A low resource framework for Multi-lingual ESG Impact Type Identification

N Harsha Vardhan[*], **Sohom Ghosh**[†], **Ponnurangam Kumaraguru**[*], **Sudip Kumar Naskar**[†]
[*]International Institute of Information Technology, Hyderabad, India
[†]Jadavpur University, Kolkata, India
nemani.v@research.iiit.ac.in, sohom1ghosh@gmail.com
pk.guru@iiit.ac.in, sudip.naskar@gmail.com

Abstract

With the growing interest in Green Investing, Environmental, Social, and Governance (ESG) factors related to Institutions and financial entities has become extremely important for investors. While the classification of potential ESG factors is an important issue, identifying whether the factors positively or negatively impact the Institution is also a key aspect to consider while making evaluations for ESG scores. This paper presents our solution to identify ESG impact types in four languages (English, Chinese, Japanese, French) released as shared tasks during the FinNLP workshop at the IJCNLP-AACL-2023 conference. We use a combination of translation, masked language modeling, paraphrasing, and classification to solve this problem and use a generalized pipeline that performs well across all four languages. Our team ranked 1st in the Chinese and Japanese sub-tasks.

1 Introduction

In recent times, the focus on Institutions' Environmental, Social, and Governance factors (ESG) has garnered increased interest from the global investment and corporate governance communities. People have also grown to be socially responsible and environmentally conscious while investing. ESG serves as a third dimension beyond risk and return. Research also indicates that Institutions with better ESG performance directly correlate to better stock performance and risk management (Whelan and Atz, 2021). Keeping this in mind, many rating agencies quantify the nature and impact of ESG aspects of an institution and publish ratings (Serafeim and Yoon, 2022). Apart from ESG investing, Impact investing (Berk and van Binsbergen, 2021) has also gained traction where investors, instead of investing solely based on ESG benefits, would look for a combination of better returns as well as a positive influence in society. Hence, impact identification is crucial to determine whether statements are an opportunity or a risk for the Institution.

Most of the scoring processes involved in ESG and Impact assessments are extremely time-consuming and require expert involvement and manual annotations. To automate this, we propose a generalized pipeline capable of predicting the impact types of ESG-related news articles (as shown in Figure 1). This generalized pipeline can be scaled to other low-resource datasets as well.

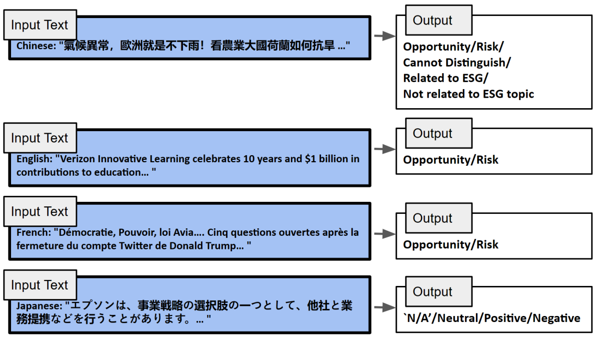
Input Text
Chinese: "氣候異常，歐洲就是不下雨！看農業大國荷蘭如何抗旱 ..."
Output
Opportunity/Risk/
Cannot Distinguish/
Related to ESG/
Not related to ESG topic
Input Text
English: "Verizon Innovative Learning celebrates 10 years and $1 billion in contributions to education... "
Output
Opportunity/Risk
Input Text
French: "Démocratie, Pouvoir, loi Avia.... Cinq questions ouvertes après la fermeture du compte Twitter de Donald Trump... "
Output
Opportunity/Risk
Input Text
Japanese: "エプソンは、事業戦略の選択肢の一つとして、他社と業務提携などを行うことがあります。... "
Output
`N/A'/Neutral/Positive/Negative

Figure 1: The Multilingual ESG Impact Assessment Task.

The labels primarily indicate if the given news is an opportunity or a risk from the ESG aspect. In this shared task, we participated in all four languages and were ranked 1st in Chinese and Japanese sub-tasks, 4th in French, and 7th in English.

2 Related Work

With the advent of green investing, many approaches and models have been developed to automate processes in Financial and ESG-based NLP research, including the development of models like FinBERT (Araci, 2019), ESGBert (Mehra et al., 2022), etc. While there has been much work on ESG-type classifications, including on multilingual datasets, more work needs to be done on impact-type classifications. FinNLP 2023 (Chen et al., 2023) focuses on a similar task where participants

Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pages 57–61
November 1, 2023. ©2023 Association for Computational Linguistics

were required to classify multilingual data into the ESG issue type, where the best results were obtained by using language-specific BERT models along with data augmentations using Large Language Models. Furthermore, it's important to note that extensive research has been conducted on sentiment analysis (Pasch and Ehnes, 2022; Aue et al., 2022), which can be considered a fundamental aspect of impact identification. Attempts have also been made for impact identification in Chinese (Tseng et al., 2023) and Japanese (Kannan and Seki, 2023).

3 Task Description

The task is primarily a classification task where given a text, classify whether the text poses a risk or an opportunity for the company. As shown in Figure 1, there are multiple languages with differences in classes.

4 Data

The dataset primarily contains news articles collected from four different languages, English (en), Chinese (zh), Japanese (ja), and French (fr), along with their impact types.

Language	Train	Test	C	W_c	W_h
English	808	218	2	412.48	76.83
Chinese	1400	156	5	-	33.68
Japanese	896	225	4	-	78.82
French	818	200	2	564.88	96.17

Table 1: Metrics across languages. C denotes the number of Classes, W_c denotes the average character length of content and W_h denotes the average character length of headline. Chinese and Japanese datasets do not have content columns.

Given that the dataset across languages is small, and the classwise distribution is highly skewed. To overcome these challenges, we use a combination of translation and data paraphrasing on minority classes.

5 Approaches

We primarily used encoder-based models for this classification task. Given the limited sample size of the dataset, variations in languages, and disparity with class distribution among different languages, We tried to make a pipeline that accounted for such differences and performed consistently well across all languages. We tried a variety of approaches like Masked Language Modelling (MLM), Paraphrasing for augmenting the minority classes, Translation, and Multilingual Models and used a combination to finalize our pipeline based on empirical experiments.

All of the experiments have been run using a batch size of 32, a learning rate of $2e^{-5}$, weight decay of 0.01, and for ten epochs. The reported metrics are based on $80 : 20$ train-test set splits with a constant random seed and not on the final validation sets used for the leaderboard. The code, data, and models used for inferences are available at the link.

5.1 Masked Language Modelling

We performed several experiments to decide the necessary models for classification. Also, we experimented with pre-training the models beforehand on the ESG corpus, which was the English dataset for the Multi-Lingual ESG Issue Identification (ML-ESG) (Chen et al., 2023) and then using the fine-tuned models for classifications. We noticed that across all languages, the models pre-trained on the ESG corpus and then fine-tuned for classification outperformed those fine-tuned for classification.

Approach	Title	Content
Classification	74.89%	92.48%
MLM + Classification	85.48%	93.16%

Table 2: Comparison across Classification and MLM + Classification approaches along with news headlines and content using bert-base-cased (Devlin et al., 2019) model. These reported numbers are the weighted $F1$ with the English dataset.

From Table 2, we also observe that using news content for training over title performs better. The French dataset exhibits similar trends, and hence, for all further analysis, we use the news content for English and French and the news title for Chinese and Japanese since they do not have news content available in the dataset.

5.2 Translation and Multilingual Models

We have also experimented with specific language models vs. translating and English-based models primarily due to a larger number of specialized models pre-trained on ESG data being available in English. We used Google Translate to translate data from French, Chinese, and Japanese and leveraged this data as additional data while training for models. Also, by using English, we were able to use paraphrasing tools to augment and extend the minority classes of our dataset.

Approach	F1
Translated	68.92%
Chinese	68.45%

Table 3: Comparison of weighted F1 scores while using translated Chinese to train a bert-base-cased model vs. using Chinese data to train a bert-base-multilingual-cased model (Devlin et al., 2019).

While the disparity between the translated text and the original language may not seem substantial, there exists a possibility that employing more specialized language models tailored to the Chinese language could have potentially delivered better results. However, this approach would have restricted our ability to employ paraphrasing-based techniques, as such tools are not as readily available in non-English languages. Furthermore, it would have limited our access to English models predominantly trained on ESG data. Accordingly, our primary strategy revolved around using translated text for classification.

5.3 Paraphrasing for Data Augmentation

Given that the dataset across languages is small and the classwise distribution is highly skewed, one of the approaches we considered for improving the classification task is to augment the minority classes and extend the dataset. While rule-based paraphrasers are popular and widely used for such tasks, the variation within sentences is frequently minor and only offers a slight improvement during training. Hence, we considered a T5-based paraphraser (Vladimir Vorobev, 2023), primarily fine-tuned on ChatGPT paraphrases. It offers a better range of sentence variations than any other approaches tried. We first translated the dataset from the respective languages to English and then generated paraphrased data on minority class data (For each minority instance, approximately 3-4 paraphrases were created, depending on the specific count of instances for that particular label. For the same reason we did not paraphrase for french language since the label distribution was already uniform.The paraphrased data can be accessed here.) and used this along with the original data for training the classification model.

We observe that across languages, paraphrased data improved the F1 metrics of models to a great extent. This effect was more prominent in Chinese and Japanese datasets, where the number of classes was more prominent, and there was a wider class disparity. This supports our choice of using translated text rather than the original despite lackluster results while just translating and using that data for classification.

Approach	F1 (en)	F1 (zh)
Paraphrased Data	98.91%	84.98%
Original Data	93.16%	68.45%

Table 4: Comparison of weighted F1 while using paraphrased text vs. original dataset for MLM + Classification on the English dataset and The original dataset for Chinese and the translated + Paraphrased version of the Chinese dataset. bert-base-cased model was used for English and bert-base-multilingual-cased for Chinese.

6 Final System Description

For the final system that was used, based on the empirical studies performed above, We used a pipeline that initially translated all of the given text into English using Google Translate. Then we use the T5-based paraphraser (Vladimir Vorobev, 2023) to generate new minority class instances. We also use an ESG corpus to initially pre-train a model on this corpus and then fine-tune it for classification on the translated and augmented dataset. Figure 2 shows the exact process.

We also performed more experiments to decide which models best performed on the English dataset and chose bert-base-cased (Devlin et al., 2019), Finbert (Araci, 2019), and finbert-tone (Huang et al., 2023). We used the same models for the other languages as well. The model hyper-parameters are the same as mentioned in the methodology.

We observe that despite using a generalized pipeline and models for all the languages, the results are good. Table 5 shows the performance of models for all of the languages and models used.

Models	F1(en)	F1 (zh)	F1 (ja)	F1 (fr)
BBC	98.91%	84.98%	89.64%	78.25%
FB	97.82%	85.31%	91.13%	78.55%
FBT	98.91%	82.26%	89.54%	70.79%

Table 5: Final weighted F1 metrics for the models used for submission. BBC = bert-base-cased, FB = FinBERT, FBT = FinBERT Tone

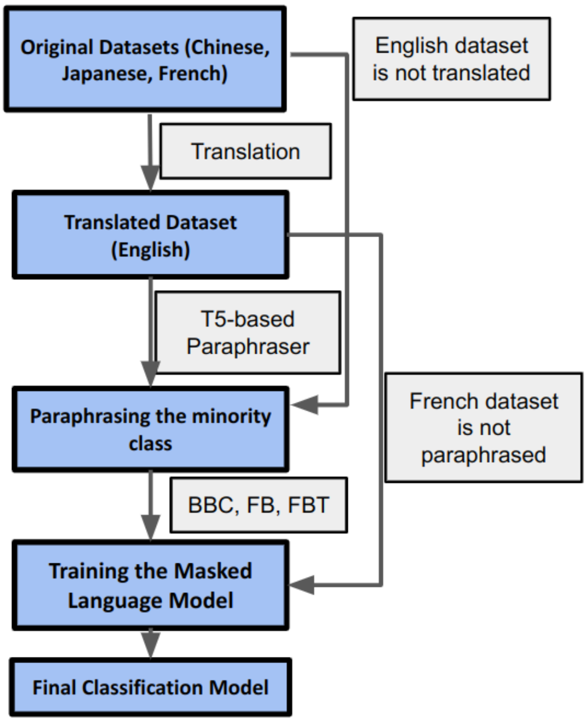
Original Datasets (Chinese, Japanese, French)
English dataset is not translated
Translation
Translated Dataset (English)
T5-based Paraphraser
Paraphrasing the minority class
French dataset is not paraphrased
BBC, FB, FBT
Training the Masked Language Model
Final Classification Model

Figure 2: The final system pipeline. BBC = bert-base-cased, FB = FinBERT, FBT = FinBERT Tone

7 Conclusion

Comparing the performance of our models with that of other participants, we conclude that our models performed consistently well. We outperformed all other teams in the Chinese and Japanese sub-tasks. One unique feature is despite four different languages, we were able to use the same pipeline and same set of models and achieve consistently good results across languages, which leads us to believe that the pipeline is performant for low resource settings. All of the data generated and code used can be accessed here.

Limitations

The primary challenge highlighted in the paper's approaches is the translation process. While it expands the possibilities, it also comes with a drawback - the loss of language-specific nuances and information. Integrating language-specific paraphrasing tools and access to Environmental, Social, and Governance (ESG) datasets tailored to those languages could enable us to adapt the existing pipeline. This adaptation would involve incorporating regional language models instead of relying solely on English models, potentially resulting in improved performance.

We also did not evaluate larger models due to time and feasibility constraints, but larger models would have provided better results. Also, since the number of classes differed across languages, training a singular multilingual model or similar approaches resulted in poor metrics for some languages. Hence we did not pursue this direction.

One of the initial choices for selecting news content as the primary choice for the classification approach could also have been flawed. Since headlines are generally more read and captivating, it might have provided a polarized view of the instance and might have been easier to categorize as an Opportunity or Risk. Idealistically, some form of ensemble modeling between headlines and content might improve the performance of the present approach.

Ethics Statement

In conducting this research, we have not encountered any significant ethical concerns or considerations that would require special attention in this paper. Our study focuses on impact type classification of ESG-related publically available news instances, and the data and methods employed adhere to established ethical guidelines and standards within the field of computational linguistics.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Tanja Aue, Adam Jatowt, and Michael Färber. 2022. Predicting companies' esg ratings from news articles using multivariate timeseries analysis.

Jonathan B. Berk and Jules H. van Binsbergen. 2021. The Impact of Impact Investing. Research Papers 3981, Stanford University, Graduate School of Business.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Overview of the FinNLP-2023 ML-ESG task: Multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Allen H. Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text*. *Contemporary Accounting Research*, 40(2):806–841.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for esg scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. ESGBERT: Language model to help with classification tasks related to companies' environmental, social, and governance practices. In *Embedded Systems and Applications*. Academy and Industry Research Collaboration Center (AIRCC).

Stefan Pasch and Daniel Ehnes. 2022. Nlp for responsible finance: Fine-tuning transformer-based models for esg. pages 3532–3536.

George Serafeim and Aaron Yoon. 2022. Stock price reactions to esg news: the role of esg ratings and disagreement. *Review of Accounting Studies*, 28.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management*, CIKM '23, New York, NY, USA. Association for Computing Machinery.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

Tensie Whelan and Ulrich Atz. 2021. Esg and financial performance : Uncovering the relationship by aggregating evidence from 1 , 000 plus studies published between 2015 – 2020.

A Appendix

In this section, we present the classwise F1 metric for models based on the language. The models in consideration are:

- bert-base-cased (BBC)
- Finbert (FB)
- Finbert-Tone (FBT)
- **English**

Models	Label 0	Label 1
BBC (LIP1)	99%	99%
FB (LIPI2)	98%	98%
FBT (LIPI3)	99%	99%
Support	146	130

Table 6: English Language Model Metrics

Note: For English, Label 0 denotes Opportunity and Label 1 signifies Risk.

- **French**

Note: Label 0 stands for Opportunity and Label 1 represents Risk. The Support is 20% of the training set as the French dataset has an almost equal class distribution.

Models	Label 0	Label 1
BBC (LIP1)	82%	74%
FB (LIPI2)	81%	76%
FBT (LIPI3)	76%	65%
Support	88	76

Table 7: French Language Model Metrics

- **Chinese**

Models	0	1	2	3	4
BBC (LIP1)	85%	95%	75%	84%	83%
FB (LIPI2)	84%	93%	86%	83%	86%
FBT (LIPI3)	80%	89%	88%	80%	83%
Support	132	58	29	122	55

Table 8: Chinese Language Model Metrics

Note: In Chinese, Label 0 is Opportunity, 1 is Risk, 2 is Cannot Distinguish, 3 is Related to ESG but unrelated to the company, and 4 is Not Related.

- **Japanese**

Models	Label 0	Label 1	Label 2	Label 3
BBC (LIP1)	88%	86%	92%	97%
FB (LIPI2)	91%	88%	91%	96%
FBT (LIPI3)	90%	85%	88%	97%
Support	86	69	57	49

Table 9: Japanese Language Model Metrics

Note: For Japanese, Label 0 is Positive, 1 is "Not Available", 2 is Neutral, and 3 is Negative.

All metrics are derived from the paraphrased testset, which forms part of the publicly accessible training set. For further details on the training data, refer to this repository.