Multi-Lingual ESG Impact Type Identification

Chung-Chi Chen,[1] **Yu-Min Tseng,**[2] **Juyeon Kang,**[3] **Anaïs Lhuissier,**[3] **Yohei Seki,**[4]
Min-Yuh Day,[5] **Teng-Tsai Tu**,[6] **Hsin-Hsi Chen**[7]
[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]Graduate Institute of Information Management, National Taipei University, Taiwan
[6]Graduate Institute of International Business, National Taipei University, Taiwan
[7]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

Abstract

Assessing a company's sustainable development goes beyond just financial metrics; the inclusion of environmental, social, and governance (ESG) factors is becoming increasingly vital. The ML-ESG shared task series seeks to pioneer discussions on news-driven ESG ratings, drawing inspiration from the MSCI ESG rating guidelines. In its second edition, ML-ESG-2 emphasizes impact type identification, offering datasets in four languages: Chinese, English, French, and Japanese. Of the 28 teams registered, 8 participated in the official evaluation. This paper presents a comprehensive overview of ML-ESG-2, detailing the dataset specifics and summarizing the performance outcomes of the participating teams.

1 Introduction

In the rapidly shifting global business milieu, relying solely on traditional financial metrics to gauge companies is no longer adequate. The imperative for companies to make meaningful contributions to society and the environment, underpinned by robust governance, has intensified. These principles, often denoted as Environmental, Social, and Governance (ESG) factors, have surged in significance for stakeholders from investors to consumers. Evaluating a company's adherence to these principles is crucial not only for its sustainable growth but also for setting the benchmark for corporate responsibility in the modern era.

The increasing need to precisely appraise and contrast ESG ratings across corporations has ignited extensive research and discourse. The influence of news and contemporaneous events on these ratings offers a captivating subject for study. This dynamic environment necessitates a resilient and malleable approach. While numerous rating systems exist, our focus leans towards the MSCI ESG rating guidelines as the foundation for the Multi-Lingual ESG (ML-ESG) shared task series.

In this backdrop, the ML-ESG shared task series was initiated to cultivate a deeper understanding of news-driven ESG ratings. In a world growing more intertwined, the imperative of incorporating multiple languages in these assessments is paramount. In the inaugural ML-ESG, our focus was on pinpointing ESG issues (Chen et al., 2023). Progressing to the second iteration, ML-ESG-2 hones in on the classification of impact types, featuring datasets in Chinese, English, French, and Japanese.[1] The objective of ML-ESG-2 is discerning whether news pieces indicate risks or opportunities for company operations through an ESG lens.

This paper aims to shed light on the methodologies, datasets, and discoveries of ML-ESG-2. Through the concerted effort of diverse teams, we aspire to offer a comprehensive perspective on the latest advancements in multi-lingual ESG impact type identification. Our endeavor is to equip both scholars and industry professionals with insights into this emerging domain.

2 Dataset and Task Setting

Table 1 presents the dataset statistics for ML-ESG-2. The Chinese (Tseng et al., 2023) and Japanese (Kannan and Seki, 2023) datasets are previously published, while the English and French datasets make their debut in ML-ESG-2.

2.1 Chinese, English, and French

The Chinese dataset is derived from ESG-centric news articles found on ESG-BusinessToday (in Chinese)[2]. Meanwhile, the English and French datasets amalgamate articles from sources including ESGToday (in English)[3], RSEDATANEWS (in

[1]Japanese dataset used different guidelines and data sources from other datasets. Please refer to Section 2.2 for details.

[2]https://esg.businesstoday.com.tw/

[3]https://www.esgtoday.com/category/esg-news/companies/

Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pages 46–50

		English	French	Chinese	Japanese
Train	Opportunity (Positive)	694	458	536	460
	Risk (Negative)	114	360	58	49
	Other	-	-	666	387
Development	Opportunity (Positive)	-	-	60	-
	Risk (Negative)	-	-	6	-
	Other	-	-	74	-
Test	Opportunity (Positive)	191	89	67	115
	Risk (Negative)	27	111	7	13
	Other	-	-	82	97
	Total	1,026	1,018	1,556	1,121

Table 1: Data statistics.

Team	Best Performing Model
LIPI	FinBERT, Translate to English, T5-Based Data Augmentation
231	ChatGPT Summarization. RoBERTa-Chinese, Convert to Simplfied Chinese
FinNLU	FlauBERT, mBERT, ALBERT, TF-IDF features, and LSA features

Table 2: Best performing model proposed by each team for Chinese dataset.

French)[4], and Novethic (in French)[5].

The annotation schemes for the Chinese, English, and French datasets categorize news articles as either opportunities or risks. Within the Chinese dataset, additional delineations are made for articles that do not fit the aforementioned categories: they are classified as "Cannot Distinguish (related to company)", "Related to ESG, but not related to company", and "Not related to ESG topic'.' In total, there are 1,556 instances for the Chinese dataset, 1,026 for the English, and 1,018 for the French.

2.2 Japanese

The Japanese dataset is sourced from EDINET[6], which were published from Financial Services Agency of Japan (JFSA). Unlike the other datasets, the Japanese collection emphasizes company annual reports (called annual securities reports, and known as *Yuuka Shoken Houkokusho* in Japanese).

During the annotation phase, annotators allocated sentiment labels at the sentence level. These labels encompassed categories such as "Positive," "Negative," "Neutral," and "None" (indicating the sentence's irrelevance to ESG topics).

In delving deeper into the nuances of these annotations, we observed that the majority of sentences labeled as "positive" or "negative" resonate with the "opportunity" or "risk" ESG impact types, respectively. Interestingly, some of sentences deemed "neutral" also aligned with the "opportunity" impact type. Our examination of the inter-annotator agreement, gauged via Cohen's κ coefficient between the type of ESG impact and sentiment annotations, yielded a result of 0.980, signifying an almost perfect agreement.

For the Japanese analysis, a total of 1,121 instances are available. For an exhaustive description and analysis, we direct the reader to Kannan and Seki (2023).

3 Method

3.1 English and French

The participants bring forth a variety of methodologies, including pre-trained transformers, fine-tuning, prompt engineering, and ensemble learning, demonstrating a diverse set of approaches. Most participant systems underscore the significance of mitigating class imbalances through various data augmentation techniques, notably through translation. The translation of data from French, Japanese, or Chinese into English serves to enhance model performance by augmenting the sample size. Furthermore, we observe that applying the same architectural pipeline across all four languages yields successful results for some languages but not for others. This implies the need to consider more language-specific resources or models to improve overall performance. Qiu et al. (2023) and Vardhan et al. (2023) use pre-trained transformer models, incorporating various data and feature augmentation techniques to enhance performance, including translation, summarization, and data paraphrasing. Polyanskaya and Brillet (2023) demonstrates the superiority of GPT-3.5 Turbo over BERT for En-

[4] https://www.rsedatanews.net/

[5] https://www.novethic.fr/actualite/environnement.html

[6] https://disclosure2.edinet-fsa.go.jp/WEEK0010.aspx

glish dataset which was enhanced by the translated French dataset. (Winatmoko and Septiandri, 2023) explores ST5 and SBERT for generating embeddings and in Mishra (2023), fine-tuning Llama2 on the English dataset with prompts detailing the classification criteria gives the best result, on both the English and French datasets. The Veeramani et al. (2023) introduces an ensemble learning method with mBERT, FlauBERT, ALBERT, and MLP models, incorporating feature representations (LSA and TF-IDF), demonstrating superior performance with early fusion ensemble across all four languages. Billert and Conrad (2023) describes an adapter-based framework designed to enhance the capture of ESG-aspect-specific knowledge and language-specific knowledge present in the training data.

3.2 Chinese and Japanese

In Table 2, the approaches proposed by different teams for the Chinese dataset are outlined. Team 231 (Qiu et al., 2023) suggests converting Traditional Chinese to Simplified Chinese and employing the summarization of ChatGPT as input, rather than the entire news article. On the other hand, Team LIPI (Vardhan et al., 2023), translates the entire dataset into English and uses a T5-based model[7] for paraphrasing the data. Meanwhile, Team FinNLU (Veeramani et al., 2023) integrates four embeddings from various language models, combined with TF-IDF and LSA features, for their predictions.

LIPI and FinNLU also joined the Japanese subtask with the proposed methods, and SPEvFT (Mishra, 2023) applies prompt engineering to Japanese articles.

4 Experimental Results

Tables 3, 4, 5, and 6 report the performances of ML-ESG-2 participants on English, French, Chinese, and Japanese datasets, respectively.

Drawing from the outcomes of both English and French datasets, it becomes evident that fine-tuning RoBERTa using these datasets produces the most optimal results, as documented in AnakItik (Winatmoko and Septiandri, 2023). When juxtaposed with the outcomes of fine-tuning Llama-2 (SPEvFT) (Winatmoko and Septiandri, 2023), there is a pronounced disparity in performance for English, though this difference is less marked for French.

Submission	Micro-F1	Macro-F1	Weighted-F1
AnakItik_English_2	0.9817	0.9548	0.9810
BrothFink_English_3	0.9771	0.9445	0.9765
NeverCareU_English_2	0.9633	0.9227	0.9648
FinNLU_English_1	0.9633	0.9180	0.9639
231_English_3	0.9633	0.9127	0.9627
SPEvFT_English_3 (Late)	0.9587	0.9118	0.9602
231_English_1	0.9633	0.9096	0.9620
231_English_2	0.9633	0.9096	0.9620
BrothFink_English_2	0.9541	0.8870	0.9525
BrothFink_English_1	0.9450	0.8645	0.9430
AnakItik_English_1	0.9220	0.8537	0.9289
LIPI_English_2	0.9312	0.8335	0.9294
NeverCareU_English_1	0.9312	0.8211	0.9267
LIPI_English_3	0.9266	0.8127	0.9226
HHU_English_1	0.9174	0.8098	0.9174
HHU_English_3	0.9174	0.8098	0.9174
LIPI_English_1	0.9083	0.7741	0.9051
AnakItik_English_3	0.9083	0.6246	0.9495
SPEvFT_English_1	0.9174	0.5574	0.9256
SPEvFT_English_2	0.8716	0.4657	0.8160
HHU_English_2	0.4908	0.4225	0.5719

Table 3: Results in English dataset.

Submission	Micro-F1	Macro-F1	Weighted-F1
SPEvFT_French_3 (Late)	0.8700	0.8661	0.8686
AnakItik_French_2	0.8550	0.8547	0.8554
AnakItik_French_1	0.8400	0.8368	0.8393
HHU_French_1	0.7550	0.7548	0.7555
LIPI_French_2	0.7550	0.7547	0.7556
HHU_French_3	0.7500	0.7457	0.7493
LIPI_French_3	0.7200	0.7182	0.7157
LIPI_French_1	0.7100	0.7090	0.7109
HHU_French_2	0.6250	0.6169	0.6231
AnakItik_French_3	0.7500	0.5545	0.8310
FinNLU_French_1	0.5500	0.5292	0.5184
SPEvFT_French_1	0.7100	0.4918	0.7367
SPEvFT_French_2	0.4450	0.3080	0.2741

Table 4: Results in French dataset.

In the evaluation of the Chinese dataset, Team LIPI (Vardhan et al., 2023) achieved the highest performance during the formal evaluation. Their strategy employed all datasets available in ML-ESG-2 to facilitate data augmentation. The predictions from their model are generated using FinBERT with English inputs. A noteworthy observation from their study is the superior performance achieved using translated inputs compared to the original data. Subsequent to the formal evaluation, Team 231 (Qiu et al., 2023) embarked on further exploration. Their findings indicate that a simple conversion from Traditional Chinese to Simplified Chinese can enhance performance. Furthermore, they observed that utilizing a summary from ChatGPT yields better results than employing the complete news content. However, there were contrasting findings among the participants: while Team 231's experiments suggest that TF-IDF fea-

[7] https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base

Submission	Micro-F1	Macro-F1	Weighted-F1
LIPI_Chinese_3	0.6859	0.5279	0.6773
LIPI_Chinese_2	0.7564	0.4585	0.7321
LIPI_Chinese_1	0.6731	0.2897	0.6508
231_Chinese_1	0.3718	0.1853	0.3725
231_Chinese_3	0.3654	0.1833	0.3593
231_Chinese_2	0.3590	0.1792	0.3593
FinNLU_Chinese_1	0.4103	0.1728	0.3881

Table 5: Results in Chinese dataset.

Submission	Micro-F1	Macro-F1	Weighted-F1
LIPI_Japanese_2	0.6889	0.6340	0.6786
LIPI_Japanese_1	0.6400	0.5436	0.6242
LIPI_Japanese_3	0.6222	0.5366	0.6033
SPEvFT_Japanese_1	0.4800	0.3792	0.4776
FinNLU_Japanese_1	0.5378	0.3043	0.4943

Table 6: Results in Japanese dataset.

tures offer limited utility, Team FinNLU's results underscore their significance.

In the evaluation of the Japanese dataset, Team LIPI's performance emerged superior compared to other participating teams. This notable advantage is attributed to their T5-based paraphrasing module, which demonstrably bolstered the system's efficacy. Notably, despite the evident disparity in document genres — transitioning from news to annual securities reports — their methodology retained its effectiveness. This suggests that certain topics, especially those pertinent to "opportunity" (positive) and "risk" (negative) annotations, exhibit shared characteristics across different document genres and across languages to some degree. Team FinNLU adopted an early ensemble fusion approach, which proved to be equally effective for the Japanese context. The inclusion of LSA features also enhanced their performance metrics, mirroring the positive outcomes observed in other languages. On the other hand, Team SPEvFT explored a prompt engineering approach tailored to the Japanese language. They employed a sentence similarity method with training examples, revealing its efficacy, albeit to a limited extent.

5 Conclusion

In this study, we provided an overview of the ML-ESG-2 shared task featured at FinNLP@IJCAI-2023. Our findings indicate that for the English and French datasets with paragraph-based annotations, the top-performing models were fine-tuned RoBERTa or Llama-2. Conversely, for the Chinese and Japanese datasets that used article-based and sentence-based structures, translating the content into English yielded commendable outcomes. Such findings bring forth pivotal questions. Our investigation into baseline models suggests that their adeptness, even in their simplicity, might surpass certain specialized methods. Additionally, it hints at the possibility that massive datasets might not be essential for fine-tuning (large) language models in ESG impact type identification, particularly for languages like English and French.

Having delved into ESG issue identification and ESG impact type identification tasks, our next venture will be into ESG impact duration inference under ML-ESG-3. Through the series of ML-ESG shared tasks, we pave the way for a holistic approach to dynamic ESG ratings based on news articles.

Acknowledgments

This research is supported by National Science and Technology Council, Taiwan, under grants 110-2221-E-002-128-MY3, 110-2634-F-002-050-, and 111-2634-F-002-023-. 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).

References

Fabian Billert and Stefan Conrad. 2023. Exploring knowledge composition for esg impact type determination. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for esg scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.

Soumya Mishra. 2023. Predicting esg impact types of multi-lingual news articles: Leveraging strategic

prompt engineering and llm fine-tuning. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Anna Polyanskaya and Lucas Fernández Brillet. 2023. Gpt-based solution for esg impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Le Qiu, Bo Peng, Jinghang Gu, Yu-Yin Hsu, and Emmanuele Chersoni. 2023. Identifying esg impact with key information. In *Proceedings of the Sixth Workshop on Financial Technology 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)*.

Harsha Vardhan, Sohom Ghosh, Ponnurangam Kumaraguru, and Sudip Naskar. 2023. A low resource framework for multi-lingual esg impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023. Enhancing esg impact type identification through early fusion and multilingual models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Yosef Ardhito Winatmoko and Ali Septiandri. 2023. The risk and opportunity of data augmentation and translation for esg news impact identification with language models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.