Team HHU at the FinNLP-2023 ML-ESG Task: A Multi-Model Approach to ESG-Key-Issue Classification

Fabian Billert and **Stefan Conrad**
Heinrich-Heine University of Düsseldorf
{fabian.billert, stefan.conrad}@hhu.de

Abstract

In this paper, we discuss our submission to the Multi-Lingual ESG Issue Identification (ML-ESG) 2023, where we classify news articles into different ESG key-issues defined by MSCI. We use an adapter-based approach and evaluate different approaches and configurations, finally showing that it is advantageous to use multiple models in order to first classify articles into E/S/G classes before determining the final sub-issues.

1 Introduction

The surge in Environmental, Social, and Governance (ESG) research over the past few years is a testament to the growing importance of these issues in the corporate world (Zumente and Bistrova, 2021). Companies are increasingly recognizing that ESG-related matters can pose significant risks if not addressed properly (Aue et al., 2022). Beyond risk management, ESG topics are also crucial for a company's reputation, as they often reflect the company's values and commitment to sustainable practices (Schramm-Klein et al., 2016), (Islam et al., 2021).

Investors, too, are becoming more attentive to the ESG behaviors of companies. One common method of evaluating a company's ESG practices is through the human-curated scores provided by major rating agencies like MSCI[1] or Sustainalytics[2]. These agencies assess whether a company adheres to good ESG practices and assign a numerical value to represent the company's ESG performance. MSCI does this by considering 35 key-issues which they combine in different ways depending on the specific industry a company operates in (Nagy et al.). The final weight of a key-issues in the ESG score calculation is determined by quantitatively assessing each industry and consulting with investment practitioners.

Upon closer scrutiny of the evaluation metrics employed by the different ESG rating agencies, it becomes evident that these metrics do not completely incorporate sustainability principles into their process of assessing corporate sustainability (Escrig-Olmedo et al., 2019). In addition, (Crona, 2021) raises several concerns with the traditional rating agencies. One point of critique mentioned by them is that companies might self report data on positive environmental initiatives that are not connected to their negative environmental impact, but are similarly considered by the rating agencies. On the other hand, scoring mechanisms like the one used by MSCI are problematic in the sense that the weighting mechanism might not consider key ESG issues, depending on how the weights were created. These uncertainties in the evaluation process underscore the need for more comprehensive and nuanced methods of assessing ESG practices.

In order to create independent analyses, machine learning techniques, particularly those in Natural Language Processing (NLP), can be used. Over the past few years, NLP research has seen a significant uptick, with advancements in this field offering promising solutions for more in-depth ESG analysis (Min et al., 2021), (Chen et al., 2022), (Fischbach et al., 2022). By leveraging NLP, investors can conduct their own research to determine the sustainability of potential investment companies.

The Machine Learning for ESG (ML-ESG) task (Chen et al.) aims to motivate research in this direction and has annotated news articles in English, French, and Chinese based on the 35 key-issues used by MSCI. The task challenges participants to develop a system capable of classifying

[1] https://www.sustainalytics.com/esg-data
[2] https://www.msci.com/our-solutions/esg-investing/esg-ratings

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.

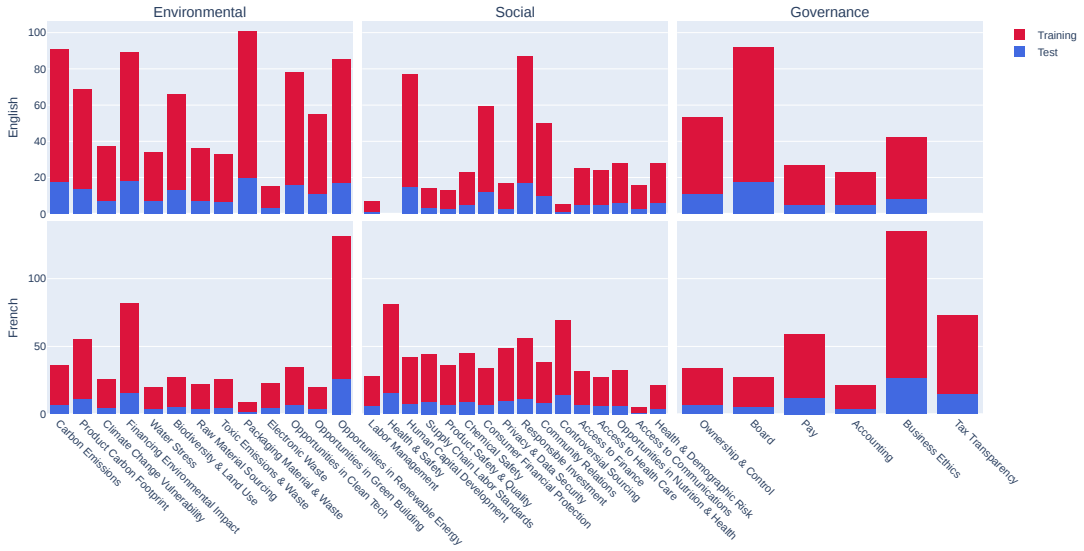

Figure 1: Occurrences of the different labels in the training- and test-data of the task. The top row represents the English data, the bottom row the French data. Each column represents one out of E/S/G as written above the figure. Training data is shown in red, test data in blue. **Note:** The test data was released after the task deadline and was not used during training unless mentioned.

articles in these languages into the appropriate ESG issues. This represents a significant stride towards more nuanced and comprehensive ESG analysis, ultimately enabling more informed and sustainable investment decisions.

In this paper, we present our solution for the ML-ESG task for the English and French datasets. We train a set of adapters for each language and try out different approaches to classify the news articles, ultimately showing that it is preferable to first classify a news article into one of the three main classes (Environmental, Social or Governance), before further classifying into the key-issues belonging to each category. Our approach achieves third place for the French language and nineteenth for the English language.

2 Task Description and Dataset

The dataset contains 1200 French and 1199 English news articles. Each article has the following properties: "URL", "news_title", "news_content", "ESG_label" (Chen et al.). The "ESG_label" is one of the 35 key-issues described in the ESG Industry Materiality Map of MSCI[3]. Each key-issue is attributed to one of the three top ESG components, "environmental", "social" and "governance". In Figure 1, different histograms for each ESG component show the occurrence of all the key-issues per language. Most news articles are classified as one of the environmental key-issues while the least articles belong to governance key-issues.

3 Experimental Approach

3.1 Adapters

Adapters are an efficient and flexible method for fine-tuning a foundational model for unique tasks (Houlsby et al., 2019) or transferring task-specific knowledge across different languages (Pfeiffer et al., 2020b). These tools are particularly useful when dealing with a dataset composed of multiple languages.

Adapter modules, which are incorporated into the layers of pre-existing models, are designed to master a particular task without altering the weights of the original model (Pfeiffer et al., 2020a). They are more parameter efficient than fine-tuning the full model while achieving nearly the same performance (Houlsby et al., 2019). Much like adapters that are trained for specific tasks, we can also train language-specific adapters. This is achieved by adding an adapter to a multilingual base model and then training it using Masked Language Modeling (MLM) (Pfeiffer et al., 2020b). If a task adapter is being trained with a multilingual base model, it is beneficial to also utilize a fixed language adapter because it captures and applies language-specific

[3] https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map

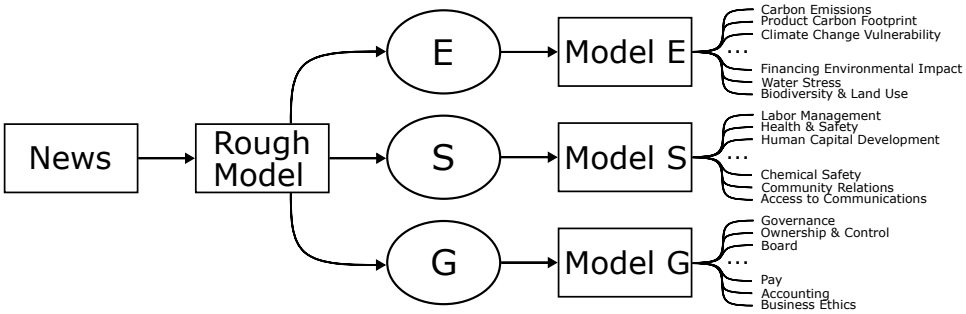

Figure 2: Schematic description of method 3. We first train a model to determine if a news article describes environmental, social or governance thematics. A second model then classifies the article with regards to the different key-issues belonging to that component. Each model in this case is constituted by a base-model, a language-adapter and a task-adapter as explained in subsection 3.1.

knowledge, which can enhance the final performance (Pfeiffer et al., 2020b).

3.2 Title or Content?

Classific. Target	Title	Content
Key-issues	0.63	0.41
Components	0.88	0.80

Table 1: Comparison of the F_1-macro scores when using the title vs using the content of the news articles for the French dataset. The first row shows results when directly classifying for the 35 key-issues while the second row classifies only for the rough E/S/G components.

We performed several experiments in order to determine if it is better to use the title or the content in order to classify the news articles. In Table 1, we show the F_1-macro scores when classifying the 35 key-issues directly in the first row, and the results for classifying the rough compontents (E/S/G) in the second row for the French dataset. In both cases, we achieve the best result when simply using the title to train the adapter. Since we observed a similar result for the English dataset, we decided to continue working without the "news_content" element.

3.3 Data Augmentation

Classific. Target	EN	FR
No Augmentation	0.58	0.69
Augmentation	0.68	0.67

Table 2: Comparison of the F_1-macro scores when augmenting the data by translating from the dataset in the other language and training on classifying the 35 key-issues directly. The top row shows the results without augmentation, the bottom row with augmentation.

Since we have data in two languages, we tried augmenting the data of each language by translating the data of the other language. For this, we use the OPUS-MT models published by (Tiedemann and Thottingal, 2020) from the huggingface-hub[4]. In Table 2, we show results on the key-issue classification for French and English. The results are conflicting, as we can see an improvement for the English dataset, while the French dataset performs slightly worse with the augmentation.

3.4 Configurations

We designed three different configurations:

- Method 1: Train an adapter on the 35 key-issues directly.
- Method 2: Augment the data by translating from the other language, then train on the 35 key-issues.
- Method 3: First train an adapter to classify an article into the ESG component, then train a set of three adapters, one for each ESG component, in order to classify the key-issues.

A rough schema for the third approach is shown in Figure 2. Since we achieved a better performance when translating for the English dataset (see Table 2), we decided to augment the data for the third approach for this language.
For all configurations, we evaluate the approach on 10% of the original training dataset. This is without augmentation, meaning the size of the eval dataset is only 5% the size of the total dataset for method 2.
As a base model, we use mBERT (Devlin et al.,

[4] https://huggingface.co/Helsinki-NLP

	Pre-Deadline			Post-Deadline	
Language	Method 1	Method 2	Method 3	Method 1	Method 3
EN	0.56	0.35	**0.57**	**0.61**	0.61
FR	0.75	0.73	**0.77**	0.78	**0.80**

Table 3: Weighted F_1-scores of the trained adapters for the test-set. On the left side, the official results. On the right side, the post-deadline results in which we evaluate on the test set during training.

2018). We then stack a pre-trained language adapter with fixed weights (from the AdapterHub [5], (Pfeiffer et al., 2020a)) on top of it followed by a task adapter (this setup is explained with more detail in (Pfeiffer et al., 2020b)). For training, we used a learning rate of $5 \cdot 10^{-5}$ and a simple cross-entropy loss-function.

4 Results

The submitted results of the three approaches are displayed on the left side of Table 3. Note that we display the weighted F_1-score here as opposed to F_1-macro we used in the previous section. For both languages method 3 shows the best performance, followed by method 1. However, while we placed third for French, the final placement for English was much worse. Our first guess at a reason for this was the imbalance of the training dataset, which we did not consider during the training. However, as we can see in Figure 1 in blue, the test data (which was released after the task-deadline) is similarly distributed as the training data.

4.1 Augmenting Key-Issues

Classific. Target	EN	FR
No Augmentation	**0.60**	0.77
Aug. Key-issue Model	0.58	**0.78**
Aug. Rough Model	0.59	0.77
Augmentation Both	0.57	0.77

Table 4: Comparison of weighted F_1-scores of the test-set when augmenting different parts of the data for method 3 (Figure 2). The first row shows the results without augmentation, the second row when augmenting only during training of the models classifying into the key-issues, the third row when augmenting only the first (rough) model and the fourth row when augmenting for all models.

In Table 3 we can see that the approach augmenting the data with the dataset of the other language (method 2) performs the worst for both languages. But since we augmented the English data for the method 3 and still measure good results (compared to method 1), we are unsure of the impact of the augmentation here. For that reason, we performed several tests where we train a model using augmented data at different stages. The results are shown in Table 4. We can see that the results are very similar among each language, especially for the French dataset where the configuration which augments the second models (classifying the key-issues) performs slightly better than the rest of the configurations. For the English dataset, the configuration without augmentation shows the best performance, while augmenting both models performs worst.

4.2 Evaluation on Test-Set

In order to determine the best performance possible with our setup, we train adapters on the whole training set, using the labelled test set to evaluate. The results are displayed in the right part of Table 3. Because method 2 performed the worst before, we do not include it here anymore. In addition, we do not augment the English approach for method 3 since we saw a better performance not augmenting in Table 4. We observe that the results improve slightly, but don't account for the difference in F_1-scores between the two languages.

5 Conclusion

We successfully trained several configurations capable of classifying news articles into the 35 key-issues defined by MSCI, showing that using the title instead of the content of the news article is more performant. We also tried to augment the datasets by translating from the other language but saw that this has little impact, even decreasing the performance in some cases. Out of the three different approaches, we observe that it is generally best to first classify the news articles into their rough ESG components (environmental, social & governance) before using a second model in order to determine the final key-issue.

[5] https://adapterhub.ml/

References

Tanja Aue, Adam Jatowt, and Michael Färber. 2022. Predicting Companies' ESG Ratings from News Articles Using Multivariate Timeseries Analysis. *arXiv*.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022. An Overview of Financial Technology Innovation. *Companion Proceedings of the Web Conference 2022*, pages 572–575.

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

Beatrice Crona. 2021. Sweet Spots or Dark Corners? An environmental sustainability examination of Big Data and AI in ESG. *SSRN Electronic Journal*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

Elena Escrig-Olmedo, María Ángeles Fernández-Izquierdo, Idoya Ferrero-Ferrero, Juana María Rivera-Lirio, and María Jesús Muñoz-Torres. 2019. Rating the Raters: Evaluating how ESG Rating Agencies Integrate Sustainability Principles. *Sustainability*, 11(3):915.

Jannik Fischbach, Max Adam, Victor Dzhagatspanyan, Daniel Mendez, Julian Frattini, Oleksandr Kosenkov, and Parisa Elahidoost. 2022. Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool. *arXiv*.

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. *arXiv*.

Tahir Islam, Rauf Islam, Abdul Hameed Pitafi, Liang Xiaobei, Mahmood Rehmani, Muhammad Irfan, and Muhammad Shujaat Mubarak. 2021. The impact of corporate social responsibility on customer loyalty: The mediating role of corporate reputation, customer satisfaction, and trust. *Sustainable Production and Consumption*, 25:123–135.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *arXiv*.

Zoltan Nagy, Linda-Eling Lee, and Guido Giese. ESG Ratings: How the Weighting Scheme Affected Performance.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A Framework for Adapting Transformers. *arXiv*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. *arXiv*.

Hanna Schramm-Klein, Joachim Zentes, Sascha Steinmann, Bernhard Swoboda, and Dirk Morschett. 2016. Retailer Corporate Social Responsibility Is Relevant to Consumer Behavior. *Business & Society*, 55(4):550–575.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Ilze Zumente and Jūlija Bistrova. 2021. ESG Importance for Long-Term Shareholder Value Creation: Literature vs. Practice. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(2):127.