Evaluating Multilingual Language Models for Cross-Lingual ESG Issue Identification

Wing Yan Li[1,2], Emmanuele Chersoni[2], Cindy Sing Bik Ngai[2]
University of Sussex[1], The Hong Kong Polytechnic University[2]
justina.li.106@gmail.com {emmanuele.chersoni,cindy.sb.ngai}@polyu.edu.hk

Abstract

The automation of information extraction from ESG reports has recently become a topic of increasing interest in the Natural Language Processing community. While such information is highly relevant for socially responsible investments, identifying the specific issues discussed in a corporate social responsibility report is one of the first steps in an information extraction pipeline. In this paper, we evaluate methods for tackling the Multilingual Environmental, Social and Governance (ESG) Issue Identification Task. Our experiments use existing datasets in English, French and Chinese with a unified label set. Leveraging multilingual language models, we compare two approaches that are commonly adopted for the given task: off-the-shelf and fine-tuning. We show that fine-tuning models end-to-end is more robust than off-the-shelf methods. Additionally, translating text into the same language has negligible performance benefits.

Keywords: ESG Reports, Pre-trained Language Models, Cross-lingual Transfer, Text Classification, Multilingual NLP

1. Introduction

Yearly releases of Environmental, Social and Governance (ESG) reports represent an important part of a financial company's life cycle. Such reports are used to guide the decisions of responsible investors, by guaranteeing that the company satisfies measurable and objective criteria that have a positive impact on society (Van Marrewijk, 2003; Sheehy and Farneti, 2021; Serafeim and Yoon, 2022). Complying with ESG practices is a requirement for corporations, for example, SEC filings in the United States have to follow the standard for Climate Change and Human Governance, and every European company providing investment products must disclose how its economic activity aligns with sustainability norms (Kang et al., 2022).

ESG reports address various issues and correspond to labels in internationally-defined standards[1]. Modern language models (LMs) can potentially play an important role in processing such reports by extracting ESG-relevant sections and automating the analysis of sustainability aspects emphasised by a company.

In this work, we propose a comprehensive evaluation of the task of multilingual ESG issue identification, using existing datasets that are written in English (EN), French (FR) and Chinese (ZH) (Chen et al., 2023a) but unifying them in a single label space and treating the labels as non-mutually exclusive (*multi-label* classification). We evaluate two commonly-used approaches to the task, namely off-the-shelf (embedding-based classification) and fine-tuning. In the former, a conventional classifier such as a Support Vector Machine (SVM), is trained on representations encoded by a pre-trained LM; in the latter, a pre-trained LM is fine-tuned end-to-end on the given task.

Using multilingual LMs, we compare the two aforementioned methods. Additionally, we test a translation-based approach by translating the FR and ZH datasets into English, the most resource-rich language. Our evaluation shows that fine-tuning is more robust than training with off-the-shelf representations, and that translation has a limited effect on model performance. We will also release our code and data, in order to allow other researchers to evaluate ESG issue identification systems in a unified multilingual setting[2].

2. Related Work

Recently, the Natural Language Processing (NLP) community has increased interest in automating the identification of issues in ESG reports where these issues are organised into taxonomies.

A dedicated workshop has been organized in conjunction with LREC 2022 (Wan and Huang, 2022), and related shared tasks are regular events in financial NLP workshops such as the FinNLP workshop series (Kang et al., 2022; Chen et al., 2023a,c). In particular, the organisers of the shared task co-located with the FinNLP IJCAI workshop 2023 have made available a multilingual dataset for English, French and Chinese. They

[1] https://www.msci.com/esg-and-climate-methodologies.

[2] https://github.com/justinaL/ML-ESG-Eval

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 50–58
20 May, 2024.

were annotated with labels defined on the basis of the MSCI ESG standard rating guidelines.

While the English and French datasets are fully comparable, the Chinese dataset includes additional labels and exhibits variations in the naming of the common label set. Moreover, in the Chinese dataset, the labels are not mutually exclusive, making it difficult to experiment with Chinese in a multilingual setting.

In the shared task, given the relatively limited size of the data, augmentation approaches relying on ChatGPT (OpenAI, 2022) to generate new instances were the most successful (Glenn et al., 2023), together with methods combining traditional classifiers (e.g. SVMs) and multilingual sentence representations (Linhares Pontes et al., 2023).

A recent and highly-relevant research trend in NLP involves the *domain adaptation* of LMs to specific domains. For instance, FinBERT models (Araci, 2019; Yang et al., 2020a) trained specifically on the financial domain and ESG-BERT models (Mukherjee, 2020; Mehra et al., 2022) trained on ESG reports in the sustainability investing field.

Essentially, these models further pre-train a general-purpose LM such as BERT (Devlin et al., 2019), on an in-domain corpus (e.g. ESG reports). Then, the domain-adapted LM is fine-tuned on the given issue identification task. Some of these models "inherit" a general-domain vocabulary from the original architecture, while others create a new in-domain vocabulary from scratch. This choice has been shown to significantly affect performance on several tasks (Peng et al., 2021, 2022).

However, current ESG-adapted models are limited to the English language. While translation is a common approach to re-adapt monolingual models to other languages, Mashkin and Chersoni (2023) showed that this approach is not significantly better than simpler classifier baselines.

3. Experiments

In this paper, we frame the Multilingual ESG Issue Identification (ML-ESG) task as a multi-label classification task. The task assigns instances (ESG-related news articles) to non-exclusive labels (ESG key issues categories), while not constraining the number of categories per instance. Given the multilinguality (EN, FR and ZH) of the datasets, the investigation is conducted with multilingual encoders where representations of various languages are mapped into a shared semantic space.

3.1. Dataset

The dataset is obtained from the ML-ESG task of FinNLP-2023, containing ESG-related news articles. According to Chen et al. (2023a), the articles were sourced from ESGToday[3] (EN), RSE-DATANEWS (FR)[4], Novethic (FR)[5] and ESG-BusinessToday (ZH)[6]. ESG-BusinessToday is a Taiwanese website, where every article is written in traditional Chinese. The articles are annotated by human experts following the MSCI ESG rating guidelines and are categorised into 35 pre-defined ESG key issues across three main topics: Environment, Social and Corporate Governance. Table 1 shows the sample size of each dataset split.

Language	Train	Test
EN	1199	300
FR	1200	300
ZH	653	131

Table 1: Sample size of dataset splits.

From the table, the ZH dataset is shown to possess the smallest sample size compared to EN and FR. During the annotation process of ZH articles, the SASB Standard[7] are merged with the original MSCI guidelines. As a result, there are extra labels in the original ZH dataset. We identify similarities between the ZH labels and those of the other two languages. Additionally, we re-analyse the set of labels of the ZH dataset, mapping missing labels to the corresponding ones in the shared set. For the labels without close correspondences, we discard the corresponding instances. Details on the label mappings are provided in Appendix B. Given that many of the ZH governance labels cannot be mapped to the shared set, the final dataset has unfortunately a limited number of governance-related labels. This is a problem that will have to be addressed by future studies, as governance labels are particularly relevant for Chinese ESG reports.

In the original task, the labels for the EN and FR datasets are mutually exclusive, while multiple labels can be assigned to the ZH instances. To facilitate cross-lingual learning, we unify the label space of all languages during data pre-processing. We treat each task as a multi-label classification by binarising the labels in every dataset. That is, given a dataset instance, the model has to carry out a binary classification for every possible label.

We focus on the actual multilingual identification of ESG issues, by unifying the task and dataset across languages. Our results are not directly comparable to Chen et al. (2023a) due to: i)

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

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

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

[6] 5https://esg.businesstoday.com.tw/.

[7] https://sasb.org/standards/materiality-finder/?lang=en-us

after mapping the labels and filtering the instances, the ZH dataset is no longer the same; ii) the task on EN and FR is different. We do not assume the labels of the CSR reports to be mutually exclusive for the purpose of uniformity with the ZH data.

3.2. Implementation Details

Leveraging multilingual large LMs, we compare two popular approaches in tackling the 2023 ML-ESG shared task: off-the-shelf and fine-tuning.

Off-the-shelf. Representations are derived from the encoder and passed to a classifier for the issue identification task. We use Support Vector Machine (SVM) as the classifier. Since SVM is designed for binary classification, we utilise the `MultiOutputClassifier` from scikit-learn that fits one SVM per target, extending SVM to support multi-label classification.

Hyper-parameters of the SVM are optimised with Bayesian optimisation[8]. For the optimisation process, we apply a 5-fold stratified sampling and constrain the search space to the following hyper-parameters: `C, gamma, degree` and `kernel`.

Fine-tuning. While off-the-shelf approaches require less training data and parameters for optimisation, they often underutilise the model capacity of the encoders. To address this, we also fine-tune the encoder on the given task. Encoders are fine-tuned end-to-end with a classification layer stacked on top. The weights of encoder and stacked classifier are updated during training. Given the small dataset size, we utilize dropout to prevent over-fitting (Srivastava et al., 2014). Further training details are provided in Appendix A.

Translation. Given that LMs are typically trained on a larger share of English data, it may be advantageous to translate other languages to English before fitting the data. To analyse the impact of translation, we re-run our models with the two aforementioned approaches after translating the FR and ZH datasets to English using Google Translate[9] and DeepL Translate[10].

3.2.1. Encoders

We leverage the following encoders: Sentence-BERT (Reimers and Gurevych, 2019) with distilled multilingual Universal Sentence Encoder (Yang et al., 2020b) (SBERT-DUSE) as the base model, a pre-trained multilingual BERT (mBERT) (Devlin et al., 2019) and a multilingual E5 model (mE5) (Wang et al., 2024b).

SBERT-DUSE follows the SBERT framework by training DUSE on the Stanford Natural Language Inference Corpus (SNLI) (Bowman et al., 2015) and Multi-Genre Natural Language Inference Corpus (MultiNLI) (Williams et al., 2018) for better sentence representations. The sentence encodings are obtained by mean pooling of all the vectors of the final layer.

mBERT shares the same structure as BERT with 12 transformer encoder layers in the base version. The model is pre-trained on Wikipedia pages of 104 languages instead of monolingual English data only. mBERT uses the same pre-training objectives as BERT, namely masked language modelling (MLM) and next sentence prediction (NSP).

mE5 is a variation of the E5 model with XLM-R (Conneau et al., 2020) as the base model. E5 is a general-purpose encoder that aims to yield robust off-the-shelf representations in both zero-shot or fine-tuned settings (Wang et al., 2022). Following the training recipe of the English E5, mE5 is trained using a contrastive loss with weak supervision, leveraging data from Wang et al. (2024a).

For the off-the-shelf approach, sentence representations of mBERT and mE5 are mean pooled vectors of the final layer as done by SBERT.

3.2.2. Evaluation Metric

Macro-F1 scores are used as the performance metric. Given the highly imbalanced classes, the macro score treats each class equally regardless of the number of samples. Thus, the model has to perform well in both majority and minority classes. The class distribution is plotted in Appendix B.

4. Results

Table 2 and 3 are the results with and without translation applied. *EU Lang.* refers to training data including EN and FR only; *All Lang.* indicates that training data from all languages are used. Compared to the results of the original shared task (Chen et al., 2023a), the scores for EN and FR are lower, but this is not surprising, since we are working in a multi-label classification setting.

In Table 2, a first noticeable trend is that models using *All Lang.* have significant improvements on ZH compared to those using *EU Lang.* only. While performance on EN and FR drop observably for SBERT-DUSE and mE5, this is not the case for mBERT, which shows more robust performance. This suggests that using multilingual data is, as expected, very helpful for languages in a low-resource setting for this task. However, the per-

[8] https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html

[9] https://translate.google.com/

[10] https://www.deepl.com/en/docs-api

Encoder	Lang.	EN	FR	ZH
SBERT-DUSE	EU	**0.52**	**0.67**	0.06
	All	0.45	0.59	**0.18**
mBERT	EU	**0.48**	**0.47**	0.04
	All	**0.48**	0.46	**0.22**
mE5	EU	**0.53**	**0.60**	0.09
	All	0.48	0.54	**0.20**

(a) Off-the-shelf approach. SVM as the classifier with Bayesian optimisation.

Encoder	Lang.	EN	FR	ZH
SBERT-DUSE	EU	**0.49**	**0.59**	0.01
	All	0.44	0.53	**0.19**
mBERT	EU	**0.55**	0.64	0.05
	All	**0.55**	**0.66**	**0.23**
mE5	EU	0.56	**0.67**	0.09
	All	**0.58**	**0.67**	**0.28**

(b) Fine-tuning approach. Attention dropout for regularisation.

Table 2: Macro-F1 on ML-ESG per language (average across 3 seeds). **No translation** is applied. Best performance per model is highlighted in **bold**.

Encoder	Translator	EN	FR	ZH
SBERT-DUSE	Google	0.48	0.57	**0.16**
	DeepL	**0.51**	**0.61**	**0.16**
mBERT	Google	0.44	**0.47**	**0.24**
	DeepL	**0.45**	**0.47**	0.23
mE5	Google	0.45	**0.50**	0.18
	DeepL	**0.46**	**0.50**	**0.21**

(a) Off-the-shelf approach on translated text. SVM as the classifier with Bayesian optimisation.

Encoder	Translator	EN	FR	ZH
SBERT-DUSE	Google	**0.49**	**0.58**	0.17
	DeepL	0.47	0.55	**0.20**
mBERT	Google	**0.55**	**0.65**	0.22
	DeepL	**0.55**	0.60	**0.23**
mE5	Google	0.55	0.65	**0.26**
	DeepL	**0.58**	**0.68**	**0.26**

(b) Fine-tuning approach on translated text. Attention dropout for regularisation.

Table 3: Macro-F1 on ML-ESG per language (average across 3 seeds). All inputs are **translated** to English, all models are trained with ***All Lang.***. Best performance per model is highlighted in **bold**.

formance can be detrimental for higher-resource ones, especially in cases where the training data mix languages that have deep typological differences, as in the case of Chinese and the two European languages.

While mBERT performs more stably across the board, the overall performance is slightly lower than the other models. Plausibly, this is due to both SBERT-DUSE and mE5 having taken advantage of their extensive training and their exposure to more training data compared to a standard pretrained mBERT.

Table 2b also shows that, despite the limited size of our training data, fine-tuning models end-to-end tends to yield better performance than the off-the-shelf approach. Fine-tuning modifies representations to be more task-specific, in contrast to the off-the-shelf approach where the encoder representation space remains static throughout the training process. Finally, it can be noticed that mE5 achieves the top overall performances after fine-tuning, with a marked improvement on ZH compared to the competitors.

Table 3 shows the results using the translated text of *All Lang.* for training. One would hypothesize that the task gets easier after translation as the models have to handle a single language only. Yet, this step often exhibits insignificant or even detrimental effects.

Also with translation, performance remains generally higher for the fine-tuning approach. This highlights the robustness of the feature learning with this technique. Once again, mE5 is the model achieving the overall highest scores for all the three languages as shown in Table 3b. Google Translate and DeepL Translator demonstrate comparable performance, regardless of the encoder utilised. Despite the slight bias towards DeepL translations in the off-the-shelf setting, the choice of the translator should be subject to the specific task and target language.

5. Conclusion

In this work, we evaluate methods for tackling the Multilingual ESG Issue Identification. To facilitate cross-lingual learning, we have modified the ML-ESG dataset (Chen et al., 2023a) and unified the sets of labels across languages. Moreover, the evaluation is carried out in a multi-label, non-exclusive classification setting, in order to make the task in English and French similar to Chinese. In our view, the multi-label setting allows for a more natural evaluation of this task, since in real-world ESG reports often cover more than one issue.

We have also studied the differences between the off-the-shelf and fine-tuning approaches. The latter consistently outperformed the former on multilingual and translated datasets, demonstrating its advantage of learning task-specific features. Furthermore, translation has minimal impact on both methods, suggesting that it may be an optional step for the given task.

Acknowledgements

The authors acknowledge the support from the project “Analyzing the semantics of Transformers representations for financial natural language processing” (ZVYU), sponsored by the Faculty of Humanities of the Hong Kong Polytechnic University.

The authors would like to thank Ivan Mashkin for his help in data preparation, and the two anonymous reviewers for their constructive feedback. Also, we would like to thank Leonidas Gee for proofreading the paper.

6. Bibliographical References

Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of EMNLP*.

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 IJCAI Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023b. 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.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023c. Multi-Lingual ESG Impact Type Identification. In *Proceedings of the IJCNLP-AACL Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

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. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

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 IJCAI Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Juyeon Kang, Mehdi Kchouk, Sandra Bellato, Mei Gan, and Ismail El Maarouf. 2022. FinSim4-ESG Shared Task: Learning Semantic Similarities for the Financial Domain. Extended Edition to ESG insights. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging BERT Language Models for Multi-lingual ESG Issue Identification. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Lu Lu, Jinghang Gu, and Chu-Ren Huang. 2022. Inclusion in CSR Reports: The Lens from a Data-driven Machine Learning Model. In *Proceedings of the LREC Workshop on Computing Social Responsibility*.

Ivan Mashkin and Emmanuele Chersoni. 2023. HKESG at the ML-ESG Task: Exploring Transformer Representations for Multilingual ESG Issue Identification. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices. *arXiv preprint arXiv:2203.16788*.

Mukut Mukherjee. 2020. ESG-BERT: NLP Meets Sustainable Investing. *Towards Data Science Blog*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is Domain Adaptation Worth your Investment? Comparing BERT and FinBERT on Financial Tasks. In *Proceedings of the EMNLP Workshop on Economics and Natural Language Processing*.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2022. Discovering Financial Hypernyms by Prompting Masked Language Models. In *Proceedings of the LREC Workshop on Financial Narrative Processing Workshop*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks. In *Proceedings of EMNLP*.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of EMNLP*.

George Serafeim and Aaron Yoon. 2022. Stock Price Reactions to ESG News: The Role of ESG Ratings and Disagreement. *Review of Accounting Studies*, pages 1–31.

Benedict Sheehy and Federica Farneti. 2021. Corporate Social Responsibility, Sustainability, Sustainable Development and Corporate Sustainability: What Is the Difference, and Does It Matter? *Sustainability*, 13(11):5965.

US SIF. 2020. Sustainable Investing Basics. *Available on https://www. ussif. org/sribasics (last accessed on 20 May 2022)*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*.

Marcel Van Marrewijk. 2003. Concepts and Definitions of CSR and Corporate Sustainability: Between Agency and Communion. *Journal of Business Ethics*, 44(2-3):95–105.

Mingyu Wan and Chu-Ren Huang. 2022. Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference. In *Proceedings of the LREC Workshop on Computing Social Responsibility*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving Text Embeddings with Large Language Models. *arXiv preprint arXiv:2401.00368*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of NAACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP: System Demonstrations*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020a. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020b. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of ACL: System Demonstrations*.

A. Training Details

We implement the models using the Hugging Face `transformers` library (Wolf et al., 2020) for fine-tuning models end-to-end. We use a batch size of 16 and learning rate of $2e-5$. We carry out a grid search on the probability value for attention dropout, $p \in$ {0.1, 0.2, 0.3}. In Table 4, the best attention dropout value per model is highlighted in bold, these values are found using *All Lang.* and are applied to the experiments on the translated datasets. Results reported are from the corresponding `best_model`, where we define the `best_model_metric` as the macro-F1 score.

Encoder	p	EN	FR	ZH
SBERT-DUSE	**0.1**	0.44	0.53	0.19
	0.2	0.43	0.53	0.18
	0.3	0.42	0.50	0.18
mBERT	0.1	0.54	0.66	0.23
	0.2	0.55	0.66	0.23
	0.3	0.55	0.64	0.23
mE5	0.1	0.58	0.66	0.28
	0.2	0.57	0.67	0.27
	0.3	0.58	0.67	0.28

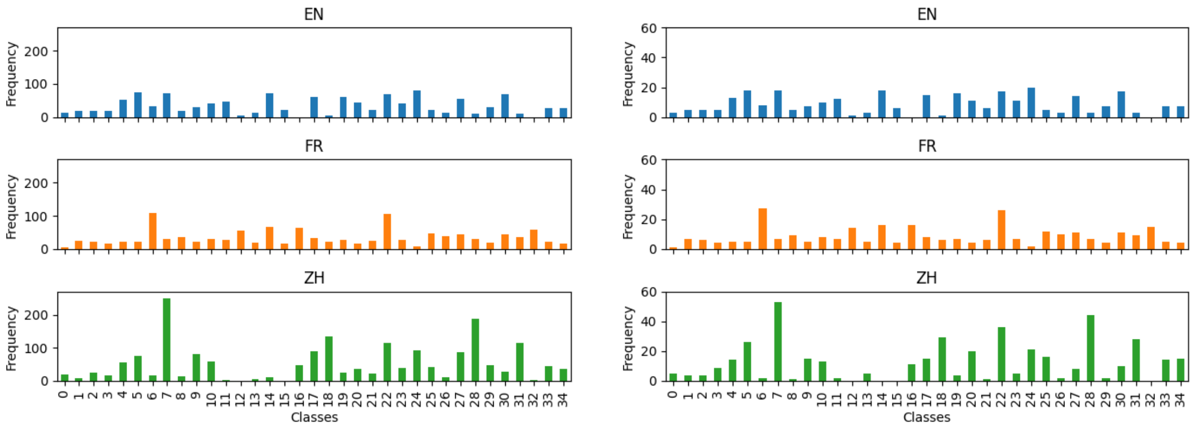
Table 4: Fine-tune models end-to-end with different probability values (p) for attention dropout. 0.1 is the default value. Models are trained with *All Lang.*. The best attention dropout value per model is highlighted in **bold**.

B. Dataset Details

Figure 1 shows the plots of the class distribution of the training and test sets per language (EN, FR and ZH). As ZH instances have multiple labels, the total number of counts is higher than EN and FR. Table 5 provides the labels of ESG key issues. Table 6 list the mappings of original ZH labels to the unified label space across the languages.

Index	Label
0	Access to Communications
1	Access to Finance
2	Access to Health Care
3	Accounting
4	Biodiversity & Land Use
5	Board
6	Business Ethics
7	Carbon Emissions
8	Chemical Safety
9	Climate Change Vulnerability
10	Community Relations
11	Consumer Financial Protection
12	Controversial Sourcing
13	Electronic Waste
14	Financing Environmental Impact
15	Health & Demographic Risk
16	Health & Safety
17	Human Capital Development
18	Labor Management
19	Opportunities in Clean Tech
20	Opportunities in Green Building
21	Opportunities in Nutrition & Health
22	Opportunities in Renewable Energy
23	Ownership & Control
24	Packaging Material & Waste
15	Pay
26	Privacy & Data Security
27	Product Carbon Footprint
28	Product Safety & Quality
29	Raw Material Sourcing
30	Responsible Investment
31	Supply Chain Labor Standards
32	Tax Transparency
33	Toxic Emissions & Waste
34	Water Stress

Table 5: Labels of ESG key issues.



(a) Training set.

(b) Test set.

Figure 1: Class distribution of the EN, FR and ZH datasets.

Original ZH Label		New Label
S12	產品責任 | 銷售模式和產品標示 (Selling Practices & Product Labeling)	Product Safety & Quality
S13	產品責任 | 產品設計與生命週期管 (Product Design & Lifecycle Management)	Product Safety & Quality
S14	產品責任 | 供應鏈管理 (Supply Chain Management)	Supply Chain Labor Standards
G11	公司行為 | 競爭行為 (Competitive Behavior)	None
G05	公司治理 | 重大事件風險管理 (Critical Incident Risk Management)	None
G08	公司治理 | 商業模式靈活度 (Business Model Resilience)	None
G06	公司治理 | 風險管理系統 (Systemic Risk Management)	None
G06	公司治理 | 風險管理系統 (Systemic Risk Management)	None
S05	人力資源 | 人權與社區關係 (Human Rights & Community Relations)	Community Relations
S11	產品責任 | 健康與人口風險 (Insuring Health & Demographic Risk)	Access to Health Care
E06	自然資源 | 原材料採購 (Raw Material Sourcing)	Raw Material Sourcing
S03	人力資源 | 人力資本發展 (Human Capital Development)	Human Capital Development
S19	社會機會 | 衛生保健管道 (Access to Health Care)	Access to Health Care
E11	環境機會 | 可再生能源的機會 (Opportunities in Renewable Energy)	Opportunities in Renewable Energy
S17	社會機會 | 溝通管道 (Access to Communication)	Access to Communication
E13	環境機會 | 綠色建造的機會 (Opportunities in Green Building)	Opportunities in Green Building
S08	產品責任 | 責任投資 (Responsible Investment)	Responsible Investment
G10	公司行為 | 納稅透明度 (Tax Transparency)	Tax Transparency
G07	公司治理 | 法律和法規環境的管理 (Management of the Legal & Regulatory Environment)	Management of the Legal & Regulatory Environment
S10	產品責任 | 金融產品安全性 (Consumer Financial Protection)	Consumer Financial Protection
S15	股東否決權 | 有爭議的採購 (Controversial Sourcing	Controversial Sourcing
S20	社會機會 | 營養與健康的機會 (Opportunities in Nutrition & Health)	Opportunities in Nutrition & Health

Table 6: Mapping of labels from the original ZH dataset to the unified label space with EN and FR.

	Original ZH Label	New Label
E12	環境機會 | 清潔技術的機會 (Opportunities in Clean Tech)	Opportunities in Clean Tech
G01	公司治理 | 董事會 (Board)	Board
E10	汙染與浪費 | 用於包裝的材料及浪費 (Packaging Material & Waste)	Packaging Material & Waste
S01	人力資源 | 人力資源管理 (Labor Management)	Labor Management
E02	氣候變化 | 產品碳足跡 (Product Carbon Footprint)	Product Carbon Footprint
G04	公司治理 | 會計 (Accounting)	Accounting
E07	自然資源 | 生物多樣性與土地利用 (Biodiversity & Land Use)	Biodiversity & Land Use
S02	人力資源 | 員工健康和安全 (Health & Safety)	Health & Safety
S04	人力資源 | 供應鏈勞動標準 (Supply Chain Labor Standards)	Supply Chain Labor Standards
E01	氣候變化 | 碳排放量 (Carbon Emissions)	Carbon Emissions
S09	產品責任 | 產品安全與品質 (Product Safety & Quality)	Product Safety & Quality
G03	公司治理 | 所有權 (Ownership & Control)	Ownership & Control
E04	氣候變化 | 氣候變化脆弱性 (Climate Change Vulnerability)	Climate Change Vulnerability
E03	氣候變化 | 融資環境影響 (Financing Environment Impact)	Financing Environment Impact
S18	社會機會 | 融資管道 (Access to Finance)	Access to Finance
E09	汙染與浪費 | 電子廢物 (Electronic Waste)	Electronic Waste
G09	公司行為 | 商業道德 (Business Ethics)	Business Ethics
S16	股東否決權 | 社區關係 (Community Relations)	Community Relations
E08	汙染與浪費 | 有毒物排放及浪費 (Toxic Emissions & Waste)	Toxic Emissions & Waste
S06	產品責任 | 化學物質安全性 (Chemical Safety)	Chemical Safety
S07	產品責任 | 隱私和數據安全 (Privacy & Data Security)	Privacy & Data Security
E05	自然資源 | 水資源壓力 (Water Stress)	Water Stress
G02	公司治理 | 薪酬 (Pay)	Pay
	Not related to ESG	None

Table 6: Mapping of labels from the original ZH dataset to the unified label space with EN and FR (continued).