HKESG at the ML-ESG Task: Exploring Transformer Representations for Multilingual ESG Issue Identification

Ivan Mashkin
City University of Hong Kong
ivan.mashkin2018@gmail.com

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Abstract

Environmental, Social and Governance reports have to be periodically released by financial companies, as they represent an essential guide for the potential, socially-responsible new investors. Therefore, automatizing the analysis of reports and extracting the main ESG issues mentioned in the text is a goal of primary importance for financial Natural Language Processing (NLP) systems.

In this paper, we report our experiments for the FinSim4-ESG Shared Task, dedicated to the problem of multilingual ESG issue identification in English and French. Our results show that even simple classifiers trained on multilingual data and using crosslingual Transformer representations can achieve a strong performance in the task.

1 Introduction

Sustainable, Responsible and Impact investing (SRI) has gained a lot of prominence in the last decades (Serafeim and Yoon, 2022; Mehra et al., 2022). As a discipline, one of its primary goals is to specify environmental, social and governance criteria to generate long-term financial returns and produce a positive impact on the society (Mukherjee, 2020). For corporations, adherence to Environmental, Social and Governance (ESG) practices has become a requirement: for example, SEC filings in the US have to follow standard for Climate Change and Human Governance, and the European Commission stipulated, at the end of 2022, that all the companies providing investment products will have to disclose how their economic activity align with the taxonomy of the European Union and with the ESG regulations for sustainability (Kang et al., 2022). It is thus not a surprise that the demand for language technologies to automatize the analysis of ESG reports is correspondingly increasing.

With the rising popularity of machine learning and NLP technologies for Natural Language Processing (Loughran and McDonald, 2016), there is also a number of academic initiatives dedicated to research on development of systems for extracting relevant issues from ESG reports. The present paper aims at reporting our findings on the multilingual datasets of the FinSim4-ESG Shared Task (Chen et al., 2023). We participated in the English and in the French track and our best model, a simple SVM classifier relying on the crosslingual representations of the Distilled Universal Sentence Encoder (Reimers and Gurevych, 2019), achieves a F1-score of 0.62 and 0.71 on the English and the French test data, respectively.

2 Related Work

2.1 ESG and NLP

The field of corporate sustainability is interested in the set of self-regulatory acts that international business perform to mitigate the negative impacts on the society (Van Marrewijk, 2003; Sheehy, 2015; Feng and Ngai, 2020). Such practices are regulated by international standards and policies (Sheehy and Farneti, 2021). The issues ESG reports have to deal with are organized in taxonomies, and their automatic identification recently attracted attention in the NLP research community, in the form of the organization of a dedicated workshop at the LREC conference (Wan and Huang, 2022) and a shared task co-located with the IJCAI conference (Kang et al., 2022).

In the former, the topics of the contributions showed a varied interest in analyzing the language data in ESG reports, including machine learning models to fight stereotypes and improve inclusivity (Lu et al., 2022), corpus-based analyses of the metaphors in the legitimation strategies for the business of oil companies in China and in the United

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.

Category	Labels
Environment	Carbon Emissions, Climate Change Vulnerability, Product Carbon Footprint Biodiversity & Land Use, Water Stress, Electronic Waste, Packaging Material & Waste Toxic Emissions & Waste, Opportunities in Renewable Energy, Opportunities in Clean Tech, Opportunities in Green Building, Opportunities in Renewable Energy
Social	Health & Demographic Risk, Human Capital Development, Labor Management, Supply Chain Labor Standards, Chemical Safety, Consumer Financial Protection, Privacy & Data Security, Product Safety & Quality, Community Relations, Raw Material Sourcing, Access to Health Care, Opportunities in Nutrition & Health, Health & Safety
Governance	Ownership & Control, Accounting, Board, Tax Transparency, Business Ethics, Pay, Responsible Investment

Table 1: Map of the dataset labels, divided into the three main categories of Environment, Social and Governance.

States (Chen et al., 2022), and diachronic distributional methods to identify changes in the usage of ESG terms over time (Purver et al., 2022).

The shared task organized by Kang et al. (2022) was challenging the teams on two different subtasks: a taxonomy enrichment task, in the form of unsupervised discovery of hypernyms (Camacho-Collados et al., 2018) in sentences from ESG reports; and a binary classification task of the sustainability (sustainable / not sustainable) of excepts from the same type of reports.

2.2 Language Models for Financial Natural Language Understanding

Language models based on the Transformer architecture have been taking NLP by storm in recent years (Vaswani et al., 2017; Devlin et al., 2019), and a consequence of the success of Transformers, researchers working on NLP for specialized domains turned to domain adaptation techniques to exploit the full potential of such architectures (Guo and Yu, 2022). The financial domain makes no exception: the recently-developed models include adaptations of BERT (e.g. FinBERT, Araci (2019); Yang et al. (2020); Liu et al. (2020)), ELECTRA (FLANG-ELECTRA, Shah et al. (2022)) and even large language models such as BloombergGPT (Wu et al., 2023).

Transformers for financial NLP have been evaluated on a variety of tasks, either supervised (e.g. sentiment analysis, named entity recognition, numeral understanding; Peng et al. (2021); Shah et al. (2022); Wu et al. (2023)) or unsupervised ones (e.g. hypernym detection; Chersoni and Huang (2021); Peng et al. (2022)), showing important gains over the performance of general domain models.

Interestingly, models specialized for dealing with ESG issue identification have also been developed and made publicly available (Yang et al., 2020; Mukherjee, 2020; Mehra et al., 2022). Such models benefit from additional training on corpora of annual sustainability reports.

2.3 Multilingual Language Models

Transformers also led to impressive improvements in multilingual NLP, thanks to the introduction of large architectures that have been pretrained with language modeling objectives on multiple languages at the same time (e.g Multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020); BLOOM (Scao et al., 2022)). Such models are initialized with a large shared vocabulary, and utilize sophisticated sampling methods to balance the representation of high-resource and low-resource languages in the same semantic space.

In parallel with the development of Sentence Transformers (Reimers and Gurevych, 2019, 2020), which are able to generate vector representations of entire sentences and paragraphs, NLP researchers also introduced *multilingual sentence embeddings*. Those are based on the idea of having first a monolingual model generating sentence embeddings for a source language, and then having multiple student models trained on the translated sentences in other languages to mimic the original model.

3 Experimental Settings

3.1 Dataset Description

The organizers of the shared task made available training datasets in French and English, containing respectively 1200 and 1199 labeled examples. The 35 labels were defined on the basis of the MSCI ESG standard rating guidelines [1], and were generally related to three macro categories: Environment, Social and Governance (see Table 1). Finally, they released test sets for the two languages, each one

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

including 300 examples without labels. The gold labels were later made available for further evaluation and analysis.

The raw materials of the dataset were multilingual news articles, which were labeled by experts in ESG annotation: the news were collected, respectively, from ESGToday for English [2] and from RSEDATANEWS[3] and Novethic[4] for French. The English and French datasets are annotated by experts (2 annotators and 1 reviewer) in Fortia's Data & Language Analyst teams. The dataset instances include both the title and the main body of the news and the labels are mutually exclusive (in the cases where multiple labels could apply to one article, the texts were split into multiple instances). Noticeably, two of the labels ("Health & Safety" and "Tax Transparency") are present in the French but not in the English dataset, and thus we just excluded the examples with those labels (16 instances, in total) in the experiments in which we use French training data to make predictions on the English test set.

3.2 Systems Description and Settings

As a preprocessing step, we concatenated the text of the title and the text of the body of the news. Next, we adopted two different approaches for representing the ESG news with Transformers.

3.2.1 Approach 1: ESG Transformers with Sentence Translation

In the first approach, we used Transformer models that are specialized for ESG data, in particular the **ESG-BERT** model by Mukherjee (2020) [5] and the **FinBERT** model by Yang et al. (2020) with a previous fine-tuning on a dataset of 2000 ESG reports and three output labels (Environment, Social and Governance). [6] We chose to use ESG Transformers as they were fine-tuned on a similar type of textual data. Since both models are available only for English, we translated the French dataset with the help of the Google Translator API. [7]

We initially fine-tuned the models via 5-fold stratified sampling, to be sure that each fold had similar class distribution. However, we realized that the models were underfitting, probably because of the small size of the dataset. [8] Therefore we decided to use the fine-tuned Transformer models to generate vector representations of the dataset instances and to utilize different types of classifiers on top of them.

In particular, we used Logistic Regression (LR), Random Forests (RF), Support Vector Machine (SVM), all of them in the standard implementation in the Scikit-learn library (Pedregosa et al., 2011). For the classifiers, the parameters were optimized via the Skopt library for Bayesian optimization [9] and using 5-fold stratified sampling, similarly to what we originally did for model fine-tuning.

3.2.2 Approach 2: Sentence Transformers and Multilingual Training

In the second approach, we adopted a multilingual training approach: we used the Distilled Universal Sentence Encoder (**DUSE**) from the Sentence Transformers library [10] to encode directly the English and the French sentences, and then we simply merged the two datasets for multilingual training. We chose this approach because it maps the French-English data onto a unified vector space, so it allows us to simply merge the two datasets for training. Given the previous results, this time we used directly the pretrained Transformers to generate the input vectors for the classifiers (LR, RF and SVM) and we did not try to fine-tune the models. For finding the best parameters for the classifiers, we used the same procedure described above, combining 5-fold stratified sampling and the Skopt library for Bayesian optimization.

4 Results

The metrics for all the systems can be seen in Table 2 for English and in Table 3 for French.

A first notable finding in our result is that the multilingual representation of the Universal En-

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

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

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

[5] https://huggingface.co/nbroad/ESG-BERT.

[6] https://huggingface.co/yiyanghkust/finbert-esg.

[7] For both approach 1 and approach 2, when the test dataset was the English one we excluded from the training data the 16 French instances with either "Health & Safety" or the "Tax Transparency" gold standard labels.

[8] With fine-tuned models, the preliminary results on the validation data always showed Accuracy scores in the low 40s, while the classifiers on top of the Transformer vectors performed more closely to the reported scores on the test set.

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

[10] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1.

System	Accuracy	Precision	Recall	F1-score (Macro)
LR-ESG-BERT	0.59	0.55	0.59	0.56
RF-ESG-BERT	0.58	0.52	0.57	0.52
SVM-ESG-BERT	0.60	0.61	0.62	0.60
LR-FinBERT	0.59	0.54	0.56	0.53
RF-FinBERT	**0.62**	0.53	0.53	0.53
SVM-FinBERT	**0.62**	0.59	0.59	0.58
LR-DUSE	0.58	0.56	0.57	0.55
RF-DUSE	0.58	0.57	0.60	0.56
SVM-DUSE	0.58	**0.63**	**0.63**	**0.62**

Table 2: Results for all the systems on the English test dataset (300 examples, best scores per metric are in **bold**).

System	Accuracy	Precision	Recall	F1-score (Macro)
LR-ESG-BERT	0.67	0.66	0.67	0.66
RF-ESG-BERT	0.59	0.56	0.59	0.56
SVM-ESG-BERT	0.64	0.65	0.62	0.62
LR-FinBERT	0.70	0.69	0.70	0.68
RF-FinBERT	0.67	0.63	0.64	0.62
SVM-FinBERT	0.69	0.69	0.68	0.67
LR-DUSE	0.62	0.56	0.58	0.56
RF-DUSE	0.64	0.57	0.61	0.58
SVM-DUSE	**0.71**	**0.72**	**0.72**	**0.71**

Table 3: Results for all the systems on the French test dataset (300 examples, best scores per metric are in **bold**).

coder generally perform better than the domain-adapted ones of FinBERT and ESG-BERT. The translation of the French sentences to English to fit in the English-language domain-adapted Transformers does not seem to affect the trend too much. Among the classifiers that we explored, SVM is consistently the best option in the English dataset; it performs closely to LR on the French data with the two ESG Transformers, but it outperforms the other classifiers by a large margin with DUSE.

It is noticeable that when we compare Accuracy and F1-Score Macro, which is computed by using the mean of the F1-score of the single classes, most systems tend to have a higher value of Accuracy. We interpret this as an effect of imbalanced classes. However, SVM-DUSE is the only system for which F1-Score is the same, or even higher than Accuracy.

In the French dataset, the more frequent classes are generally predicted better: the accuracy is at least above 0.6 for all the classes with at least 10 examples in the test data. This does not hold for English, where we noticed that, for several classes with relatively high support, the accuracy is below chance level, e.g. **Electronic Waste, Health & Demographic Risk, Financing Environmental Impact, Privacy & Data Security**, which are all in the top-10 of the most frequent classes. We hypothesized that this might be due to diverging label distributions between the English test set and the joint training set. A Pearson correlation test revealed that, indeed, the class frequency correlation between the joint training and the English test data is lower than for the French data ($r = 0.62$ vs. $r = 0.74$), so this could be a partial explanation of the different performance across languages.

For the shared task, we submitted our systems with the following names: SVM-ESG-BERT as HKESG1, SVM-FinBERT as HKESG2 and SVM-DUSE as HKESG3. Our best ranks, both obtained by SVM-DUSE, are the 9th place out of 23 systems in the English track and the 10th place out of 21 systems of in the French track.

5 Conclusions

In this paper, we presented the systems that we used to compete in the ML-ESG shared task on Multilingual ESG Issue Identification. We took part in both the English and the French track, and our best system was -perhaps surprisingly- a simple linear SVM model relying on the sentence vector representation generated by the Distilled Universal Sentence Encoder (Reimers and Gurevych, 2019).

The dataset size was too small for the fine-tuning of Transformers. However, multilingual training was sufficient to obtain robust results on both dataset (0.62 of F1-score for English and 0.71 for French). The scores for the English dataset are generally higher, probably due to a more diverging distribution between training and test sets.

Acknowledgements

This work has supported by the Faculty of Humanities of the Hong Kong Polytechnic University (project "Analyzing the semantics of Transformers representations for financial natural language processing", code 1-ZVYU). We would like to thank the anonymous reviewers for their constructive feedback.

References

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

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of SemEval*.

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

Jieyu Chen, Kathleen Ahrens, and Chu-Ren Huang. 2022. Framing Legitimacy in CSR: A Corpus of Chinese and American Petroleum Company CSR Reports and Preliminary Analysis. In *Proceedings of the LREC Workshop on Computing Social Responsibility*.

Emmanuele Chersoni and Chu-Ren Huang. 2021. PolyU-CBS at the FinSim-2 Task: Combining Distributional, String-Based and Transformers-Based Features for Hypernymy Detection in the Financial Domain. In *Companion Proceedings of the Web Conference*.

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*.

Penglan Feng and Cindy Sing-bik Ngai. 2020. Doing More on the Corporate Sustainability Front: A Longitudinal Analysis of CSR Reporting of Global Fashion Companies. *Sustainability*, 12(6):2477.

Xu Guo and Han Yu. 2022. On the Domain Adaptation and Generalization of Pretrained Language Models: A Survey. *arXiv preprint arXiv:2211.03154*.

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*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. In *Proceedings of IJCAI*.

Tim Loughran and Bill McDonald. 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4):1187–1230.

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*.

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*.

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*.

Matthew Purver, Matej Martinc, Riste Ichev, Igor Lončarski, Katarina Sitar Šuštar, Aljoša Valentinčič, and Senja Pollak. 2022. Tracking Changes in ESG Representation: Initial Investigations in UK Annual Reports. In *Proceedings of the LREC Workshop on Computing Social Responsibility*.

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*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176b-parameter Open-access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*.

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.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE Meets FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain. In *Proceedings of EMNLP*.

Benedict Sheehy. 2015. Defining CSR: Problems and Solutions. *Journal of Business Ethics*, 131:625–648.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*.

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*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv preprint arXiv:2303.17564*.

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