Multi-Lingual ESG Impact Duration Inference

Chung-Chi Chen,[1] **Yu-Min Tseng,**[2] **Juyeon Kang,**[3] **Anaïs Lhuissier,**[3]
Yohei Seki,[4] **Hanwool Lee**,[5] **Min-Yuh Day**,[6] **Teng-Tsai Tu**,[7] **Hsin-Hsi Chen**[8]
[1]AIST, Japan
[2] Data Science Degree Program, National Taiwan University and Academia Sinica, Taiwan
[3]3DS Outscale, France, [4]University of Tsukuba, Japan, [5]NCSOFT, South Korea
[6]Graduate Institute of Information Management, National Taipei University, Taiwan
[7]Graduate Institute of International Business, National Taipei University, Taiwan
[8]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

Abstract

To accurately assess the dynamic impact of a company's activities on its Environmental, Social, and Governance (ESG) scores, we have initiated a series of shared tasks, named ML-ESG. These tasks adhere to the MSCI guidelines for annotating news articles across various languages. This paper details the third iteration of our series, ML-ESG-3, with a focus on impact duration inference—a task that poses significant challenges in estimating the enduring influence of events, even for human analysts. In ML-ESG-3, we provide datasets in five languages (Chinese, English, French, Korean, and Japanese) and share insights from our experience in compiling such subjective datasets. Additionally, this paper reviews the methodologies proposed by ML-ESG-3 participants and offers a comparative analysis of the models' performances. Concluding the paper, we introduce the concept for the forthcoming series of shared tasks, namely multi-lingual ESG promise verification, and discuss its potential contributions to the field.

Keywords: argument relation, argument mining, cross-lingual

1. Introduction

In recent years, the Environmental, Social, and Governance (ESG) criteria have emerged as vital measures for evaluating a company's impact on the world. These criteria not only inform investors about the sustainability and ethical implications of investing in a company but also help consumers and employees align with organizations that share their values. However, accurately assessing a company's performance in these areas remains a complex challenge, exacerbated by the dynamic and multifaceted nature of ESG-related information. To address this challenge, our research community has initiated the ML-ESG series of shared tasks.

Given the increasing importance of ESG for accounting departments and investors, many rating companies have emerged, such as DJSI, CDP, FTSE, MSCI, and Sustainalytics. In the ML-ESG shared tasks series, we selected MSCI's rating standard for annotations on ESG-related news articles. In ML-ESG-1 (Chen et al., 2023a), we explored the ESG Issue Identification task. In ML-ESG-2 (Chen et al., 2023b), we focused on ESG Impact Type Identification. After understanding the issue (up to 44 aspects) and the type (opportunity or risk), ML-ESG-3 goes a step further to infer the impact duration. This task aims to estimate how long the effects of certain events or actions taken by a company will last, impacting its ESG scores. It involves not only interpreting the immediate effects of an event but also predicting its long-term consequences—something that even experienced human analysts find challenging.

This paper presents an overview of ML-ESG-3, including the datasets developed, and the insights gained from compiling these datasets. ML-ESG-3 includes news articles in five different languages, acknowledging the global nature of ESG issues and the importance of diverse linguistic representation in ESG analysis. Moreover, we summarize the methodologies proposed by participants in ML-ESG-3, offering an analysis of their models' performance. Finally, we conclude with a discussion on the next series of shared tasks, focusing on multi-lingual ESG promise verification. This forthcoming task is designed to further the field's understanding of how companies' promises regarding ESG performance align with their actual actions and impacts. By exploring the verification of these promises across different languages, we aim to enhance the transparency and accountability of companies on a global scale.

2. Dataset

2.1. Guidelines

The MSCI guidelines delineate the timeline for impact duration as follows: short-term is under 2 years, long-term is 5+ years, and medium-term encompasses the period in between. Given that all actions carry long-term consequences, the following advice is provided to avoid indiscriminately assigning the label "long" to each time frame:

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 219–227

	Train			Test		
	within 2 years	2 to 5 years	longer than 5 years	within 2 years	2 to 5 years	longer than 5 years
Chinese	97	69	226	11	8	25
English	82	198	265	6	47	83
French	122	222	293	31	32	83
Korean	446	212	142	96	40	64
Japanese	15	7	5	291	167	715

Table 1: Statistics of impact duration dataset.

	Train			Test		
	High	Medium	Low	High	Medium	Low
English	196	243	106	60	59	17
French	198	317	122	45	53	48

Table 2: Statistics of impact-level.

- Pay attention to any time indications within the text, as these can serve as reliable indicators of the intended duration, such as references to political agendas or statements from scientists.
- Consider the subject matter of the sentence: if the focus is on contract negotiations or diplomacy rather than the issue itself, it may be appropriate to classify the paragraph as short-term, despite potential long-term benefits or harms.
- Recognize that some topics inherently imply a specific impact duration based on common sense. For issues that cannot be predicted with absolute certainty, opting for a safe, neutral mean or the most likely impact duration is advisable.
- In the absence of explicit date references or common-sense driven topics, focus on keywords that indicate the type of issue being discussed or the nature of the debate, rather than the overarching topic.

In addition to the impact duration, English and French datasets provide additional impact-level annotations. Since evaluating the impact of an event can be utterly subjective, to minimize this, here are some pieces of advice to remain objective and indications as to what could be considered low, medium, and high impact.

- Take into consideration the broader issue at stake and not only the discussed matter, to get a better picture of the potential impact.
- Reference similar previous events as a benchmark.
- National or international events do not always signify high impact. Decision-makers can take small steps towards their goals, and these should be assessed as such for the sake of our shared task.

	Train	Test
Opputunity	462	105
Risk	229	66
Cannot distinguish	109	29

Table 3: Statistics of Korean impact type annotations.

- The impact level may be adjusted according to a balance of positive and negative impacts. For example, a highly impactful/problematic event may be partially resolved.

Korean is the new language of ML-ESG, and impact-type labels are also provided at this time. Please refer to our previous paper (Chen et al., 2023b; Tseng et al., 2023) for more details.

2.2. Statistics

The Cohen's Kappa coefficient (Cohen, 1960) for datasets in Chinese, Korean, and Japanese yielded values of 0.21, 0.26, and 0.31, respectively. This variation underscores the challenges inherent in inferring the duration of the impact. To ensure the quality of the training and testing data, we exclusively utilized instances from the Chinese dataset that received uniform labels from the annotators. Table 1 details the statistics of the annotation results. The distribution of impact levels and types are presented in Tables 2 and 3, respectively. Table 1 and 2 demonstrate that low impact duration and length data are less abundant for the English and French languages.

2.3. Challenges

This edition faced a double challenge due to the previously mentioned nature of ESG news: unbalanced label distribution and annotation disagreements. For the first issue, the detailed guidelines guaranteeing a certain objectivity cannot ignore the fact that annotators having different backgrounds can still interpret the guidelines with a biased view, adjusting the impact level and duration accordingly during the annotation process. Thus, we accentuated our efforts on both cross- and group reviews to reach a high level of objectivity and coherence. For the latter, as most of ESG-related actions carry relatively long-term consequences with a medium to

	Best-Performing Method	Paper
Chinese	Longformer (Beltagy et al., 2020)	Tseng et al. (2023)
English	DeBERTa-v3 (He et al., 2023)	Dakle et al. (2024)
French	BERT (Devlin et al., 2019) & FinBERT (Araci, 2019)	Banerjee et al. (2024)
	GPT4 (OpenAI et al., 2024)	Tian and Chenn (2024)
Korean	XLM-RoBERTa (Conneau et al., 2020)	Dakle et al. (2024)
	KF-DeBERTa (jeo, 2023)	Kim et al. (2024)
Japanese	DeBERTaV3 (He et al., 2023)	Dakle et al. (2024)
	XLM-RoBERTa (Conneau et al., 2020)	Abburi et al. (2024)

Table 4: Best-performing methods.

English			French			Korean			Japanese		
	Micro-F1	Macro-F1		Micro-F1	Macro-F1		Micro-F1	Macro-F1		Micro-F1	Macro-F1
Jetsons_3	64.71%	52.47%	LIPI_1	56.16%	49.19%	Jetsons_1	70.00%	66.24%	Jetsons_2	36.50%	25.60%
Team Tredence_3	66.18%	50.12%	upaya_2	52.05%	48.73%	3idiots_1	67.50%	61.98%	Drocks_1	36.30%	25.60%
LIPI_1	58.82%	49.62%	French-_kaka_1	46.58%	47.42%	3idiots_3	67.50%	61.54%	Jetsons_3	36.50%	25.50%
fin-turbo_2	69.12%	46.89%	fin-turbo_1	56.16%	46.22%	3idiots_2	66.50%	61.02%	kaka_1	34.90%	25.50%
Jetsons_1	61.03%	46.70%	Drocks_3	50.00%	45.77%	Team Tredence_2	64.00%	58.18%	Team Tredence_1	43.10%	24.80%
Team Tredence_2	58.09%	45.45%	Team Tredence_2	53.42%	45.66%	Jetsons_3	64.00%	57.39%	Drocks_2	34.40%	24.50%
upaya_3	60.29%	44.23%	fin-turbo_3	48.63%	44.88%	Drocks_3	62.50%	55.17%	Albatross_1	31.90%	23.70%
Drocks_2	59.56%	44.14%	fin-turbo_2	57.53%	43.35%	kaka_1	56.00%	52.94%	Drocks_3	32.70%	23.40%
CriticalMinds_3	65.44%	43.86%	Drocks_1	50.00%	43.31%	Team Tredence_3	57.50%	52.36%	Team Tredence_2	39.90%	21.80%
Drocks_1	58.82%	43.37%	upaya_1	46.58%	42.86%	Team Tredence_1	59.50%	51.58%	Jetsons_1	30.90%	21.50%
Drocks_3	57.35%	43.01%	Drocks_2	49.32%	42.52%	fin-turbo_3	65.50%	51.26%	LIPI_3	27.90%	19.20%
CriticalMinds_1	64.71%	42.81%	Jetsons_3	54.11%	42.23%	Drocks_2	57.50%	48.39%	LIPI_1	29.90%	18.60%
upaya_1	57.35%	42.75%	Team Tredence_3	49.32%	40.54%	Drocks_1	60.50%	48.02%	Team Tredence_3	29.90%	18.00%
DICE_2	55.88%	42.53%	Team Tredence_1	41.78%	39.70%	fin-turbo_1	65.00%	47.32%	LIPI_2	24.30%	16.10%
Jetsons_2	56.62%	42.28%	Jetsons_1	47.95%	37.06%	fin-turbo_2	64.50%	47.30%	ABC_1	18.90%	11.80%
CompLx_1	56.62%	42.07%	SamNLP_2	43.84%	36.84%	FIT_2	61.50%	43.98%	IMNTPU_2	11.90%	7.10%
SamNLP_2	57.35%	41.94%	LIPI_3	37.67%	36.41%	FIT_1	52.50%	43.82%	IMNTPU_1	11.10%	5.00%
MLG-TRDDCPune_1	52.21%	41.75%	Jetsons_2	46.58%	34.62%	Jetsons_2	42.00%	38.11%			
MLG-TRDDCPune_3	52.21%	41.75%	DICE_1	34.93%	34.45%	FinNLP_1	49.00%	36.87%			
MLG-TRDDCPune_2	52.21%	41.75%	SamNLP_1	46.58%	33.70%	FinNLP_2	49.50%	36.75%			
CriticalMinds_2	59.56%	41.53%	CriticalMinds_3	54.11%	32.88%	FinNLP_3	49.00%	36.47%			
LIPI_3	52.21%	40.73%	CriticalMinds_2	46.58%	32.19%	LIPI_1	3.50%	4.38%			
fin-turbo_3	61.76%	40.35%	upaya_3	41.10%	32.09%						
fin-turbo_1	58.82%	39.83%	CriticalMinds_1	54.79%	30.33%						
SamNLP_1	61.76%	39.59%	LIPI_2	41.10%	30.02%						
FinTwin_1	62.50%	38.90%									
Team Tredence_1	61.03%	38.74%									
upaya_2	51.47%	38.55%									
DICE_3	55.15%	37.84%									
DICE_1	44.85%	37.07%									
kaka_1	52.94%	36.36%									
LIPI_2	50.00%	32.70%									

Table 5: Performance — Impact Duration.

high impact on society and industries, a substantial analytical work was conducted to reveal which topics and impact type could entail a low impact level and/or low impact duration in order to obtain quality datasets. Overcoming these challenges evidences the necessity to assist human analysts.

3. Methods

A total of 12 teams share their methods in ML-ESG-3. We show the best-performing method in Table 4, and provide an overview of participants' methods in this section.

3.1. Impact Duration

In the Korean subtask of the ML-ESG-3 challenge, two teams, Jetsons (Dakle et al., 2024) and 3idiots (Kim et al., 2024), showcased strategies for improving ESG impact duration prediction accuracy amidst challenges like class imbalance and data scarcity. The Jetsons team led the field by implementing a data augmentation strategy that utilized self-training with supplementary English and French ESG articles to generate pseudo labels, thus enriching their training dataset. This approach, coupled with the fine-tuning of an XLM-RoBERTa model (Jetsons_1) (Conneau et al., 2020) , showcased the effectiveness of integrating sophisticated language models with data augmentation to improve multilingual ESG impact duration predictions. The 3idiots team distinguished themselves with a semi-supervised learning (SSL) approach, utilizing a finance-specialized pre-trained language model, KF-DeBERTa (jeo, 2023), along with advanced data augmentation techniques (Wei and Zou, 2019). By enriching their dataset with unlabeled ESG-related news articles, they achieved significant results, illustrating the potential of SSL and domain-specific models in enhancing NLP tasks with limited labeled data.

In the Japanese impact duration subtask, both Jetsons_2 (Dakle et al., 2024) and Drocks_1 (Abburi et al., 2024) achieved first place with the highest Macro F1 score. Dakle et al. (2024) implemented three strategies in the Japanese subtask: the English translation approach (Jetsons_2), the ensemble approach (Jetsons_3), and

the fine-tuned multilingual model approach (Jetsons_1). For the English translation approach, Japanese texts were translated into English using the Google API, followed by a fine-tuning of the DeBERTa-v3-small model (He et al., 2023) on the class labels using the translated text. In the ensemble approach, they combined three models: XLM-RoBERTa, Longformer, and DeBERTa. The comparative results indicated that both the English translation and ensemble approaches outperformed the fine-tuned multilingual model approach, which was based on XLM-RoBERTa. Abburi et al. (2024) employed a data augmentation approach based on English text translated using the DeepL service, augmented with PEGASUS and GPT-mix, and then translated back into Japanese. They also trained an ensemble model that combined transformers (XLM-RoBERTa), CNN, and Voyage AI embeddings. It is noteworthy that a common characteristic of both teams was their reliance on English translation.

3.2. Impact Level

In the English impact duration and level subtasks, Jetsons_3 and Jetsons_1 (Dakle et al., 2024), respectively, proposed the best performing model with the highest Macro F1 score, while LIPI_1 (Banerjee et al., 2024) acheived the best score for the French impact duration task and kaka_1 (Tian and Chenn, 2024) for the French impact level task.

To handle multilingual datasets with relatively low volume and issues of label imbalance, most participants translated all datasets into English using tools like DeepL and Google Translate and explored data augmentation techniques using recent LLMs (e.g. GPT, Gemini, T5) to generate more samples. Those efforts on the dataset show improvements in some cases (Banerjee et al., 2024) (Dakle et al., 2024) but not in others (Atanassova et al., 2024). This observation indicates that processing ESG-related information seems to be language-dependent, so that it requires a strategy determining the relevance of data to each specific language (Dakle et al., 2024).

Most participants largely explored pre-trained transformer-based models, particularly, BERT, RoBERTa, DeBERTa and Longformer, by fine-tuning them on the ESG dataset. We observe that training various transformer models separately and subsequently combining them through an ensembling process has proven to yield the best results in impact duration and level classification (Yang and Rong, 2024) (Kao et al., 2024) (Bougiatiotis et al., 2024) (Dakle et al., 2024). An alternative approach involves fine-tuning Mistral-7B on a dataset generated by GPT-4, which contains articles along with information on the impact level, length, and rationale behind the classification (Rajpoot et al., 2024).

Another approach relies on classical machine learning classification algorithms such as Random Forest, XGBoost and KNN, which have shown less optimal performance in theses tasks due to challenges related to data imbalance (Shetty, 2024) (Atanassova et al., 2024).

3.3. Impact Type

Building upon their successful semi-supervised learning (SSL) approach for predicting ESG impact duration, the 3idiots team (Kim et al., 2024) applied a similar methodology to classify the impact type of ESG-related events on companies. Employing the same finance-specialized pre-trained language model, KF-DeBERTa (jeo, 2023), the team enriched their dataset with additional unlabeled ESG news articles, paralleling their strategy in the impact duration challenge. Through the use of advanced data augmentation techniques, including both weak (Wei and Zou, 2019) and strong augmentations, they effectively leveraged the model's capabilities to capture domain-specific nuances.

4. Performances

4.1. Impact Duration

Table 5 shows the performance of the official evaluation of participants' models.

In Korean Impact Duration, the application of advanced NLP models, notably KF-DeBERTa (jeo, 2023) and XLM-RoBERTa (Conneau et al., 2020), showcased exemplary performance among encoder models such as FinBERT (Araci, 2019), BERT (Devlin et al., 2019), and so on. Particularly, the integration of semi-supervised learning (SSL) (Tarvainen and Valpola, 2018) and diverse augmentation strategies (Wei and Zou, 2019; Lee et al., 2023) played a crucial role, enhancing model robustness and comprehension of ESG-related news articles, thereby leading to superior outcomes in classification tasks. Moreover, a noteworthy innovation was observed from a team (Yun Hyojeong and Son, 2024) employing GPT-4 (OpenAI et al., 2024), which diverged from traditional methodologies by leveraging prompting and dynamic in-context learning without direct model fine-tuning on the provided datasets. This approach highlighted how advanced generative language models can understand and tackle specialized area.

Banerjee et al. (2024)[1] proposed an English translation approach using Google Translate for the Japanese subtask and augmented the translated dataset with a T5-based model. They utilized the

[1]Their team ID is "LIPI."

English			French		
	Micro-F1	Macro-F1		Micro-F1	Macro-F1
Jetsons_1	65.44%	60.90%	kaka_1	63.70%	63.29%
Team Tredence_3	58.09%	57.69%	upaya_1	58.22%	56.78%
LIPI_1	60.29%	56.57%	upaya_2	58.22%	56.69%
Jetsons_2	60.29%	56.51%	Team Tredence_3	54.79%	53.80%
Team Tredence_2	59.56%	56.16%	Team Tredence_2	50.00%	51.06%
DICE_2	55.88%	55.27%	Drocks_1	48.63%	48.81%
DICE_3	58.82%	55.08%	Drocks_2	48.63%	48.70%
Drocks_1	57.35%	55.03%	LIPI_2	48.63%	48.30%
IMNTPU_2	58.82%	55.03%	Team Tredence_1	47.95%	47.56%
DICE_1	55.15%	53.11%	IMNTPU_1	47.26%	47.16%
CompLx_1	60.29%	51.88%	DICE_1	49.32%	44.80%
LIPI_2	58.09%	51.48%	Drocks_3	43.15%	42.90%
LIPI_3	56.62%	51.42%	LIPI_3	41.78%	40.45%
kaka_1	51.47%	51.07%	SamNLP_2	43.15%	38.00%
Drocks_2	53.68%	48.65%	CriticalMinds_2	39.04%	37.96%
upaya_1	54.41%	48.40%	upaya_3	42.47%	37.64%
Team Tredence_1	50.00%	48.10%	SamNLP_1	42.47%	37.63%
MLG-TRDDCPune_3	52.21%	47.78%	IMNTPU_2	37.67%	34.46%
Drocks_3	52.21%	46.41%	LIPI_1	41.10%	26.89%
SamNLP_2	50.74%	46.30%	CriticalMinds_3	36.30%	26.21%
upaya_3	51.47%	46.09%	CriticalMinds_1	36.30%	22.48%
upaya_2	53.68%	45.93%			
SamNLP_1	52.21%	45.24%			
MLG-TRDDCPune_1	49.26%	44.74%			
MLG-TRDDCPune_2	50.00%	43.95%			
FinTwin_1	50.00%	43.55%			
CriticalMinds_1	47.06%	43.16%			
CriticalMinds_3	45.59%	40.64%			
CriticalMinds_2	42.65%	39.59%			
IMNTPU_3	19.12%	17.22%			
IMNTPU_1	18.38%	15.54%			

Table 6: Performance — Impact Level.

pretrained BERT-base multilingual uncased model for content concatenated with the impact type feature and classified it using a linear layer. Kao et al. (2024)[2] also employed the BERT-base multilingual-cased model for the Japanese subtask and augmented the dataset using GPT-3.5-turbo. Shetty (2024) explored the efficacy of various classifiers using the scikit-learn library and demonstrated that the decision tree approach was effective for the Japanese subtask. One reason for the comparative deficiency in performance against the top teams appeared to be their lack of use of state-of-the-art pretrained models such as DeBERTa-v3-xsmall or XLM-RoBERTa.

4.2. Impact Level

Table 6 shows the performance of participants' methods on the impact level task.

4.3. Impact Type

Table 7 shows the results of the impact type task in the Korean dataset.

[2]Their team ID is "IMNTPU."

	Micro-F1	Macro-F1
3idiots_3	84.00%	79.85%
FIT_2	81.50%	76.13%
Team Tredence_2	82.50%	75.95%
3idiots_2	81.50%	73.98%
3idiots_1	80.50%	73.43%
Team Tredence_1	80.00%	73.17%
Team Tredence_3	80.00%	71.76%
FIT_1	78.50%	64.46%
FinNLP_2	79.50%	62.46%
FinNLP_3	79.50%	62.46%
FinNLP_1	79.50%	62.46%
kaka_1	63.00%	55.53%
LIPI_1	64.00%	45.53%

Table 7: Performance — Impact Type

In the MLESG-3 shared task, the approaches to the Korean Impact Type mirrored those of the Korean Impact Duration, leveraging advanced NLP models such as KF-DeBERTa (jeo, 2023) with consistent effectiveness. This parallel strategy was reinforced by the adoption of semi-supervised learning (SSL) (Tarvainen and Valpola, 2018) or data augmentation, enhancing both tasks. Further-

more, the use of GPT-4 (OpenAI et al., 2024) by a team (Yun Hyojeong and Son, 2024) showcased in-context learning and prompting techniques, proving that specialized tasks like Impact Type classification can achieve significant outcomes without conventional fine-tuning.

5. Verifying Virtue — Promise Verification

In the ML-ESG shared tasks series, we focus on analyzing news articles from various countries to understand ESG-related events, thereby dynamically scoring a company's ESG performance based on third-party news. To advance our research, the upcoming shared tasks series will concentrate on the ESG-related promises made by companies. This series will encompass tasks such as (1) identifying ESG-related promises, (2) linking evidence to these promises, (3) determining the type of promise-evidence relationship, and (4) inferring the timing for verifying these promises. Our goal is to continue enhancing our multilingual and cross-country datasets.

For the forthcoming series, participants are encouraged to utilize ML-ESG datasets to improve their task performances. For instance, the dataset from ML-ESG-1 can aid in understanding the types of promises, which is crucial for the promise-evidence type task. Similarly, the ML-ESG-3 dataset can be instrumental in inferring the duration of events, a key factor in the task of verifying timing inference.

6. Conclusion

In the ML-ESG series of shared tasks, we have explored three tasks for dynamically scoring a company's ESG score based on news articles. ML-ESG-3, in particular, introduced the challenge of inferring the duration of impacts. Unlike ESG issue identification (ML-ESG-1) and impact type (ML-ESG-2), the impact duration (ML-ESG-3) is much more subjective, evidenced by low agreements in the annotation results across different languages. The performance in ML-ESG-1 and ML-ESG-2 is significantly better than in ML-ESG-3. Based on participants' findings, we observe that pre-trained LMs and LLMs perform well in well-defined tasks but still face challenges with this kind of subjective task. Thus, one of our suggestions is for ESG scoring companies to share more details about the assessment results of experts' discussions and experiences. This would help make the process more transparent and increase the possibility of models performing the task automatically.

Furthermore, we reveal our plan for the next shared task series, which focuses on multi-lingual ESG promise verification. This future direction promises to further refine our understanding of corporate ESG performance, enhancing transparency and accountability across languages and borders. We hope the ML-ESG task series will contribute to promoting sustainability and equity in the financial sector.

7. Acknowledgments

This work was supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3, NSTC 112-2634-F-002-005 -, and Ministry of Education (MOE) in Taiwan, under grants NTU-112L900901. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The work of Yohei Seki 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).

8. Bibliographical References

2023. *KF-DeBERTa: Financial Domain-specific Pre-trained Language Model*. Korean Institute of Information Scientists and Engineers.

Harika Abburi, Ajay Kumar, Edward Bowen, and Balaji Veeramani. 2024. Multilingual esg news impact identification using ensemble approach. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Iana Atanassova, Marine Potier, Maya Mathie, Marc Bertin, and Panggih Kusuma Ningrum. 2024. Criticalminds: Enhancing ml models for esg impact analysis categorisation using linguistic resources and aspect-based sentiment analysis. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Neelabha Banerjee, Anubhav Sarkar, Swagata Chakraborty, Sohom Ghosh, and Sudip Naskar.

2024. Fine-tuning language models for predicting the impact of events associated to financial news articles. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Konstantinos Bougiatiotis, Andreas Sideras, Elias Zavitsanos, and Georgios Paliouras. 2024. Dice @ ml-esg-3: Esg impact level and duration inference using llms for augmentation and contrastive learning. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023a. Multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115, Macao. -.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023b. Multi-lingual esg impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Parag Pravin Dakle, Alolika Gon, Sihan Zha, Liang Wang, Sai Krishna Rallabandi, and Preethi Raghavan. 2024. Jetsons at finnlp 2024: Towards understanding the esg impact of a news article using transformer-based models. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Yu Han Kao, Vidhya Nataraj, Ting-Chi Wang, Yu-Jyun Zheng, Hsiao-Chuan Liu, Wen-Hsuan Liao, Chia-Tung Tsai, and Min-Yuh Day. 2024. Imntpu at ml-esg-3: Transformer language models for multi-lingual esg impact type and duration classification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Jungdae Kim, Eunkwang Jeon, and Jeon Sang Hyun. 2024. Leveraging semi-supervised learning on a financial-specialized pre-trained language model for multilingual esg impact duration and type classification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. Easyguide : Esg issue identification framework leveraging abilities of generative large language models.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling,

Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Pawan Rajpoot, Ashvini Jindal, and Ankur Parikh. 2024. Adapting llm to multi-lingual esg impact and length prediction using in-context learning and fine-tuning with rationale. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Poorvi Shetty. 2024. Esg impact inference in english, french, korean and japanese. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Antti Tarvainen and Harri Valpola. 2018. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Ke Tian and Hua Chenn. 2024. Esg-gpt: Gpt4-based few-shot prompt learning for multi-lingual esg news text classification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of The 32nd ACM International Conference on Information and Knowledge Management (CIKM'23)*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks.

Weijie Yang and Xinyun Rong. 2024. Duration dynamics: Fin-turbo's rapid route to esg impact insight. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Moonjeong Hahm Kyuri Kim Yun Hyojeong, Chan-Yeong Kim and Guijin Son. 2024. Esg classification by implicit rule learning via gpt-4. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.