Multi-Lingual ESG Issue Identification

Chung-Chi Chen,[1] **Yu-Min Tseng,**[2] **Juyeon Kang,**[3] **Anaïs Lhuissier,**[3]
Min-Yuh Day,[4] **Teng-Tsai Tu**,[5] **Hsin-Hsi Chen**[2]
[1]AIST, Japan
[2]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[3]3DS Outscale (ex Fortia), France
[4]Graduate Institute of Information Management, National Taipei University, Taiwan
[5]Graduate Institute of International Business, National Taipei University, Taiwan

Abstract

This paper introduces an innovative approach for incorporating environmental, social, and governance (ESG) factors into AI-based financial decision-making processes. Recent developments in AI and NLP have predominantly focused on financial outcomes, often disregarding the significant impacts that corporations can have on society and the environment. This perspective overlooks potential business risks associated with environmental and social issues. We propose a task, the Multilingual ESG Issue Identification Task (ML-ESG), that seeks to integrate the ESG paradigm into financial NLP systems. The ML-ESG is designed according to the MSCI ESG rating methodology and requires systems to classify news articles into 35 key ESG issues. Moreover, systems must identify the target company and its industry, as the weighting of each issue varies accordingly. This paper presents an overview of the ML-ESG shared task, implemented as part of the FinNLP-2023 workshop, detailing the datasets, methods, and participant performances.

1 Introduction

Finance often brings to mind a world dominated by monetary transactions and market forecasts. The environmental and social implications of investment decisions, significant factors in today's business environment, have been largely overlooked in machine learning models. For instance, even in scenarios where a corporation is reported to be engaged in environmentally harmful practices, such as improper waste disposal, AI models may still recommend purchasing the corporation's stock following a market overreaction to the news. Such decisions, while potentially profitable in the short-term, can lack foresight into potential long-term risks associated with the corporation's practices.

To address this concern, we introduce the concept of ESG (environmental, social, and governance) into our shared task, aiming to help AI models consider the broader impacts of investment decisions. By integrating insights from the financial domain into NLP research, we hope to promote long-term, value-driven investments that also account for non-monetary factors like environmental and social impacts.

The ESG concept, initially proposed by the UN Global Compact in 2005, has gained increasing attention over the past few years, particularly since 2020. The idea of ESG has matured over time, with a growing body of research analyzing and evaluating these non-monetary factors (Amel-Zadeh and Serafeim, 2018; Matos, 2020). In last year's FinNLP workshop, we proposed the first step towards integrating ESG considerations into NLP with the FinSim-2022 task, which focused on learning semantic similarities. This task aimed to classify given words into ESG-related taxonomies and sentences into sustainable or unsustainable descriptions, thereby evaluating models' understanding of ESG narratives.

Building on this foundation, this year's FinNLP workshop presents a more detailed task: the Multilingual ESG Issue Identification Task (ML-ESG). This task is designed according to the MSCI ESG rating methodology and requires systems to classify news articles into 35 key ESG issues, as depicted in Figure 1. The ESG Industry Materiality Map provides these weights thus, the system's primary task is to identify the topic. In this shared task, we offer multilingual datasets (English, Chinese, French) to identify ESG issues in news articles. This paper provides an overview of the ML-ESG shared task in the FinNLP-2023 workshop, detailing the dataset, participant methods, and performances. Twenty-seven teams registered, ten of which submitted their system outputs for the official evaluation.

2 Dataset and Task Setting

This section outlines the composition of our proposed datasets and elucidates the corresponding

Proceedings of the Joint Workshop of the 5th Financial Technology and Natural Language Processing (FinNLP) and 2nd Multimodal AI For Financial Forecasting (Muffin), Macao, August 20, 2023.

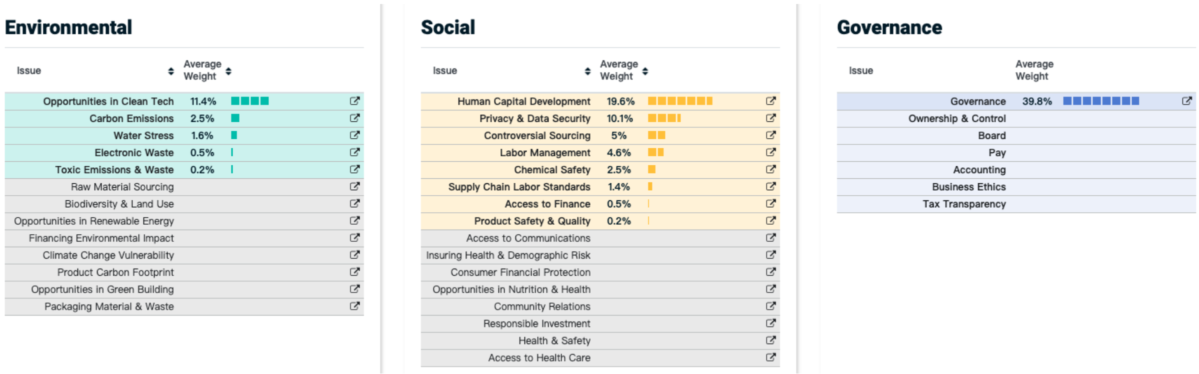

Figure 1: List of the ESG issues and examples of the weighting. This is the screenshot of the ESG Industry Materiality Map.

	English	French	Chinese
Train	1,199	1,200	900
Development	-	-	100
Test	300	300	238
Total	1,499	1,500	1,238

Table 1: Statistics of Datasets

task settings, as depicted in Table 1.

2.1 English and French Datasets

The English and French datasets are collected from ESG-related news articles acquired from ESGToday (English)[1], RSEDATANEWS (French)[2], and Novethic (French)[3]. Given a news article, annotators are asked to select the related issues from the 35 pre-defined ESG Key Issues by MSCI[4], and then label it with the most relevant issues. The English and French datasets are annotated by experts (2 annotators and 1 reviewer) in Fortia's Data & Language Analyst team.

Many events comprise multiple components, including various pillars (e.g., Environment + Social), themes within the same pillar (e.g., Environment > Natural Capital + Pollution & Waste), or even within the same theme (e.g., Environment > Pollution & Waste > Toxic Emissions & Electronic Waste). Although key issues are clearly defined to establish boundaries between somewhat similar themes, real-life events are not always so clear-cut.

For that reason, we have chosen to divide a news article into multiple paragraphs based on the topic. In both the English and French task settings, the objective is to predict one of the ESG issues based on a specific paragraph extracted from a news article.

2.2 Chinese Dataset

Our Chinese dataset is sourced from ESG-related news articles available on ESG-BusinessToday (Chinese)[5]. Seven postgraduate students from the Graduate Institute of Information Management at National Taipei University undertake the annotation of this dataset. To maintain consistency and accuracy in annotation, we organize bi-weekly meetings to address arising issues and ensure a consensus on the guidelines and labels.

Given that Chinese news articles are annotated on an article-based framework, each article may pertain to more than one ESG issue, which calls for a multi-label task setting in the Chinese dataset.

Furthermore, we noted that some articles on the ESG news platform do not truly align with ESG or ESG scoring principles. To account for this discrepancy, we have included an additional label to identify articles that are not related to ESG.

To gain a more comprehensive understanding of ESG issues, we have merged the SASB Standard with MSCI's guidelines, which has yielded 44 issues.[6]

3 Methods

3.1 French and English

Exploring diverse BERT language model strategies, such as SVM (Cortes and Vapnik, 1995) with

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

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

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

[4] https://www.msci.com/our-solutions/esg-investing/esg-ratings/esg-ratings-key-issue-framework

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

[6] For a more detailed definition, please refer to the following document: https://drive.google.com/file/d/12ia_CF3nrjv_R8s_e44SLnZnNcHH-D0_/view?usp=sharing

Submission	Precision	Recall	F1-Score
NCMU_English_1	0.69	0.70	0.69
TradingCentralLabs_English_1	0.67	0.68	0.67
NCMU_English_2	0.68	0.66	0.66
kaka-ML-ESG_English_Test_gpt	0.67	0.67	0.65
Jetsons_English_1	0.64	0.65	0.64
Jetsons_English_2	0.63	0.64	0.63
LASTI_English_2	0.64	0.63	0.63
NCMU_English_3	0.65	0.63	0.63
HKESG_English_3	0.63	0.63	0.62
Jetsons_English_3	0.63	0.64	0.62
kaka-ML-ESG_English_Test_word2vec_tfidf	0.62	0.63	0.61
LASTI_English_3	0.62	0.62	0.61
TradingCentralLabs_English_2	0.61	0.63	0.61
HKESG_English_1	0.61	0.62	0.60
kaka-ML-ESG_English_Test_roberta	0.62	0.62	0.60
LASTI_English_1	0.61	0.60	0.60
HKESG_English_2	0.59	0.59	0.58
TradingCentralLabs_English_3	0.59	0.59	0.58
HHU_English_3	0.60	0.58	0.57
HHU_English_1	0.55	0.59	0.56
HHU_English_2	0.42	0.36	0.35
LivermoreSXI_English_1	0.36	0.33	0.30
wwy_test_English_1	0.28	0.37	0.30

Table 2: Experimental results in English Dataset.

SBERT embeddings (Reimers and Gurevych, 2019) and RoBERTa, Linhares Pontes et al. (2023) conduct experiments on monolingual and multilingual data. Their findings reveal that RoBERTa performs best on monolingual data for the English dataset, while on the French dataset, RoBERTa excels on multilingual data, achieving superior results. Glenn et al. (2023) generate synthetic data using a large language model - gpt-3.5-turbo - in order to augment the training data which is then used to fine-tune the multilingual BERT for classification. Hanwool et al. (2023) use generative models like Pythia (Biderman et al., 2023), CerebrasGPT (Dey et al., 2023), and OPT (Zhang et al., 2022), along with the zero-shot (Xian et al., 2017), GPT3Mix (Yoo et al., 2021) and translation as augmentation techniques to tackle the data imbalance issue; then, explore encoder models, RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and FinBERT (Araci and Genç, 2020). Mashkin and Chersoni experiment with ESG Transformers (Mukut, 2020), and for classification, Logistic Regression, Random Forests and Support Vector Machine achieving the best results with SVM classifier for both languages. Billert and Conrad introduce adapter modules (Houlsby et al., 2019) to a multilingual base model, mBERT (Devlin et al., 2019), then train it using Masked Language Modeling (MLM) (Pfeiffer et al., 2020).

3.2 Chinese

Wang et al. (2023) leverage MacBERT (Cui et al., 2020)—a contrastive learning framework—enhancing performance using both unlabeled and pseudo-labeled data. Linhares Pontes et al. (2023) explores the performance of SVM (Cortes and Vapnik, 1995) when combined with SentenceBERT's embeddings (Reimers and Gurevych, 2019) (SBERT). Additionally, Glenn et al. (2023) outlines a method for utilizing synthetic data generated by a large language model, ChatGPT,[7] to enhance the performance of multilingual BERT (mBERT).

4 Results

Performance metrics, including precision, recall, and F1-score, were utilized to evaluate the English and French datasets. Given the distinctive task settings of the Chinese dataset, micro-averaged F1, macro-averaged F1, and weighted F1 were adopted for evaluation. Tables 2, 3, and 4 display the experimental results from the participants' system outputs in the official evaluation round.

We find that BERT-like language models with data augmentation by LLMs perform well for the English and French results. NCMU (Hanwool et al., 2023) ranks first and second in these two datasets. Jetsons (Glenn et al., 2023) also uses

[7] gpt-3.5-turbo: https://platform.openai.com/docs/models/gpt-3-5

Submission	Precision	Recall	F1-Score
Jetsons_French_2	0.80	0.79	0.78
NCMU_French_1	0.80	0.79	0.78
HHU_French_3	0.80	0.77	0.77
Jetsons_French_1	0.78	0.78	0.77
HHU_French_1	0.78	0.75	0.75
TradingCentralLabs_French_2	0.76	0.76	0.75
kaka-ML-ESG_French_Test_gpt	0.75	0.75	0.74
HHU_French_2	0.76	0.74	0.73
TradingCentralLabs_French_3	0.74	0.74	0.73
HKESG_French_3	0.72	0.72	0.71
TradingCentralLabs_French_1	0.73	0.72	0.71
Jetsons_French_3	0.70	0.71	0.70
NCMU_French_2	0.71	0.70	0.69
HKESG_French_1	0.69	0.68	0.67
HKESG_French_2	0.65	0.62	0.62
kaka-ML-ESG_French_Test_word2vec_tfidf	0.62	0.61	0.60
LASTI_French_1	0.60	0.59	0.59
LASTI_French_2	0.61	0.60	0.59
LASTI_French_3	0.56	0.56	0.55
LivermoreSXI_French_1	0.32	0.33	0.28
kaka-ML-ESG_French_Test_roberta	0.16	0.25	0.18

Table 3: Experimental results in French Dataset.

Submission	Micro F1	Macro F1	Weighted F1
CheryFS_Chinese_2 (Wang et al., 2023)	0.391	0.180	0.392
TradingCentralLabs_Chinese_3 (Linhares Pontes et al., 2023)	0.279	0.137	0.263
TradingCentralLabs_Chinese_2 (Linhares Pontes et al., 2023)	0.267	0.103	0.233
TradingCentralLabs_Chinese_1 (Linhares Pontes et al., 2023)	0.212	0.073	0.179
Jetsons_Chinese_1 (Glenn et al., 2023)	0.134	0.042	0.102
Jetsons_Chinese_3 (Glenn et al., 2023)	0.134	0.042	0.102
Jetsons_Chinese_2 (Glenn et al., 2023)	0.121	0.038	0.091
CheryFS_Chinese_1 (Wang et al., 2023)	0.089	0.074	0.123

Table 4: Experimental results in Chinese Dataset.

synthetical data to get the best performance in the French dataset.

For the Chinese dataset, the performance is lower due to the multiple-label task setting. The MacBERT with data augmentation method proposed by Wang et al. (2023) gets the best performances.

5 Conclusion

This paper presents the findings of the ML-ESG shared task and highlights the impact of data augmentation methods on performance, regardless of the language employed. It is worth noting, however, that the effectiveness of data generated by LLMs may not always yield favorable outcomes. Selecting the optimal LLM for data augmentation remains an unresolved challenge, with participants opting for a practical approach of utilizing data generated by diverse augmentation methods. Moving forward, our next objective within the ML-ESG initiative is to determine whether a given news event can be classified as an opportunity or risk within the realm of ESG considerations.

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.

References

Amir Amel-Zadeh and George Serafeim. 2018. Why and how investors use esg information: Evidence from a global survey. *Financial Analysts Journal*, 74(3):87–103.

Dogu Araci and Zülküf Genç. 2020. Financial sentiment analysis with pre-trained language models.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

Fabian Billert and Stefan Conrad. 2023. Team hhu at the finnlp-2023 ml-esg task: A multi-model approach to esg-key-issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. 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.

Nolan Dey, Gurpreet Gosal, Zhiming, Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. 2023. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the finnlp-2023: Using synthetic data and transfer learning for multilingual esg issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.

Lee Hanwool, Choi Jonghyun, Kwon Sohyeon, and Jung Sungbum. 2023. Easyguide : Esg issue identification framework leveraging abilities of generative large language models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging bert language models for 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)*.

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.

Ivan Mashkin and Emmanuele Chersoni. 2023. Hkesg at the ml-esg task: Exploring transformer representations for multilingual 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)*.

Pedro Matos. 2020. Esg and responsible institutional investing around the world: A critical review.

Mukherjee Mukut. 2020. Esg-bert: Nlp meets sus- tainable investing. In *Towards Data Science Blog*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Weiwei Wang, Wenyang Wei, Qingyuan Song, and Yansong Wang. 2023. Leveraging contrastive learning with bert for 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)*.

Yongqin Xian, Christoph Lampert, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.