Exploring Knowledge Composition for ESG Impact Type Determination

Fabian Billert[1,2] and **Stefan Conrad**[1]
[1]Heinrich-Heine University of Düsseldorf
[2]GET Capital AG
{fabian.billert, stefan.conrad}@hhu.de

Abstract

In this paper, we discuss our (Team HHU's) submission to the Multi-Lingual ESG Impact Type Identification task (ML-ESG-2). The goal of this task is to determine if an ESG-related news article represents an opportunity or a risk. We use an adapter-based framework in order to train multiple adapter modules which capture different parts of the knowledge present in the training data. Experimenting with various Adapter Fusion setups, we focus both on combining the ESG-aspect-specific knowledge, and on combining the language-specific-knowledge. Our results show that in both cases, it is possible to effectively compose the knowledge in order to improve the impact type determination.

1 Introduction

The substantial rise in Environmental, Social, and Governance (ESG) research over the past few years underscores the growing significance of sustainability in the corporate world (Zumente and Bistrova, 2021). Both investors and companies have a growing interest in ESG issues, as it becomes clearer that they are vital for a company's brand and, consequently, also value (Schramm-Klein et al., 2016), (Islam et al., 2021). Studies show that investor interest in a company's stock depends significantly on whether an estimation of their ESG-practices is available or not (Zumente and Lāce, 2021).
Different agencies have developed a variety of ESG-scoring mechanisms in order to quantify the sustainability practices of companies which offers investors an easier way to determine the ESG-related risk a company represents in their portfolio. Popular examples of this are MSCI[1] and Sustainalytics[2]. However, inconsistencies in the scoring practices and discrepancies between different score-providers cast doubt on the available ESG-scores (Berg et al., 2019), (Clément et al., 2023). This doubt could be alleviated by providing a more detailed overview of the various aspects of sustainability a company is involved in. Researchers in the field of natural language processing (NLP) have recently made efforts to enhance conventional ESG scoring methodologies by incorporating alternative data sources, aimed at offering a clearer understanding of a company's sustainability practices. (Nicolas et al., 2023) underscores the impact of ESG-risk events found in social media data, illustrating how they can result in adverse financial returns. Meanwhile, (Aue et al., 2022) focuses on analyzing news articles to construct a predictive model for determining ESG scores.
The ML-ESG series of shared tasks aims to advance research in this area of using natural language processing to provide more transparency in ESG topics. ML-ESG-1 focused on determining the ESG-aspects of news articles (Chen et al., 2023b). In the next step of the ML-ESG series, ML-ESG-2 aims to find the impact type of the information in news articles, meaning detecting if they represent opportunities or risks (Chen et al., 2023a).

In this paper, we present our submission for the French and English part of the ML-ESG-2 shared task. We train adapters on different parts of the dataset and experiment with several approaches in order to combine their knowledge using Adapter Fusion. We show that this knowledge composition yields a significant improvement of the performance compared to simply finetuning an adapter for the whole dataset. Our submitted approaches achieve fourth place for French and fifteenth place for English, however we investigated additional approaches after the task deadline which improve on

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

Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pages 79–83

those results.

2 Task Description and Dataset

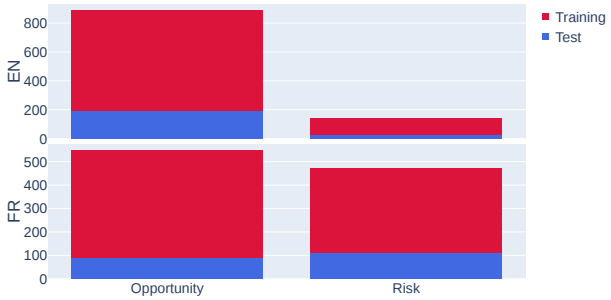


Figure 1: Occurrences of the different labels in the training- and test-data of the task. The top row represents the English data, the bottom row the French data. Training data is shown in red, test data in blue. **Note:** The test data was released after the task deadline and was not used during training.

ML-ESG-2 presents data in four different languages: English, French, Japanese and Chinese. Because the Japanese and Chinese labels are slightly different than the English and French ones, we chose to only participate in the latter languages. We display information about the training dataset in Figure 1 in red. For English, the training dataset contains 808 news articles while there are 818 news articles for French. Each news article has a news title, a news content and an associated impact type label. Different samples can have the same news title but a different news content and a different impact type, which happens a total of 100 times in the two datasets. For this reason, we didn't use the news title in our approaches.
The impact type is labelled as either "Opportunity" or "Risk". While the data is quite imbalanced for English, with the "Opportunity" label being the majority class, it is approximately even for French. The average length of the news content is 412 characters for English and 565 characters for French, which should fit well within the maximum length of 512 tokens of BERT-based models. Otherwise, the samples are truncated.

2.1 ESG-Aspect Dataset

The previous ML-ESG shared task focused on determining one out of 35 key issues defined by MSCI[3], where each key issue belongs to one ESG-aspect (E/S or G) (Chen et al., 2023b). The dataset shared there consists of 1200 news articles in English and French each, some of which overlap with the news articles used in the dataset we work on here. We present some approaches of using the knowledge available in the ML-ESG-1 dataset to improve the determination of "Opportunity" or "Risk" of the current task in subsection 3.3.

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

3 Experimental Approach

The basis of our approach consists of mBERT (Devlin et al., 2018), a multilingual BERT-based model. For fine-tuning, we use bottleneck-adapter-modules (Houlsby et al., 2019), which are a small set of weights inserted into the layers of the base-model. In order to account for the specific language we are training in, we implement an approach described in (Pfeiffer et al., 2020b), where a pre-trained language adapter is inserted before the adapter that is being fine-tuned (known as the task-adapter). We use the language adapters from the adapterhub (Pfeiffer et al., 2020a), which are available for all task-languages for the mBERT architecture. Adapter-modules allow us to be more flexible in training our model, as we will describe in subsection 3.2.
To train the adapters, we split the dataset and use 10% as the evaluation set. We use a learning rate of $5e-5$. Finally, to account for the label imbalance in the training data we can observe in Figure 1, we use a weighted loss function.

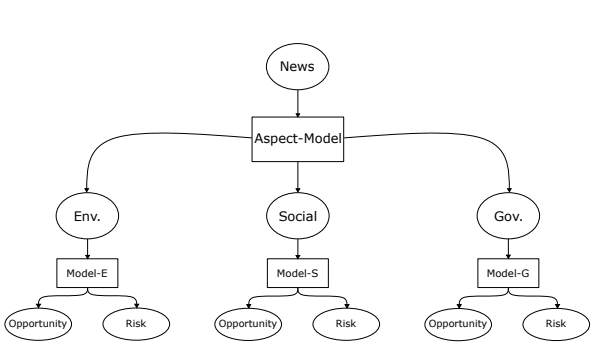
3.1 Data Augmentation

To augment the data available in the two different languages, we translate both of them to the other language respectively. Since translating the whole text at once sometimes produced artifacts, we split the news content into sentences using nltk's sentence tokenizer (Bird et al., 2009) and translate a single sentence at a time. In order to translate, we use the OPUS-MT models (Tiedemann and Thottingal, 2020) from the huggingface-hub[4].

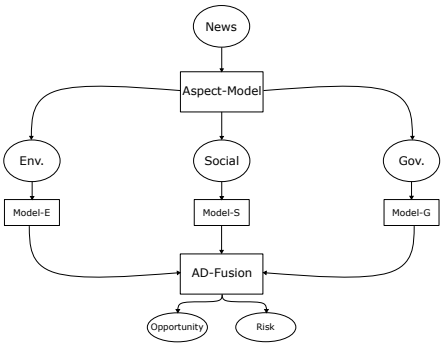
3.2 Adapter Fusion

In section 3, we explain our adapter-based approach, where we use a base model, a pre-trained language adapter and finally a task-adapter for fine-tuning. In this configuration, the different components are aligned in series. Adapters can also be used in parallel, which allows for the combination of the task-specific knowledge of each adapter.

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



(a) Schema of Method 2.



(b) Schema of Method 3.

Figure 2: Schemata for Methods 2 and 3. The Aspect Model is obtained using data from ML-ESG-1 (Chen et al., 2023b). The E/S/G-Models in Method 3 are the same ones we train in Method 2.

This is known as Adapter-Fusion. Adapter-Fusion demonstrates particular advantages in scenarios involving limited datasets, as it can place greater emphasis on the knowledge obtained from the task-adapters trained on larger datasets (Pfeiffer et al., 2021). In addition, Adapter-Fusion has shown to lead to performance-improvements in multilingual-scenarios as it can use knowledge from training in other languages (Billert and Conrad, 2023).
Since the concept of "Opportunity" and "Risk" can mean different things depending on the context, we aim to use the learnings from ML-ESG-1 (Chen et al., 2023b) in order to train aspect-specific adapters. These adapters will learn more specifically to determine "Opportunity" and "Risk" in the environmental, social and governance context. In practice this means that we will first train a model to determine the rough ESG-aspect of a news-article (environmental, social, governance), before training three different aspect-specific adapters to determine the ESG impact type for articles of each aspect. We achieve an F_1-Macro score of around 0.86 for English and 0.79 for French for the adapters classifying the ESG-aspect, meaning that there is still the possibility of misclassifying the aspect in the first place, which might be detrimental for the downstream adapters predicting the impact type.
After training the three aspect-specific adapters, it is then possible to train an Adapter Fusion layer to make a final prediction for the ESG impact type.

3.3 Configurations

In order to test the premise that the ESG impact type depends on the aspect (as we describe in the previous subsection), we designed three different configurations.

- Method 1: Train an adapter on the two impact type labels directly.
- Method 2: Determine the ESG-aspect first, then train an adapter for the news articles belonging to each aspect separately in order to determine the impact type.
- Method 3: Use the same approach as in Method 2, but use Adapter-Fusion in order to combine the knowledge of the three separate adapters.

Method 2 and method 3 are depicted in Figure 2. Note that for the governance aspect-adapters, we augment the evaluation data with the governance-evaluation data from the other language respectively, because the evaluation sets have a very small amount of samples.

4 Results

The test data is shown in Figure 1 in blue. For English, the label imbalance in the training set is also present in the test set. For French, while there is still no clear imbalance, the "Risk" label now occurs more often as opposed to the "Opportunity" label in the training set.
In Table 1, we show the F_1-Macro scores of the various approaches. The three methods we submitted can be seen on the left side of the table. For English, method 1 and method 3 have the same score, with method 2 being much worse than both of them. For French, method 1 shows the best performance, with method 3 being slightly worse. Again, method 2 trails far behind.
Interestingly, method 2 shows a big discrepancy to method 3, even though the same adapters are used, the only difference being the Adapter-Fusion layer.

	Pre-Deadline Results			AD-Fusion Experiments	
Language	Method 1	Method 2	Method 3	w/o gov	Lang-Fusion
EN	<u>0.8098</u>	0.4225	<u>0.8098</u>	**0.8771**	0.8557
FR	<u>0.7548</u>	0.6169	0.7457	0.7726	**0.8084**

Table 1: F_1-Macro scores of the trained adapters for the test-set. On the left side, the submitted results. On the right side, there are additional experiments that were done after the deadline. Bold values represent the best achieved score, while underlined values represent the best submitted score for each language.

4.1 Aspect-Fusion

Since there was not a lot of training data for the governance adapter in methods 2 and 3, we suspect that it is detrimental to the adapter fusion performance. We do several experiments without the governance adapter, using only the environmental and social adapters for the Adapter Fusion. The results are depicted on the right side of Table 1. We can see that the performance surpasses both the previous score of method 3 as well as the score of method 1, which confirms our previous suspicions about the governance adapter.

For English, we note an improvement of almost 7 points while for French, we see an increase slightly lower than 2 points. In both cases, the performance is significantly higher than the previous highest result, demonstrating that the concept of ESG impact type depends on the ESG aspect and that the trained Adapter Fusion layer successfully composes the knowledge learned by our aspect-specific adapters.

4.2 Language-Fusion

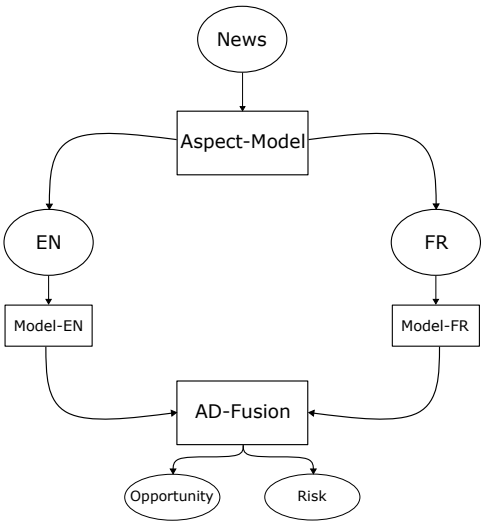


Figure 3: Adapter Fusion on the different language adapters from Method 1.

As mentioned in subsection 3.2, it is also possible to use Adapter Fusion to combine information learned from tasks in different languages. To test if this also holds true in our case, we train an Adapter Fusion layer consisting of both language-specific adapters of method 1. The setup is displayed schematically in Figure 3. This is a similar approach as was used in (Billert and Conrad, 2023), although we only work with two languages in this case.

In the rightmost column of Table 1, we can see that this configuration achieves an outperformance of about 5 points over just using the language-specific adapters of method 1. Compared to using the aspect-fusion approach described in subsection 4.1, the results are conflicting. For English, the aspect-fusion shows better results, but for French, using the language-fusion seems superior.

5 Conclusion

In this work, we present multiple approaches using adapters to determine the ESG impact type of news articles. We show that it is possible to gain significant performance increases by using Adapter Fusion, which allows us to combine the knowledge present in the separate adapter modules, by experimenting with implementing Adapter Fusion in two different ways: Firstly, using the prior knowledge about ESG aspects, we train an adapter for each aspect and fuse them to determine the ESG impact type. Secondly, we fuse the adapters trained on the different languages in order to combine the knowledge learned from the different language datasets. Further studies could focus on how to combine the two Adapter Fusions, which we did not attempt in this work due to time constraints. In addition, it might be interesting to use the Japanese and Chinese datasets to try to improve the language-fusion approach even further. Finally, putting more focus on augmenting the training data, e.g. by using paraphrases, could bridge the gap between this study and the top-placed systems.

References

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

Florian Berg, Julian F Kölbel, and Roberto Rigobon. 2019. Aggregate Confusion: The Divergence of ESG Ratings. *SSRN Electronic Journal*.

Fabian Billert and Stefan Conrad. 2023. HHU at SemEval-2023 Task 3: An Adapter-based Approach for News Genre Classification. *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1166–1171.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

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

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 (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*, pages 111–115, Macao.

Alexandre Clément, Élisabeth Robinot, and Léo Trespeuch. 2023. The use of ESG scores in academic literature: a systematic literature review. *Journal of Enterprising Communities: People and Places in the Global Economy*, ahead-of-print(ahead-of-print).

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

Maxime L D Nicolas, Adrien Desroziers, Fabio Caccioli, and Tomaso Aste. 2023. ESG Reputation Risk Matters: An Event Study Based on Social Media Data. *arXiv*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

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. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

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.

Ilze Zumente and Natalja Lāce. 2021. ESG Rating—Necessity for the Investor or the Company? *Sustainability*, 13(16):8940.