Fine-tuning Language Models for Predicting the Impact of Events Associated to Financial News Articles

Neelabha Banerjee[1], Anubhav Sarkar[1], Swagata Chakraborty[1], Sohom Ghosh[2], Sudip Kumar Naskar[2]
[1]Christ (Deemed to be University), [2]Jadavpur University
India
{neelabha.12.banerjee, sarkaranubhav2001, swagatac652}@gmail.com
{sohom1ghosh, sudip.naskar}@gmail.com

Abstract

Investors and other stakeholders like consumers and employees, increasingly consider ESG factors when making decisions about investments or engaging with companies. Taking into account the importance of ESG today, FinNLP-KDF introduced the *ML-ESG-3* shared task, which seeks to determine the duration of the impact of financial news articles in four languages - English, French, Korean, and Japanese. This paper describes our team, LIPI's approach towards solving the above-mentioned task. Our final systems consist of translation, paraphrasing and fine-tuning language models like BERT, Fin-BERT and RoBERTa for classification. We ranked first in the impact duration prediction subtask for French language.

Keywords: ESG, financial natural language processing, impact prediction, language models, ESG impact prediction

1. Introduction

The Multi-Lingual ESG Impact Duration Inference (ML ESG-3) task being organised in conjunction with the FinNLP-KDF@LREC-COLING-2024 deals with predicting the impact of events on companies. Determining the duration of an impact, an event might have on a company in the context of Environmental Social and Governance (ESG) factors could be crucial for understanding and managing the risks or opportunities associated with that event. Predicting the duration of an impact might involve fine-grained analysis of historical data, sentiment analysis, and other relevant information from news articles. In this paper, we talk about our team LIPI's approach of solving the subtasks of ML ESG- 3. This can be the first step towards achieving the long-term goal of developing multilingual systems that can assess the potential short-term and long-term effects of specific events on a company's performance, reputation, or other ESG-related aspects. We present this in Figure 1.

Our contributions

Our contributions include developing a framework that finetunes pre-trained language models for classifying the impact and duration of an event associated with multi-lingual news articles. We open-sourced the code[1] so that the research community can utilize them as baselines.

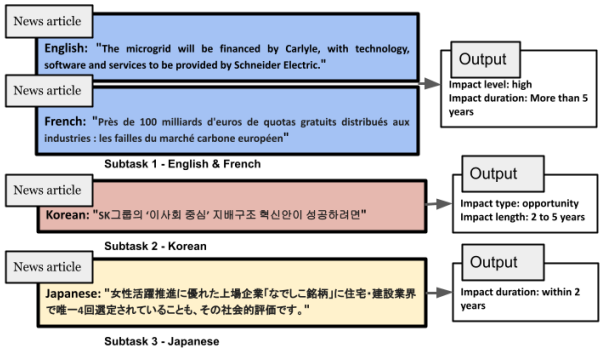

Figure 1: Overview of the ML-ESG3 task

2. Problem Statement

The multilingual data set of the shared task ML-ESG-3[2] consists of financial news articles in different languages such as English, French, Japanese, and Korean (Chen et al., 2024) (Kannan and Seki, 2023). The design of the task varies slightly across different languages. It is described as follows:

- **English and French**: Given a financial news article in English or French, the objective is to determine its *impact level* and predict its *impact length*. The impact length can be "low",

This paper got accepted at FinNLP-KDF-ECONLP workshop of LREC-COLING 2024.

[1]https://github.com/Neel-132/ML-ESG3_LIPI

[2]https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-kdf-2024/shared-task-ml-esg-3 (accessed on 3rd Feb 2024)

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 244–247

"medium" or "high". The impact length can be "Less than 2 years", "2 to 5 years", or "More than 5 years".

- **Japanese**: Given a financial news article in Japanese, the objective is to predict its *impact duration*. The impact duration can be "Less than 2 years", "2 to 5 years", or "More than 5 years".
- **Korean**: Given a financial news article in Korean, the goal is to determine its *impact type* and predict its *impact length*. The impact type can be between "opportunity", "risk", or "cannot distinguish" and the impact length can be "less than 2 years", "2 to 5 years", or "more than 5 years".

3. System Descriptions

The pipeline for handling the tasks mentioned above comprises the following steps:

- Step 1: **Translation** - Although there are several powerful multilingual encoder models present, our experiments revealed that they were not very efficient in learning the intricate patterns in the dataset and thereby correctly predicting the impact type and duration of news articles. Thus, we primarily translated the non-English datasets into English before proceeding with modelling.
- Step 2: **Paraphrase** - We found that as the given data set was small, the classification models were overfitting. To solve this, we paraphrased the translated data set returned by the translation module as mentioned in Step 1 using a T5-based model (Vladimir Vorobev, 2023).
- Step 3: **Classification** - After paraphrasing comes the final module of the pipeline. This is the classification module. Since the target variable differed slightly across different datasets, we designed two different classification modules for the three tasks given as follows:
  - Module 1 (for English, French & Korean): The English, French, and Korean dataset has two target variables. For English and French, they are *impact level* and *impact length*. For Korean, they are *impact type* and *impact length*. We used pre-trained encoder models like BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), etc. to learn the embeddings of the content as given by the paraphrase module, followed by a linear layer to predict the target which can be impact length, impact type, or impact level. The number of classes in each of these target variables is used as a hyperparameter to specify the output of the linear layer.
  - Module 2 (For Japanese): The Japanese dataset has only one target variable, *impact duration*. The *impact type* was given for this dataset. So, we developed the second module to learn the pre-trained text embeddings using the same encoder models, but for two features which are news content and impact type, followed by a concatenation operation. Finally, we added a linear layer to predict the output.

We present this in Figure 2.

Dataset	Model	macro F-1	micro F-1
English	XG-Boost	**0.35**	**0.31**
	SVM	0.29	0.26
	DNN	0.32	0.27
French	XG-Boost	0.23	0.22
	SVM	0.21	0.21
	DNN	**0.33**	**0.34**
Japanese	XG-Boost	**0.12**	**0.09**
	SVM	0.08	0.05
	DNN	0.11	0.10
Korean	XG-Boost	0.34	0.34
	SVM	0.27	0.22
	DNN	**0.42**	**0.34**

Table 1: Result of the Baselines

4. Experiments and Results

In this section, we describe the experiments we performed, and the corresponding results.

4.1. Baseline

For the baseline, we chose BERT-base uncased (for English) and BERT-base multilingual (Devlin et al., 2018) uncased (for other languages) to learn the pre-trained embeddings of news content and used them to train classical machine learning algorithms like XG-Boost (Chen and Guestrin, 2016), Support Vector Machine (Cortes and Vapnik, 1995), and deep learning based algorithms like Multi-layered Perceptron with just one hidden layer.
The results corresponding to it are presented in Table 1.

4.2. Experiment 1

The first experiment towards improving on the baseline had three stages, depending on the language of the dataset. For the non-English datasets like

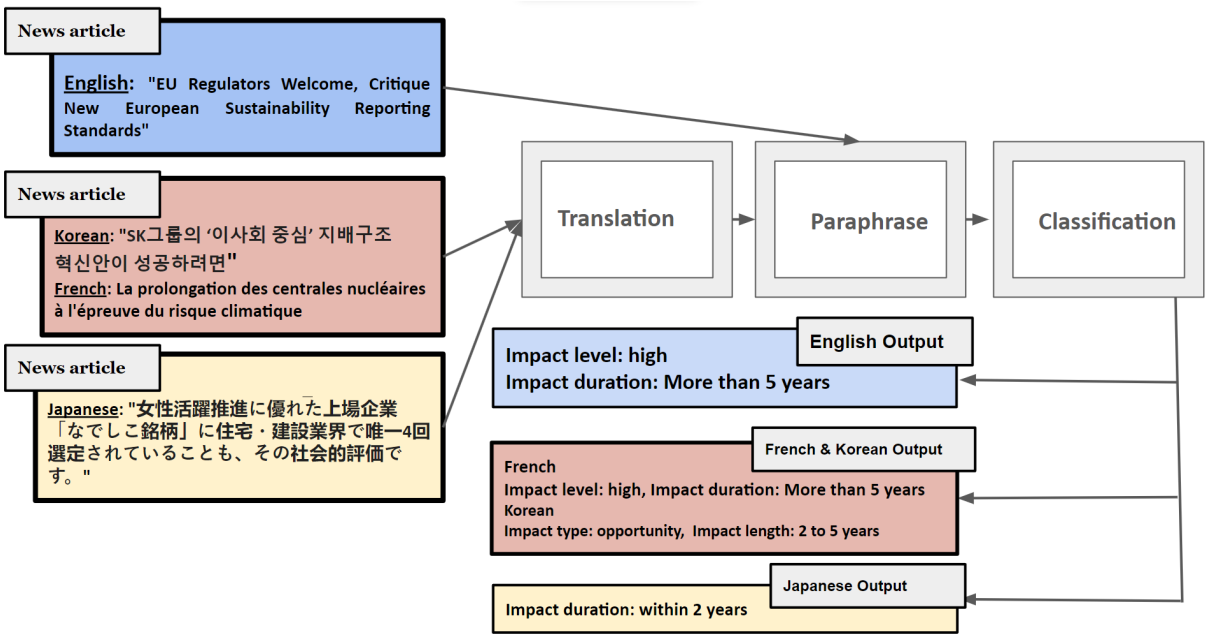

Figure 2: Our framework

French, Korean, we firstly translated the news content into English using Google Translate. In the next step, we paraphrased each data point using a T5 based paraphraser (Vladimir Vorobev, 2023) with a beam size of 5, temperature of 0.7, and repetition penalty of 10.
In the final step, we fine-tuned pre-trained encoder models like BERT (Devlin et al., 2018), DistillBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2020), Fin-BERT (Araci, 2019), etc. for the task of classifying the news articles to their respective impact type/level. We used a learning rate of e^{-5}, and a weight decay of 0.005 and fine-tuned the models for 30 epochs. Our best-performing models were BERT-base-uncased (Devlin et al., 2018) for English, RoBERTa (Liu et al., 2020) for Korean, and FinBERT (Araci, 2019) for French .
The results are presented in Table 2.

Dataset	Model	macro F-1	micro F-1
English	BERT-base-uncased	**0.99**	**0.99**
	FinBERT	0.97	0.97
	DistillBERT-multiling	0.70	0.68
	DistillBERT-base	0.68	0.69
	NLI-Distilroberta-base	0.81	0.80
	Distilroberta Financial	0.75	0.78
	XLI Roberta base	0.83	0.81
	RoBERTa-base	0.98	0.97
Korean	BERT-base-uncased	**0.95**	**0.94**
	FinBERT	0.94	0.93
	DistillBERT-multiling	0.78	0.71
	DistillBERT-base	0.76	0.69
	NLI-Distilroberta-base	0.82	0.81
	Distilroberta Financial	0.67	0.64
	XLI Roberta base	0.75	0.71
	RoBERTa-base	0.96	0.93
French	BERT-base-uncased	0.93	0.93
	FinBERT	**0.94**	**0.93**
	DistillBERT-multiling	0.57	0.49
	DistillBERT-base	0.91	0.90
	NLI-Distilroberta-base	0.51	0.45
	Distilroberta Financial	0.47	0.46
	XLI Roberta base	0.63	0.67
	RoBERTa-base	0.91	0.92

Table 2: Results of Experiment-1

4.3. Experiment 2

Since the Japanese dataset had only one objective, i.e. to predict the *impact duration*, we used the *impact type* as another feature along with the news content. Like the first experiment mentioned above, we translated the data into English, followed by paraphrasing with the same model, and configurations as mentioned in Experiment 1. Finally, we fine-tuned pre-trained models mentioned in Experiment 1 for assessing impact duration of news articles in Japanese.
Furthermore, we concatenated the embeddings of news content and impact type followed by a linear layer before the final output layer. We used a learning rate of e^{-4} and a weight decay of 0.006 and trained the models for 30 epochs. Our top performing models were BERT-base-uncased (Devlin et al., 2018), RoBERTA-base (Liu et al., 2020) and FinBERT (Araci, 2019).
The results are presented in Table 3.

4.4. Experiment 3

Since the English and French datasets had the same objective of predicting the impact level and impact length, we experimented with fine-tuning the pre-trained models (mentioned in both of the

Dataset	Model	macro F-1	micro F-1
Japanese	BERT-base-uncased	**0.67**	**0.69**
	FinBERT	0.55	0.52
	DistillBERT-multiling	0.52	0.48
	DistillBERT-base	0.36	0.32
	NLI-Distilroberta-base	0.51	0.48
	Distilroberta Financial	0.43	0.48
	XLI Roberta base	0.49	0.51
	RoBERTa-base	0.68	0.67

Table 3: Results of Experiment-2

Dataset	Model	macro F-1	micro F-1
English and French	BERT-base-uncased	**0.79**	**0.79**
	FinBERT	0.67	0.62
	DistillBERT-multiling	0.34	0.41
	DistillBERT-base	0.51	0.55
	NLI-Distilroberta-base	0.57	0.61
	Distilroberta Financial	0.47	0.45
	XLI Roberta base	0.51	0.50
	RoBERTa-base	0.76	0.76

Table 4: Results of Experiment-3

previous experiments) on the English dataset and testing them on the French dataset. The hyperparameters were the same as those of Experiment 2 and the results corresponding to it are mentioned in Table 4.

5. Conclusion

In this paper, we share our team, LIPI's approach for determining the duration of an event's impact on the company. We translated the non-English datasets into English and further paraphrased them before fine-tuning the encoder-based pre-trained language models on them. Our observations revealed the best performing models were BERT(Devlin et al., 2018) for English and Japanese; RoBERTa (Liu et al., 2020) for Korean, and FinBERT(Araci, 2019) for French. We achieved a significant increase in performance with translation and paraphrasing. Finally, we proposed a unified framework for all the languages.

Our team ranked 3^{rd} in both of the sub-tasks of the English dataset, 1^{st} in the first sub-task(impact-length) and 8^{th} in the second sub-task(impact-level) of the French dataset, 20^{th} in the first sub-task(impact-length) and 13^{th} in the second sub-task(impact-type) of the Korean dataset, and 11^{th} in the Japanese dataset.

However, we did not consider the semantic loss while paraphrasing and also had to translate the dataset into English to seek improvement. The future scope of this paper involves, designing better language models for low-resourced languages, improving the computational aspect of the algorithms, and extending the solution to cater to bigger and more important needs.

6. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-lingual esg impact duration inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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 and the Second Multimodal AI For Financial Forecasting*, pages 45–54, Macao. -.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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