The Risk and Opportunity of Data Augmentation and Translation for ESG News Impact Identification with Language Models

Yosef Ardhito Winatmoko
Nestlé
Jakarta, Indonesia
yosef.ardhito@id.nestle.com

Ali Akbar Septiandri
Nokia Bell Labs
Cambridge, UK
ali.septiandri@nokia-bell-labs.com

Abstract

This paper presents our findings in the ML-ESG-2 task, which focused on classifying a news snippet of various languages as "Risk" or "Opportunity" in the ESG (Environmental, Social, and Governance) context. We experimented with data augmentation and translation facilitated by Large Language Models (LLM). We found that augmenting the English dataset did not help to improve the performance. By fine-tuning RoBERTa models with the original data, we achieved the top position for the English and second place for the French task. In contrast, we could achieve comparable results on the French dataset by solely using the English translation, securing the third position for the French task with only marginal F1 differences to the second-place model.

1 Introduction

ESG factors have gained increasing prominence in recent years, not only among stakeholders but also in the decision-making processes of investors and financial institutions. As the awareness of ESG risks grows, so does the need for precise and real-time classification of these risks. Traditional ESG risk analysis has largely relied on structured data, such as company disclosures, financial reports, and pre-defined ESG metrics. However, these sources often provide an incomplete and lagging view of a company's ESG footprint. Moreover, they are subject to reporting biases and may lack granularity in capturing the diverse dimensions of ESG risks.

The proliferation of online news media offers a fertile ground for harvesting a more comprehensive set of data on ESG issues. News articles, in particular, often capture real-time events, public sentiment, and expert opinions, providing a more immediate and multifaceted perspective on ESG risks than can be obtained from traditional structured data. Yet, leveraging this unstructured textual data to accurately classify ESG risks presents computational challenges. These challenges include but are not limited to, natural language understanding, sentiment analysis, and the development of a robust taxonomy for ESG risk classification.

The FinNLP-2022 workshop introduced a FinSim4-ESG[1] shared task that centres on ESG issues. To deepen the understanding of these areas, FinNLP@IJCAI-2023 released a new dataset for the FinNLP community. This dataset is designed to explore the task of identifying key ESG issues in multiple languages, guided by the MSCI ESG rating framework, which includes 35 key issues for categorisation.

Expanding on this discourse, a new task dubbed Multi-Lingual ESG Impact Type Identification (ML-ESG-2) was introduced. The primary objective of this task is to identify the type of ESG impact a given piece of news may have. Specifically, models are tasked with determining whether the news presents an ESG-related opportunity or risk. This aspect of impact identification is structured as a single-choice question.

In this study, we present our methodology for tackling the ML-ESG-2 shared task, using datasets in both English and French. We conducted our experiments using two primary methods: fine-tuning language models on the dataset and training a logistic regression model using Sentence-BERT (SBERT) embeddings (Reimers and Gurevych, 2019). Our findings suggest that the optimal approach can be achieved solely by relying on the given datasets, without the need for any data augmentation. Furthermore, we found that comparable results could be obtained on the French dataset by first translating the text into English and then employing a model pre-trained specifically on English text. Using this strategy, our models secured first place in the English language dataset and second and third places in the French language dataset.

[1] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022/shared-task-finsim4-esg

Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pages 66–71
November 1, 2023. ©2023 Association for Computational Linguistics

2 Datasets

In ML-ESG-2, a new task called "ESG Impact Type Identification" was introduced to advance discussions on ESG ratings from the previous shared task (Chen et al., 2023). This task requires models to discern whether a given news article signifies an ESG "opportunity" or "risk". Each data point consists of the URL, title, content, and the assigned impact type for a given article. An overview of the English and French datasets can be seen in Table 1.

	English	French
Training-Opportunity	694 (85.9%)	458 (56.0%)
Training-Risk	114 (14.1%)	360 (44.0%)
Test-Opportunity	191 (87.6%)	111 (55.5%)
Test-Risk	27 (12.4%)	89 (44.5%)

Table 1: Summary statistics of the datasets

Data Augmentation While the French dataset exhibits a relatively balanced distribution across its classes, the English dataset has significantly more instances labelled as "Opportunity" compared to "Risk". Drawing inspiration from the successful approach employed by the previous ML-ESG task winner (Lee et al., 2023), we experimented with data augmentation using GPT3Mix (Yoo et al., 2021) to augment the "Risk" training data during the fine-tuning of the English models. For more details on this process, please refer to Appendix A.1. Our approach involved leveraging the `text-davinci-003` model from OpenAI to generate the additional training data.

Data Translation As a pivotal component of our experimental approach, we employed large language models (LLMs) to facilitate the translation of training data from French to English (see Appendix A.2). This translation step was essential to ensure that our models, primarily designed for English text processing, could effectively comprehend and learn from the French-language content within the dataset.

3 Methods

We conducted experiments using two primary methods: fine-tuning language models on the dataset and training a logistic regression model using SBERT embeddings (Reimers and Gurevych, 2019). Given the absence of development sets and the presence of imbalanced training data, we employed 5-fold cross-validation to assess the effectiveness of our approaches. Additionally, we observed that the news titles are not unique and two news contents with the same title can be both "Opportunity" and "Risk". Thus, we decided to disregard the titles altogether and used only the content as the input.

Baseline. To demonstrate the performance improvements achievable through the two methods mentioned earlier, we initially established a simple baseline. This benchmark was created by employing TF-IDF and logistic regression using scikit-learn (Pedregosa et al., 2011). We retained the top 1000 features identified by TF-IDF and conducted hyperparameter optimisation (see Appendix A.3).

Fine-tuning language models. We experimented with three well-known pre-trained encoder models: DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021). Specifically, the model names in Hugging Face (Wolf et al., 2020) were `distilbert-base-uncased`, `roberta-large`, `microsoft/deberta-v3-large`. We also experimented with the XLM-RoBERTa (`xlm-roberta-large`) model (Conneau et al., 2020) for the French dataset. We fine-tuned the pre-trained models on the ML-ESG-2 dataset and used Optuna (Akiba et al., 2019) to find the optimal hyperparameters of each model. The final list of the hyperparameters is shown in Table 2.

Model Name	Batch Size	Learning Rate	Epoch
DistilBERT	32	2.5e-5	4
RoBERTa	16	1.3e-5	2
XLM-RoBERTa	4	6.8e-6	4
DeBERTa	4	2.3e-5	4

Table 2: Hyperparameters used in the model fine-tuning.

Sentence-BERT. Unlike traditional word embeddings that represent individual words, SBERT (Reimers and Gurevych, 2019) is designed to generate embeddings for entire sentences or paragraphs. It leverages pre-trained transformer-based models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and fine-tunes them specifically for sentence-level tasks. This fine-tuning process enables SBERT to capture contextual information, semantic meaning, and the relationships between sentences. In our experiment, we used Sentence-T5 (ST5) (Ni et al., 2021), a variant based on T5

(Raffel et al., 2020), to generate the embeddings as input for logistic regression. We used ST5 because it outperformed SBERT in a range of natural language inference (NLI) and question-and-answer (Q&A) tasks (Ni et al., 2021). Additionally, we conducted experiments with different sizes of the ST5 model in this study.

4 Results

Our experiments revealed several key findings in the context of model performance. Firstly, RoBERTa demonstrated outstanding performance on the English dataset, achieving an impressive F1 score of 85.19% (Table 3)[2]. In comparison, ST5-XXL also performed well with an F1 score of 81.29%. Remarkably, both models achieved the best results without the need for data augmentation and translation.

When subjected to the test set, RoBERTa continued to shine, outperforming ST5-XXL by a significant margin. RoBERTa achieved an F1 score of 92.00%, while ST5-XXL scored 75.36% (Table 5). This discrepancy underscores RoBERTa's robustness and suitability for this task.

Shifting our focus to the French dataset, our cross-validation results favoured ST5-XXL (FR→EN) over XLM-RoBERTa (FR), with an F1 score of 76.52% versus 72.17% (Table 4). Nevertheless, when assessing the model's performance on the test set, XLM-RoBERTa emerged as the winner, achieving an F1 score of 86.12% (Table 5). These findings highlight the importance of considering both cross-validation and test set results when evaluating model performance.

Furthermore, the test results between XLM-RoBERTa and ST5-XXL differ only slightly (<1%). Note that XLM-RoBERTa was trained on the original French dataset, but ST5-XXL used the translated dataset. The results for ST5-XXL raise the potential of translating other languages to English and used English-based models for text classification.

5 Related Work

In recent studies (Lehman et al., 2023; Xu et al., 2023), it was discovered that relatively compact specialised clinical models exhibit significantly superior performance compared to all in-context learning approaches when applied to LLMs. This superior performance holds true even when these clinical models are fine-tuned on a limited amount of annotated data. Additionally, their research revealed that pretraining on clinical tokens enables the development of smaller, more parameter-efficient models that can either match or surpass the performance of much larger language models trained on general text.

On a similar note, Septiandri et al. (2020) found that classical NLP techniques, i.e. bag-of-words and TF-IDF, could produce comparable results to the more advanced word2vec (Bojanowski et al., 2017) and BiLSTM with a fraction of the training time. The study focused on a binary classification task, similar to ML-ESG-2. They suggested that even though the tiny improvement from complex models is crucial in a competition, one should consider allocating more resources to improve the quality of the dataset in practical settings.

6 Conclusion

In summary, within the English sub-task of ML-ESG-2, the RoBERTa model fine-tuned only with the original data secured the top position, even without any particular strategy to address the class imbalance. Our investigation revealed that data augmentation failed to improve F1 scores on the training set. Similarly, adding translated French news for English models did not contribute to improved performance. From these findings, we deduce that increasing data quantity through augmentation and translation may not consistently benefit model performance.

For the French sub-task, we observed that translating to English provided better results for the training set. However, the multi-lingual RoBERTa model (XLM-RoBERTa) fine-tuned on the original French dataset achieved higher F1 for the test set, albeit slightly. These results indicate an opportunity for future works: the potential of translating text to English as a preprocessing strategy in other languages.

7 Availability

The code is available at `https://github.com/aliakbars/esg-finnlp`.

[2] All F1 scores shown are calculated based on the F1-score of the "Risk" class. We also include the weighted-F1 scores for the test set as presented in the final leaderboard of ML-ESG-2 task for reference.

	Dataset		
Model	EN	EN+AUG	EN+FR
Baseline	48.45% ± 9.30%	-	-
ST5-Base	78.63% ± 5.66%	78.32% ± 7.56%	72.82% ± 11.40%
ST5-XXL	81.29% ± 2.72%	80.01% ± 2.98%	79.08% ± 5.45%
DistilBERT	79.84% ± 9.28%	76.01% ± 14.8%	71.26% ± 11.72%
RoBERTa	**85.19%** ± **8.97%**	83.68% ± 8.30%	82.48% ± 9.65%
DeBERTa	82.23% ± 10.83%	76.92% ± 14.47%	72.19% ± 19.99%

Table 3: **F1 scores on the English training dataset.** Apart from using only the English dataset (EN), we also experimented with augmenting the dataset using large language models (EN+AUG), and the English translation of the French dataset (EN+FR). We were only augmenting the training set using the Risk-labelled data points.

	Dataset	
Model	FR	FR→EN
Baseline	65.61% ± 4.66%	63.95% ± 3.38%
ST5-Base	-	71.13% ± 4.89%
ST5-XXL	-	**76.52%** ± **4.26%**
RoBERTa	-	67.33% ± 12.73%
XLM-RoBERTa	**72.17%** ± **5.43%**	71.05% ± 6.09%

Table 4: **F1 scores on the French training dataset.** FR→EN indicates the use of the English translation of the French dataset during training.

	F1		Weighted F1	
Model	EN	FR	EN	FR
ST5-XXL	75.36%	85.96%	92.89%	83.94%
RoBERTa	**92.00%**	-	98.10%	-
XLM-RoBERTa	-	**86.12%**	-	85.54%

Table 5: **F1 scores on the test sets using the best models.** All models were trained on the vanilla training set, except ST5-XXL (FR) which used the translated version of the French (FR) dataset.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

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

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 578–597. PMLR.

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. *arXiv preprint arXiv:1907.11692*.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. volume 35, pages 27730–27744.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ali Akbar Septiandri, Yosef Ardhito Winatmoko, and Ilham Firdausi Putra. 2020. Knowing right from wrong: Should we use more complex models for automatic short-answer scoring in Bahasa Indonesia? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing: System*

Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

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.

A Additional Experimental Details

A.1 Data Augmentation Prompt

We ran the augmentation 100 times, each time using three distinct samples from the actual dataset and taking five generated texts. In total, we generated 500 additional "Risk" contents for the English dataset using the following prompt:

```
Each line in the following list contains a
     snippet taken from a news article and the
     respective ESG impact identification.
ESG impact is one of 'Risk' and 'Opportunity'.
Opportunity: (a random 'Opportunity' content)
Risk: (another random 'Risk' content)
Opportunity: (another random 'Opportunity'
     content)
Risk:
```

A.2 Translating French to English

We experimented with three models to translate the text from French to English: Flan-T5 (Chung et al., 2022), DeepL [3], and GPT-3.5 Turbo (Ouyang et al., 2022)). We found that the English translation using GPT-3.5 Turbo would result in a better performance. Thus, the results reported in this paper were based on the GPT-3.5 Turbo translation only.

A.3 Hyperparameter Tuning

For the logistic regression model trained on the SBERT embeddings and the baseline approach, we tuned the hyperparameters using the values provided in Table 6.

Hyperparameter	Values tested
C	$\{0.1, 1, 10, 100\}$
class_weight	$\{1, 2, 5, 10, 20\}$

Table 6: Hyperparameters tested for logistic regression

[3] https://www.deepl.com/translator