Textual Evidence Extraction for ESG Scores

Naoki Kannan[1] and **Yohei Seki**[2]
[1]Graduate School of Comprehensive Human Sciences, University of Tsukuba
[2]Institute of Library, Information and Media Science, University of Tsukuba
s2321684@u.tsukuba.ac.jp
yohei@slis.tsukuba.ac.jp

Abstract

With the growing importance of environmental, social, and governance (ESG) information, ESG scores, which have been rated and published by various institutions, are used for investment decisions or corporate evaluation. The evidence for rating high or low ESG scores, however, is often vague and unclear. In this paper, we propose a method to extract the textual evidence of ESG scores by automatically labeling sentences with information related to ESG. Specifically, we constructed two labeling models for ESG and ESG sentiment, and extracted sentences with high confidence levels using the two models. At first, to label ESG-related information, we developed the annotation corpus using Japanese annual securities reports. Then, we constructed the labeling models by fine-tuning a large language model that was pre-trained on financial documents. The experimental results showed that the macro average F1 scores using the BERT model pre-trained on Japanese financial documents, were 0.874 for ESG labeling and 0.797 for ESG sentiment labeling respectively. These values were higher than those obtained using the comparative models that were pre-trained on Wikipedia documents only. We also confirmed that textual evidence for the ESG scores can be effectively extracted for the companies not included in the training dataset.

1 Introduction

In recent years, the global investment and corporate governance community has become increasingly interested in ESG (Environmental, Social and Governance), the three perspectives necessary for a company's long-term growth. The ESG score is an assessment of a company's level of commitment to ESG and is used by investors to determine the extent to which a company takes ESG factors into account when investing. ESG scores are provided to investors by various rating agencies. Recently, companies with higher ESG scores have been prioritised for investment, and the amount of assets under management for sustainable investments worldwide was expected to increase by 55% between 2016 and 2020 (Global Sustainable Investment Alliance (GSIA), 2021).

Despite this increase in investments taking ESG factors into account, many existing ESG scores are not open and unclear about how they are derived. They are also not consistently evaluated across rating agencies (Christensen et al., 2022). They are also incomplete, opaque and subject to considerable uncertainty (Avramov et al., 2022).

In this paper, we proposes a method for consistently extracting textual evidence from each text in the annual securities reports to assess ESG scores. Annual securities reports contain many general texts that are not related to company initiatives. Therefore, our method, which extracts only the textual evidence, can be used to help investors make decisions. This will help ESG-conscious investors to make decisions when investing in green assets.

Furthermore, "social capital" and "human capital" has become more important in assessing the social dimension ("S" in ESG) of corporate sustainability (Muñoz-Torres et al., 2019). In addition, in many ESG scores, social capital and human capital are evaluated as separate concepts in the assessment of sociability (International Sustainability Standards Board (ISSB), 2021; Toyo Keizai Inc., 2021; MSCI ESG Research LLC., 2023; FTSE Russell, 2022). On the other hand, to the best of authors' knowledge, no computational approach has been found that evaluates sociability separately into social and human capital perspectives. Therefore, we use pre-trained language models to automatically classify the sentences in annual securities reports describing a company's ESG efforts as separate labels for social and human capital, in addition to E and G, and extract textual evidence for rating the ESG score.

We specify the combination of ESG labels and their sentiment labels as the query to extract textual evidence for ESG scores. For example, to extract textual evidence for a high social capital score, we specify the attributes "social capital (S-1)" and "positive" as queries. Similarly, to extract evidence indicating a low human capital score, we specify the attribute "human capital (S-2)" and "negative" as a query. We also propose a method for extracting textual evidence of ESG scores using the confidence level in assessing ESGs and their sentiments. As confidence levels, the predicted probabilities of the ESG classifier and the ESG sentiment classifier are used. We regard sentences with high confidence levels as textual evidence for the ESG score. In this way, the textual evidence of ESG scores can be extracted from noisy securities reports to support investment decisions.

The contributions of this work can be summarized as follows:

1. We propose our method to extract textual evidence for ESG scores based on labeling ESG and ESG sentiment with their confidence levels from Japanese annual securities reports.
2. We distinguished the "S" in ESG labelling into social capital and human capital, and clarified that textual evidence on corporate social capital and human capital initiatives can be extracted respectively.
3. We also clarified that the pre-trained model on financial documents was effective in the ESG and ESG sentiment labeling tasks.

Our research revealed a strong correspondence between ESG scores and their textual evidence. This finding will be helpful in future works on automatic estimation of ESG scores from textual resources.

The structure of this paper is as follows. Section 2 presents relevant research on automatic labeling of ESG and ESG sentiment. In Section 3, we propose a method for extracting texts as evidence for rating ESG scores. Section 4 describes our dataset for labeling ESG information. In Section 5, we verify the effectiveness of large language models for labeling ESG information. In Section 6, we describe the experiment to verify the effectiveness of the proposed method in terms of extracting textual evidence. Section 7 discusses the results of the evaluation experiments. Finally, Section 8 summarizes the findings of our work.

2 Related Work

In recent years, computational approaches for analyzing ESG ratings have gradually intensified in response to the growing interest in ESG corporate activities.

2.1 ESG Labeling

Goel et al. (2022) achieved 2% higher accuracy than traditional BERT (Devlin et al., 2019) by combining various linguistic and semantic features. Dakle et al. (2022) collected a list of concepts and terms related to ESG issues in the financial domain and constructed a dataset of positive and negative term and concept pairs using a Sentence-BERT-based paraphrase detector. By fine-tuning BERT and RoBERTa (Liu et al., 2019) on this dataset, they achieved 96% accuracy on the validation set and 92.3% accuracy on the test set. Kiriu et al. (2020) analyzed corporate CSR reports and used Word2vec to obtain word-specific embedded representations and classify these words into three values: environmental, social, and governance. They then defined the quantity score and the specificity score and attempted to rate the ESG activities of the companies based on the qualitative information.

In this study, we extract textual evidence to help investors evaluate ESG scores. In contrast to the related works, we created a dataset manually annotated with ESG labels and fine-tuned several Transformer-based models, including BERT, which is pre-trained on a Japanese financial corpus. Also, to retrieve textual evidence for scores such as sociality score and human capital utilization score (Toyo Keizai Inc., 2021), this work defines S (social) concept as S-1 (social capital) and S-2 (human capital) instead of the E, S, and G three-valued classification. This approach follows the trend that human capital has recently become more important in evaluating corporate sociality.

2.2 ESG Sentiment Classification

Pasch et al. (2022) combined S&P Global's ESG score and text from annual reports to train an ESG sentiment model. Among the companies targeted, they labeled those with ESG scores above the median as "positive" and those with scores below the median as "negative" and fine-tuned based on their text. In this study, we performed labeling on a per-sentence basis using manually annotated labels and fine-tuned them to classify the sentiment of ESG-related sentences. Furthermore, by combining ESG labels and their sentiment labels, we extracted textual evidence for ESG scores. We defined the labeling strategy in Section 3.1 and demonstrated that the annotators annotated labels consistently even if they were not ESG or economic experts.

Aue et al. (2022) calculated a company's ESG rating by classifying news into sentiment categories and subtracting the percentage of negative news from the percentage of positive news. Fischbach et al. (2022) proposed ESG-Miner, a tool for analyzing the sentiment of a company's ESG-related news. By contrast, we extract textual evidence for ESG scores based on the assumption that the texts with positive (or negative) sentiments contain the reasons for high (or low) ESG scores.

The methods for fine-tuning pre-trained language models using small amounts of task-specific data are well known in the field of natural language processing in recent years (Howard and Ruder, 2018). These methods are effective in certain domains that contain many specialized words that do not appear often in general documents, such as sentences relevant to ESG. In this work, we performed automatic ESG sentiment classification by fine-tuning with BERT, RoBERTa, or ELECTRA, which have shown high performance in many natural language processing tasks, including sentiment classification.

3 Proposed Method

In this section, we describe our method for extracting textual evidence for rating ESG scores. In Section 3.1, we introduce the definition of the labeling attributes to be assigned to the texts. In Section 3.2, we describe the labeling model for each attribute to extract the evidence text of ESG scores.

3.1 Definition of ESG related attributes

We introduce and define two attributes to create the experimental dataset: ESG sentence type and ESG sentiment.

3.1.1 ESG sentence type

This attribute corresponds to the ESG type of content of the sentence. We define five labels: "Environmental (E)," "Social Capital (S-1)," "Human Capital (S-2)," "Corporate Governance (G)," "ESG General (All)," and "Other." The specific conditions for classification are defined in Table 1, referring to the SASB Standard[1], an international framework for ESG information disclosure, and actual sentences.

3.1.2 ESG sentiment

We defined three labels: "positive," "negative," and "neutral." This attribute indicates the sentiment of the sentence with respect to ESG. In addition, the annotation standards are defined so that non-experimental collaborators who are not experts in economics or investing can evaluate the annotation criteria. Table 2 shows the annotation standards, which are defined by focusing on the most commonly used phrases.

Table 1: ESG Sentence Type Requirements

Environmental (E)	• Contains expressions that consider greenhouse gases. • Contains climate-friendly phrases. • Contains phrases that consider the impact on the natural environment. • Contains other environmentally friendly phrases.
Social Capital (S-1)	• Contains statements related to product safety and user health. • Contains expressions related to privacy or information security. • Contains language about managing suppliers or trading partners (supply chain). • Contains other expressions of social responsibility outside the company.
Human Capital (S-2)	• Contains expressions related to employee recruitment, evaluation systems, and training that are not based on the old values. • Contains expressions of concern for workers' human rights and labor standards. • Contains other language that addresses the company's internal social responsibilities.
Governance (G)	• Contains expressions related to the structure of directors, executive officers, and auditors that could be considered relevant to the fairness and transparency of management. • Contains expressions of preparedness for serious risks that may affect the company's survival, compliance with laws and regulations, and ethical conduct. • Contains other language that reflects governance considerations.

Table 2: ESG Sentiment Requirements

Positive	• Contains phrases stating that the company is improving or is likely to improve in the future its long-term values. • Contains other phrases that are considered positive from an ESG perspective.
Negative	• Contains phrases that could be viewed as potentially detrimental to the long-term value of the company. • Contains other phrases that could be considered negative from an ESG perspective.
Neutral	• Contains statements about actions to be taken in the future for which it is not known whether they will actually be taken. • Contains ideas or actions of the company that cannot always be said to be generally true. • Contains statements that cannot be assessed as positive or negative from an ESG perspective.

[1] https://www.sasb.org/standards/materiality-map/

3.2 Classifiers and Ranking Model

In this work, we classify ESG-relevant sentences by fine-tuning the pre-trained language models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020). These transformer-based machine learning models are capable of acquiring context-aware embedded representations of words, allowing the same word to acquire different vector representations in different contexts.

The overall scheme of the proposed method is shown in Figure 1. First, we use the ESG dataset described in Section 4 and fine-tuned the pre-trained models for predicting ESG sentence type and ESG sentiment labels.

Next, we take sentences from the annual securities reports for the companies under evaluation and perform label prediction. The output is predicted labels and their probabilities for each classification model. The prediction probability is the output value of the Softmax function of the prediction label in the linear transformation layer of the pre-trained language model. This module extracts and ranks the sentences where the probabilities are above a certain threshold. They provide textual evidence for ESG scores.

In this paper, we use Japanese annual securities reports as a source of information to obtain a company's ESG initiatives because of the existence of uniform standards for disclosure, easy comparison with other companies in the same industry, and high machine readability of the data. We also use ESG scores provided by Toyo Keizai Inc. (2021) for major Japanese companies since 2007.

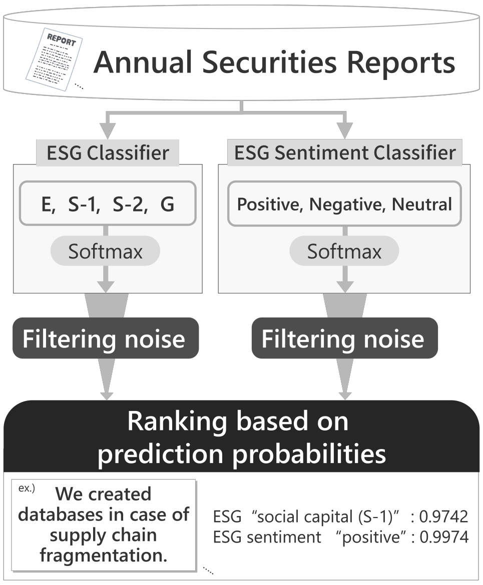
REPORT
Annual Securities Reports
ESG Classifier
E, S-1, S-2, G
Softmax
ESG Sentiment Classifier
Positive, Negative, Neutral
Softmax
Filtering noise
Filtering noise
Ranking based on prediction probabilities
ex.)
We created databases in case of supply chain fragmentation.
ESG "social capital (S-1)" : 0.9742
ESG sentiment "positive" : 0.9974

Figure 1: Proposed Method

4 ESG Label Classification Dataset

We created an ESG label classification dataset with sentences and annotated labels using annual securities reports collected from the electronic disclosure system EDINET[2].

4.1 Collection of annual securities reports (ASR)

The dataset was created by analyzing the XBML data from the annual securities reports (Financial Services Agency of Japan (JFSA), 2020) and removing unnecessary characters such as subheadings and symbols. We selected the actual companies for the dataset from Toyo Keizai Inc.'s ESG score ranking for 2021 (2021). Sentences were automatically split into sentence units at each punctuation point, and the output results were checked and corrected by the first author. The statistics of the dataset are as follows.

- Submission period: 2020/4/1 - 2021/3/31
- Data format: XBML
- Total number of sentences: 1,813 in total
- Target companies: 17 companies in 2 industries
  - Automobile manufacturers: 8 companies
  - Electrical manufacturers: 9 companies
- Target chapters in ASR
  - "Management policy, management environment, issues to address"
  - "Business risks"
  - "Research and development activities"
  - "Overview of corporate governance"

4.2 Annotation strategy for each attribute

To construct the ESG label classification dataset, the data described in Section 4.1 are manually annotated with the attributes defined in Section 3. This is done only for 744 sentences out of a total of 1,813 collected sentences. The remaining 1,069 sentences are used in Section 6 to evaluate the proposed method. At first, 562 of the 744 sentences were annotated by five annotators (the first author and four collaborators). Therefore, each of the 562 sentences has five annotations. The labels were determined by a majority vote. In cases in which we could not decide the results by majority vote because the votes were tied, we discussed until agreeing on a final decision. If all annotations for a sentence were different, it was chosen by discussion. Table 3 shows the Fleiss' κ coefficients (1971) for the five annotators as the annotator agreement degree. The κ values for all attributes were greater than 0.8 (almost perfect (Landis and Koch, 1977)).

[2] https://disclosure.edinet-fsa.go.jp/

From these results, we confirmed that there was no significant difference in the annotation results between the annotators. Based on these results, the first author annotated the remaining 182 sentences.

Table 3: Agreement for Each Attribute (Fleiss'κ)

Attribute	κ
ESG Sentence Type	0.89
ESG Sentiment	0.87

5 Language Models for Labeling Attributes

5.1 Objective

Here, we evaluate the accuracy of the proposed method for each attribute of the ESG label classification dataset created in section 4.2, and verify the effectiveness of the method. We describe comparative experiments using large language models pre-trained with different source data.

5.2 Method

First, we evaluate the classification accuracy of all 744 sentences in the ESG label classification dataset in Section 4.2 by predicting each attribute using the five-fold cross-validation method based on the model described in Section 3.2. The values used for the evaluation are precision, recall, $F1$ score, and accuracy, each of which is the macro average of each fold and each label.

The five pre-trained language models for comparison were as follows: BERT (wiki)[3], RoBERTa (wiki+CC100)[4], ELECTRA-base (wiki)[5], BERT (wiki+fin)[6], and ELECTRA-small (wiki+fin)[7]. The latter three models were developed by Suzuki et al. (2023) Note that the characters inside the brackets are the corpus name for pre-training. The hyperparameters were set as follows. The maximum token length was 128, the batch size was 32, the learning rate was 1e^{-5}, and the maximum number of epochs was 100. Learning was stopped when the loss function did not decrease for more than five epochs.

5.3 Results

Tables 4 and 5 show the classification results for ESG sentence types and ESG sentiment, respectively. The number of attribute labels for ESG sentence types used in the five-fold cross-validation is 111 for E, 162 for S-1, 70 for S-2, and 401 for G, respectively. The number of attribute labels for ESG sentiment is 566 for positive, 61 for negative, and 99 for neutral, respectively. Because BERT (wiki+fin) showed the best classification accuracy for all labels, we focus on BERT (wiki+fin) and BERT (wiki) for comparison.

Tables 6 and 7 show the F1 scores for the ESG sentence type and for the ESG sentiment label using BERT (wiki+fin) and BERT (wiki), respectively. In Table 6, the F1 scores of Social Capital (S-1) and Human Capital (S-2) are significantly improved compared to the other labels. Table 7 shows that the F1 scores of positive and neutral are significantly improved compared to the other labels.

Table 4: ESG Sentence Type Classification Results

Pre-trained Language Model	F1 score	Precision	Recall	Accuracy
BERT (wiki)	0.849	0.849	0.856	0.891
BERT (wiki+fin)	**0.874**	**0.870**	**0.885**	**0.906**
RoBERTa (wiki+CC100)	0.845	0.837	0.863	0.882
ELECTRA-base (wiki)	0.685	0.711	0.695	0.786
ELECTRA-small (wiki+fin)	0.175	0.135	0.250	0.539

Table 5: ESG Sentiment Classification Results

Pre-trained Language Model	F1 score	Precision	Recall	Accuracy
BERT (wiki)	0.785	0.807	0.778	0.885
BERT (wiki+fin)	**0.797**	**0.823**	**0.789**	**0.888**
RoBERTa (wiki+CC100)	0.783	0.802	0.774	0.876
ELECTRA-base (wiki)	0.628	0.711	0.627	0.819
ELECTRA-small (wiki+fin)	0.291	0.258	0.334	0.773

Table 6: Comparison Results of ESG Sentence Types Classification Using BERT (wiki+fin) and BERT (wiki) by Label Types

Label (Count)	F1 score	
	BERT (wiki+fin)	BERT (wiki)
Environmental (E - 111)	0.883	**0.887**
Social Capital (S-1 - 162)	**0.825**	0.796
Human Capital (S-2 - 70)	**0.832**	0.763
Governance (G - 401)	**0.955**	0.948

Table 7: Comparison Results of ESG Sentiment Classification Using BERT (wiki+fin) and BERT (wiki) by Label Types

Label (Count)	F1 score	
	BERT (wiki+fin)	BERT (wiki)
positive (566)	**0.879**	0.865
negative (61)	0.939	**0.942**
neutral (99)	**0.574**	0.553

[3] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[4] https://huggingface.co/rinna/japanese-roberta-base

[5] https://huggingface.co/izumi-lab/electra-base-japanese-discriminator

[6] https://huggingface.co/izumi-lab/bert-base-japanese-fin-additional

[7] https://huggingface.co/izumi-lab/electra-small-paper-japanese-fin-discriminator

6 Experiment: Extracting Textual Evidence for Rating ESG Scores of Companies

6.1 Objective

The purpose of this experiment is to verify whether the proposed method can extract textual evidence for rating the ESG scores of the companies. Note that the evaluation data for the companies are not used for fine-tuning BERT in Section 5.

6.2 Method

The text used in the experiment consists of 1,069 sentences from two industries and four companies that were not used in model training, as mentioned in Section 4.2 (Financial Services Agency of Japan (JFSA), 2020). Of these, we will refer to two companies as Electrical Manufacturer A and Automobile Manufacturer B for evaluation and two companies as Electrical Manufacturer α and Automobile Manufacturer β for verification.

Table 8 shows the ESG scores of Electrical Manufacturer A and Automobile Manufacturer B, which were extracted from the top 500 companies in the ESG Corporate Ranking in Japan (Toyo Keizai Inc., 2021). When ESG scores are above the mean with a large deviation from the mean, they are underlined. The scores that are below the mean with a large deviation are double-underlined. We evaluate our method to extract the textual evidence for the scores from the annual securities reports. At first, we took all 1,069 sentences into the proposed classifier and assign attributes to each sentence. The prediction probabilities of both classifiers are used as the confidence level of the sentence, and the sentences with high confidence levels are regarded as textual evidence of ESG scores. When the assigned attributes are a combination of "Environmental (E)," "Social Capital (S-1)," "Human Capital (S-2)," or "Governance (G)" and "Positive," then the sentences are textual evidence for high ESG scores, whereas when they are combined with "Negative," they serve as textual evidence for low ESG scores.

Next, we filter the noisy sentences for textual evidence. The filtering is done by setting thresholds for the prediction probabilities of the ESG classifier and the ESG sentiment classifier as the confidence level of the sentence, respectively, as described in Section 3.2. The thresholds were decided by actually checking the sentences as the positive cases with the lowest prediction probabilities in the verification data.

From the evaluation scores of the sentences of the electronics manufacturer α, we decided the threshold for electronics manufacturers, and from the evaluation scores of the sentences of the automobile manufacturer β, we decided the threshold for automobile manufacturers. Attributes for which no sentences were extracted from the verification data were given 0 as a threshold.

Finally, the extracted sentences were evaluated by three experiment participants, including the first author, to decide whether they actually served as textual evidence for the ESG scores. We used the precision at the top 1, 5, 10, and 20 ranks as the evaluation measure.

6.3 Results

The evaluation results for Electrical Manufacturer A and for Automobile Manufacturer B are shown in Tables 9 and 10. From these results, we found that we could extract the textual evidence for the companies rated as having high ESG scores. We also found that there were still challenges in extracting the textual evidence for the companies rated as having low ESG scores.

7 Discussion

In this section, we discuss the effectiveness of our proposed method based on the validity of extracted sentences as textual evidence, the effect of the distinction of S-1 (social capital) and S-2 (human capital), and the validity of the model pre-trained on the financial corpus. We also discuss failure analysis.

7.1 Textual evidence for rating ESG scores

In Figures 2 and 3, we show two example sentences that were extracted as textual evidence for high ESG scores from the annual securities reports

Table 8: ESG Scores of Companies for Evaluation

	ESG Score	E Score	S1 Score	S2 Score	G Score
	(Average: 320.4)	(Average: 76.32)	(Average: 77.11)	(Average: 77.91)	(Average: 88.96)
Electrical Manufacturer A (Rank 46)	372.2	97.4	93.1	85.9	93.8
Automobile Manufacturer B (Rank 336)	300.4	74.4	85.2	60.6	80.2

Table 9: Results of Ranking Textual Evidence for Rating ESG Scores of Electrical Manufacturer A

Query	Threshold		# of textual evidence		Precision@k			
Specified attribute label	ESG classification	ESG sentiment classification	judged by humans	extracted	k=1	k=5	k=10	k=20
"Environmental (E)" and "positive"	0.8456	0.9969	13	16	1.000	1.000	0.900	–
"Social Capital(S-1)" and "positive"	0.9285	0.9963	11	22	1.000	0.400	0.600	0.500
"Human Capital (S-2)" and "positive"	0.8796	0	9	12	1.000	1.000	0.800	–

Table 10: Results of Ranking Textual Evidence for Rating ESG Scores of Automobile Manufacturer B

Query	Threshold		# of textual evidence		Precision@k	
Specified attribute label	ESG classification	ESG sentiment classification	judged by humans	extracted	k=1	k=5
"Social Capital(S-1)" and "positive"	0.9902	0.9967	6	7	1.000	0.800
"Human Capital (S-2)" and "negative"	0	0	0	1	0.000	–

of Electrical Manufacturer A and Automobile Manufacturer B[1], respectively. For Electrical Manufacturer A, we extracted sentences describing that they were building facilities and creating new training curricula for talent development. For Automobile Manufacturer B, we extracted sentences describing that they were creating supply chain databases for each part to prepare for emergencies and manage the supply chain.

Munoz et al. (Muñoz-Torres et al., 2019) examined the rating methodologies used by the eight major ESG rating agencies and analyzed the criteria and their strength (frequency of occurrence) for each environmental, social, and governance dimension. According to their analysis of the ESG rating agencies' evaluation processes in 2017, the main social responsibility criteria, especially those related to human capital, were the quality of working conditions, health and safety, labor management, and human rights. These criteria emphasize responsibility toward employees as stakeholders. In fact, looking at the actual text, the provision of sufficient training programs for employees is considered to contribute to the human capital criteria mentioned for Electronics Manufacturer A. Therefore, the extracted text here corresponds to the addition of ESG scores from a human capital perspective and is considered evidence based on the text of the ESG scores. Similarly, for Automobile Manufacturer B, it can be said that implementing supply chain management to prepare for emergencies fulfills the company's responsibility to society by minimizing the impact on production in the event of a crisis, allowing production to continue. Therefore, the extracted text here corresponds to the addition of ESG scores from a social capital perspective and is considered evidence based on the text of the ESG scores.

Label: "Human Capital (S-2)"; "Positive"
Value: 0.9968
"We built the Academy Training Center as part of our centennial project to develop the people who will drive our growth. We also reformed personnel development programs, such as the introduction of new curricula, with focusing on the development for highly skilled technicians and professionals."

Figure 2: Textual Evidence for ESG Scores for Electrical Manufacturer A (Original Text in Japanese)

Label: "Social Capital (S-1)"; "Positive"
Value: 0.9974
"In addition, in case of an emergency, we have also performed maintenance to keep the supply chain database of our current core and secondary suppliers to mitigate the impact of supply chain disruptions. This helped us to identify potentially affected suppliers and parts at an early stage, to identify required stocks, to propose alternative manufacturing, and to support the restoration of production facilities. "

Figure 3: Textual Evidence for ESG Scores for Automobile Manufacturer B (Original Text in Japanese)

7.2 Effect of S-1 and S-2 distinction in ESG

We investigate the effect of extracting textual evidence using S-1 (social capital) and S-2 (human capital) by comparing the labeling results using "S" without distinguishing the two attributes.

We compare the classification results of the 4-value classifier for E, S-1, S-2, and G with those of the 3-value classifier for E, S, and G using the same test data. We took 1,069 sentences from the two automobile manufacturers and two electronics manufacturers used in Section 6 as input data into both the ESG 4-value classifier and the ESG 3-value classifier, respectively. Then, we checked which labels are predicted by the 4-value classifier for the top 20 sentences with the highest prediction probability in the 3-value classifier among the sentences classified as "S" in the 3-value classifier.

[1] Note that we omitted proper nouns that identified the company.

Table 11: ESG Sentence Examples Correctly Classified as S-1 / S-2 with BERT (wiki+fin)

Input (Original Text in Japanese)	Prediction Labels	
	BERT (wiki+fin)	BERT (wiki)
We used universal design for our products for wheelchair users to operate easily.	S-1	E
In the United States, our 2020 models were rated as *TOP SAFETY PICK+* from the IIHS (Insurance Institute for Highway Safety) in its 2020 safety survey.	S-1	E
We, the management team, aim to be a company where everything is communicated openly and honestly. We also take the initiative in promoting continuous improvement of our organizational environment.	S-2	G

The results showed that many of the sentences classified as "S" by the ternary classifier were actually S-1 (social capital). We examined 80 sentences, 20 sentences from each of the four companies, and found that 57 (71%) of them were S-1 (social capital). This implied that the proposed method using a 3-value classifier cannot extract enough textual evidence of S-2 (human capital). In addition, as shown in Table 6, the annotated dataset was skewed, with 162 sentences for S-1 and 70 sentences for S-2, which might have led to this result. Thus, we conclude that the proposed method using a 4-value classifier allows us to effectively extract textual evidence for both social and human capital sentences.

7.3 Effect of pre-training using financial documents

We investigate the classification results using BERT (wiki+fin), a model retrained on Japanese financial texts by comparing with BERT (wiki), a model not retrained on Japanese financial texts. It should be noted that we omitted proper nouns that could identify the company. From Table 6, the F1 scores of S-1 (Social Capital) and S-2 (Human Capital) were improved using BERT (wiki+fin). There were a certain number of sentences that were correctly classified as S-1 or S-2 in BERT (wiki+fin) but were incorrectly classified as E or G in BERT (wiki). Examples of such sentences are shown in Table 11. All of these sentences are considered to be positive promotional phrases by the company itself, which are specific to annual securities reports. As described, depending on pre-training using financial texts, we identified several contexts that were well captured by BERT (wiki+fin).

7.4 Failure Analysis

We confirmed that textual evidence for rating ESG scores can be extracted with our proposed method. However, some extracted sentences were not used as the textual evidence for ESG scores. We discuss some typical failure cases as follows.

An example of misclassification is the sentence: "AI predicts Chemical Oxygen Demand in 2 hours." It is difficult to consider this sentence as the textual evidence for rating the environmental score, but it may have been misclassified because of the inclusion of phrases such as "oxygen," which are often used in the context of environmental protection.

Another example is the following sentence: "We have added a range of composite environmental sensors." This sentence was also misclassified because of the inclusion of the word "environment." This is because of the fact that sentences annotated with labels such as "Social Capital (S-1)," "Human Capital (S-2)," and "Governance (G)" did not contain many phrases related to "social" or "governance," whereas sentences annotated with the label "Environmental (E)" often contained many phrases related to "environment."

8 Conclusion

In this work, we annotated the sentences in annual securities reports and created classification models by fine-tuning large language models pre-trained on financial documents. We proposed a method for extracting the textual evidence for ESG scores by automatically assigning both ESG sentence type labels (E, S-1, S-2, and G) and ESG sentiment labels to the sentences in annual securities reports and ranking them with their prediction probabilities as the confidence levels to filter out noisy sentences. Through experimentation, we have confirmed that it is possible to extract the textual evidence of companies rated as having a high ESG score.

In the future, we plan to automatically estimate ESG scores from textual resources based on the strong correspondence between ESG scores and their textual evidence.

Ethical sentence

This research was conducted with the approval of the Ethics Review Committee of the Institute of Library, Information and Media Science, the University of Tsukuba. The participants in the corpus creation experiment were asked to sign a consent form in advance and were allowed to quit the experiment at any time.

Acknowledgments

This work was partially supported by the Japanese Society for the Promotion of Science Grant-in-Aid for Scientific Research (B) (#23H03686), and Grant-in-Aid for Challenging Exploratory Research (#22K19822).

References

Tanja Aue, Adam Jatowt, and Michael Färber. 2022. Predicting Companies' ESG Ratings from News Articles Using Multivariate Timeseries Analysis. *arXiv preprint arXiv:2212.11765*.

Doron Avramov, Si Cheng, Abraham Lioui, and Andrea Tarelli. 2022. Sustainable investing with esg rating uncertainty. *Journal of Financial Economics*, 145(2, Part B):642–664.

Dane M. Christensen, George Serafeim, and Anywhere Sikochi. 2022. Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *The Accounting Review*, 97(1):147–175.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Parag Pravin Dakle, Shrikumar Patil, Sai Krishna Rallabandi, Chaitra Hegde, and Preethi Raghavan. 2022. Using transformer-based models for taxonomy enrichment and sentence classification. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 250–258, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186. Association for Computational Linguistics.

Financial Services Agency of Japan (JFSA). 2020. Annual Securities Report (*Yuuka Shoken Houkokusho* in Japanese).

Jannik Fischbach, Max Adam, Victor Dzhagatspanyan, Daniel Mendez, Julian Frattini, Oleksandr Kosenkov, and Parisa Elahidoost. 2022. Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

FTSE Russell. 2022. Esg-ratings-overview.pdf. https://research.ftserussell.com/products/downloads/ESG-ratings-overview.pdf. (Accessed on 04/23/2023).

Global Sustainable Investment Alliance (GSIA). 2021. "GLOBAL SUSTAINABLE INVESTMENT REVIEW 2020". https://www.gsi-alliance.org/wp-content/uploads/2021/08/GSIR-20201.pdf. (Accessed 04/21/2023).

Tushar Goel, Vipul Chauhan, Suyash Sangwan, Ishan Verma, Tirthankar Dasgupta, and Lipika Dey. 2022. TCS WITM 2022@FinSim4-ESG: Augmenting BERT with linguistic and semantic features for ESG data classification. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 235–242, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

International Sustainability Standards Board (ISSB). 2021. Materiality map screenshot). https://www.sasb.org/wp-content/uploads/2021/11/MMap-2021.png. (Accessed on 04/23/2023).

Takuya Kiriu and Masatoshi Nozaki. 2020. A Text Mining Model to Evaluate Firms' ESG Activities: An Application for Japanese Firms. *Asia-Pacific Financial Markets*, 27(4):621–632.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics. International Biometric Society*, pages 159–174.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

MSCI ESG Research LLC. 2023. Esg ratings methodology - msci esg ratings methodology.pdf. https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology+%28002%29.pdf. (Accessed on 04/23/2023).

María Jesús Muñoz-Torres, María Ángeles Fernández-Izquierdo, Juana M. Rivera-Lirio, and Elena Escrig-Olmedo. 2019. Can environmental, social, and governance rating agencies favor business models that promote a more sustainable development? *Corporate Social Responsibility and Environmental Management*, 26(2):439–452.

Pasch Stefan and Ehnes Daniel. 2022. NLP for Responsible Finance: Fine-Tuning Transformer-Based Models for ESG. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3532–3536.

Masahiro Suzuki, Hiroki Sakaji, Masanori Hirano, and Kiyoshi Izumi. 2023. Constructing and analyzing domain-specific language model for financial text mining. *Information Processing & Management*, 60(2):103194.

Toyo Keizai Inc. 2021. *CSR Corporate Social Responsibility White Paper 2021 (in Japanese)*. Toyo Keizai Inc.