]

We verified our expectation via firm-year pooled regression for companies listed in the first section of the Tokyo Stock Exchange between 2006 and 2013. We eliminated samples with missing score values and introduced a logarithm of market capitalization $Size_{i,t}$ at the end of year t as a control variable. We expect that the coefficients of $Size_{i,t}$ are positive because large firms have sufficient financial and human resources to conduct ESG activities. Regression is as follows:

[3] Please refer to Thomson Reuters (2011) for detailed information on the score.

Table 3 KPIs of Thomson Reuters Asset4 score

Score	Num. of KPIs	Examples of KPIs
Environmental	70	R&D expenditures divided by sales
		Total CO2 emission in tonnes divided by sales
Social	89	Percentage of women managers
		Percentage of employee turnover
Governance	70	Percentage of independent board members
		Company a signatory of the Global Compact

$$s_{i,t}^{A4} = \alpha + \beta_1 s_{i,t}^{qnt} + \beta_2 s_{i,t}^{spc} + \beta_3 Size_{i,t} + \sum_{t=2006}^{2012} \gamma_t D_t, \tag{6}$$

where D_t is a dummy variable for year t. Table 4 shows the results. The coefficients of both the quantity and specificity score in E, S, and G are positive and significant, except a case where a coefficient of the specificity score in S is positive but not significant. This result suggests that firms that report ESG activities with a greater amount of words and detailed descriptions tend to have an adequate ESG performance.

Next, we check relationships between our scores and subsequent improvement in ESG quantitative performance. Because the outcomes of ESG activities are likely reflected in long-term ESG performance, we use the difference in Asset4 scores for three years $\Delta s_{i,t+3Y}^{A4}$ as a proxy for subsequent improvement. We add the current Asset4 score $s_{i,t}^{A4}$ as a control variable because firms with Asset4 scores close to 100 (0) likely cannot raise (reduce) their current score significantly. In particular, we use the following regression:

$$\Delta s_{i,t+3Y}^{A4} = \alpha + \beta_1 s_{i,t}^{qnt} + \beta_2 s_{i,t}^{spc} + \beta_3 s_{i,t}^{A4} + \beta_4 Size_{i,t} + \sum_{t=2006}^{2012} \gamma_t D_t. \tag{7}$$

Table 5 shows the results. The coefficient of the specificity score in E is positive and significant. This implies that firm-specific efforts are essential to improve E-related performance because E-related issues vary from company to company. This is consistent with Liew et al. (2014), who find that environmental activities are heterogeneous among firms from the result of the topic analysis of CSR reports. On the other hand, contrary to our expectation, the coefficient of the quantity score in E is negative and significant. One potential explanation for this result is greenwashing. Namely, firms that are not vigorously engaged in activities use more words in their CSR reports to make benefits by misleading consumers (Delmas and Burbano 2011). However, Clarkson et al. (2008) and Mahoney et al. (2013) show the empirical results that do not support the existence of greenwashing. Therefore, further study regarding the reason for the negative coefficient should be conducted.

The coefficients of the quantity score are significant, and coefficients of the specificity score are not significant in S and G. This finding suggests that firms engaged

Table 4 Relationship with the current ESG performance

	Dependent variable: Asset4 score		
	E	S	G
Quantity score E	0.087***		
	(0.013)		
Specificity score E	243.824***		
	(38.926)		
Quantity score S		0.237***	
		(0.014)	
Specificity score S		26.551	
		(32.385)	
Quantity score G			0.083***
			(0.012)
Specificity score G			32.721***
			(9.229)
Size	6.445***	9.744***	4.367***
	(0.549)	(0.618)	(0.308)
Constant	− 65.213***	− 97.666***	− 55.285***
	(11.711)	(10.750)	(4.039)
Year dummy	Y	Y	Y
Observations	1882	1883	1858
Adjusted R^2	0.149	0.288	0.187

Values in parentheses are heteroscedasticity-consistent standard errors

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

in many S-related or G-related activities are more likely to improve their subsequent quantitative performance because S-related and G-related issues are relatively common among firms.

We confirm that the large quantity and high specificity of ESG activities indicate a higher current ESG quantitative performance. Moreover, the high specificity of E-related and the large quantity of S and G-related activities portend subsequent improvement of ESG quantitative performance. These results suggest the importance of qualitative information in CSR reports.

4 Conclusion

In recent years, ESG has become an essential factor in investment. This study examined qualitative information on ESG activities that is disclosed in CSR reports. Because practical applications of qualitative information are constrained by the difficulty of manual analysis, we developed a new text mining model to visualize ESG activities in structures of ESG-related words. Additionally, we proposed a score to evaluate qualitative information regarding ESG activities. We

Table 5 Relationship with the future improvement of ESG performance

	Dependent variable: improvement of ESG performance		
	E	S	G
Quantity score E	− 0.015*		
	(0.007)		
Specificity score E	99.649***		
	(23.668)		
Asset4 score E	− 0.272***		
	(0.019)		
Quantity score S		0.061***	
		(0.009)	
Specificity score S		− 9.230	
		(21.330)	
Asset4 score S		− 0.289***	
		(0.016)	
Quantity score G			0.029**
			(0.009)
Specificity score G			9.689
			(6.022)
Asset4 score G			− 0.354***
			(0.029)
Size	1.032**	0.664	1.227***
	(0.334)	(0.410)	(0.238)
Constant	− 7.831	13.043	− 14.522***
	(6.927)	(6.796)	(3.237)
Year dummy	Y	Y	Y
Observations	1882	1883	1858
Adjusted R^2	0.224	0.233	0.296

Values in parentheses are heteroscedasticity-consistent standard errors

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

found positive relationships between the proposed score and the ESG quantitative performance. In addition, we found some aspects of ESG activities contribute to subsequent improvement of the ESG performance. Although our results indicate the importance of qualitative information in CSR reports, the relationship between it and firm performance requires further clarification. That investigation remains our future task.

Acknowledgements We thank the asset management department of Mitsubishi UFJ Trust and Banking Corporation for advice and cooperation. The authors are responsible for any remaining errors.

References

Azhar, N. A., Pan, G., Seow, P. S., Koh, A., & Tay, W. Y. (2019). Text analytics approach to examining corporate social responsibility. *Asian Journal of Accounting and Governance*, *11*, 85–96.

Busch, T., & Friede, G. (2018). The robustness of the corporate social and financial performance relation: A second-order meta-analysis. *Corporate Social Responsibility and Environmental Management*, *25*(4), 583–608.

Clarkson, P. M., Li, Y., Richardson, G. D., & Vasvari, F. P. (2008). Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis. *Accounting, Organizations and Society*, *33*(4–5), 303–327.

Das, S., Choudhary, A., & Harding, J. (2016). An insight into CSR reports: A text mining approach. In *Conference paper, 3rd international conference on green supply chain*.

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review*, *54*(1), 64–87.

Fatemi, A., Glaum, M., & Kaiser, S. (2018). ESG performance and firm value: The moderating role of disclosure. *Global Finance Journal*, *38*, 45–64.

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, *5*(4), 210–233.

GSIA. (2018). *Global sustainable investment review*. Retrieved June 13, 2020, from http://www.gsi-alliance.org/wp-content/uploads/2019/03/GSIR_Review2018.3.28.pdf.

Gutsche, R., Schulz, J.F., & Gratwohl, M. (2017). Firm-value effects of CSR disclosure and CSR performance. In *EFMA-conference proceedings* (pp. 1–31).

KPMG. (2017). The road ahead, the KPMG survey of corporate responsibility reporting 2017.

Kudo, T. (2006). *Mecab: Yet another part-of-speech and morphological analyzer*. Retrieved June 13, 2020, from https://taku910.github.io/mecab/.

Liew, W., Adhitya, A., & Srinivasan, R. (2014). Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry*, *65*(3), 393–400.

Mahoney, L. S., Thorne, L., Cecil, L., & LaGore, W. (2013). A research note on standalone corporate social responsibility reports: Signaling or greenwashing? *Critical Perspectives on Accounting*, *24*(4–5), 350–359.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. arXiv preprint arXiv:1310.4546.

Sampong, F., Song, N., Boahene, K. O., & Wadie, K. A. (2018). Disclosure of CSR performance and firm value: New evidence from South Africa on the basis of the GRI guidelines for sustainability disclosure. *Sustainability*, *10*(12), 4518.

Shahi, A. M., Issac, B., & Modapothala, J. R. (2014). Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications*, *13*(1), 1450006.

Szekély, N., & Jv, Brocke. (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLoS ONE*, *12*, e0174807.

Thomson Reuters. (2011). Thomson Reuters Datastream Asset4 ESG content fact sheet.

Tremblay, M., Parra, C., & Castellanos, A. (2015). Corporate social responsibility reports: Understanding topics via text mining. In *Proceedings of Americas conference on information systems*.

Velte, P. (2017). Does ESG performance have an impact on financial performance? Evidence from Germany. *Journal of Global Responsibility*, *8*(2), 169–178.

Wang, Z., Hsieh, T. S., & Sarkis, J. (2017). CSR performance and the readability of CSR reports: Too good to be true? *Corporate Social Responsibility and Environmental Management*, *25*(1), 66–79.

Xie, J., Nozawa, W., Yagi, M., Fujii, H., & Managi, S. (2018). Do environmental, social, and governance activities improve corporate financial performance? *Business Strategy and the Environment*, *28*(2), 286–300.