Duration Dynamics: Fin-Turbo's Rapid Route to ESG Impact Insight

Xinyun Rong, Weijie Yang
roxyrong@berkeley.edu, raphaelyang1998@berkeley.edu

Abstract

This study introduces "Duration Dynamics: Fin-Turbo's Rapid Route to ESG Impact Insight", an innovative approach employing advanced Natural Language Processing (NLP) techniques to assess the impact duration of ESG events on corporations. Leveraging a unique dataset comprising multilingual news articles, the research explores the utility of machine translation for language uniformity, text segmentation for contextual understanding, data augmentation for dataset balance, and an ensemble learning method integrating models like ESG-BERT, RoBERTa, DeBERTa, and Flan-T5 for nuanced analysis. Yielding excellent results, our research showcases the potential of using language models to improve ESG-oriented decision-making, contributing valuable insights to the FinNLP community.

Keywords: ESG Impact Analysis, Financial NLP, Multilingual Data Pipeline

1. Introduction

The growing emphasis on Environmental, Social and Corporate Governance (ESG) within the financial sector underscores the necessity for better understanding and analysis of ESG-centric information. To address this need, the FinNLP community has been at the forefront of crafting natural language processing tasks on ESG-related news. Previous efforts encompassed taxonomy enrichment, semantic representation[1], ESG-issue identification and classification[2] in a variety of languages.

Building on previous work, ML-ESG-3[3] introduces a new task aimed to evaluate the potential impact duration of ESG events reported in news articles on corporations. This task challenges NLP models to pinpoint the impact timeline of ESG events on a company's performance and sustainability. Gaining this insight is vital for investment decisions, corporate strategies, and policy-making.

This paper delves into the intricacies of leveraging large language models to tackle the ML-ESG-3 challenge. By harnessing a blend of machine translation for multilingual coherence, data processing and augmentation for content uniformity, and a mix of language models like ESG-BERT, RoBERTa, DeBERTa, and Flan-T5, we aim to quantify the temporal effects of ESG-related news on companies across multiple languages. Our work not only enriches the domain's academic discourse but also offers practical insights for stakeholders in the financial sector.

2. Related Work

Despite the clear definition of Environmental, Social, and Governance principles following years of evolution, systematically identifying and assessing ESG-related news presents persistent challenges, drawing the attention of scholars aiming to address it. The annotation work by Kannan and Seki laid a comprehensive framework for categorizing ESG themes and assessing their sentiment through the analysis of Japanese corporate CSR (Corporate Social Responsibility) reports. Further advancing the field, the DynamicESG project, led by Tseng et al., compiled and analyzed an extensive collection of news articles over a twelve-year period, drawing upon MSCI ESG ratings and SASB standards to categorize news by impact type, level, and duration. The temporal dimension enables a more nuanced analysis of how news coverage could align with ESG criteria.

Domain-specific models have also played a significant role in enhancing the analysis of ESG information. Among these, FinBERT-ESG stands out as a specialized adaptation of the FinBERT model, which has been fine-tuned on 2,000 manually annotated sentences extracted from firms' ESG reports and annual reports (Huang et al., 2022). This allows FinBERT-ESG to efficiently tackle ESG classification tasks. Similarly, ESG-BERT is an environment-focused variant of BERT, initially trained through a Masked Language Model (MLM) task on Accounting for Sustainability corpus, and subsequently fine-tuned for sequence classification tasks (Mehra et al., 2022). These models have proven to be invaluable tools for both researchers and industry practitioners.

In light of these advancements, the ESG task series has been launched, starting with FinSim4-ESG at FinNLP-2022 to expand the taxonomy for semantic analysis of sustainability reports. In 2023, the ML-ESG-1 task focuses on the identification and classification of ESG-related news into 35 key is-

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

[2] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp2023/shared-task-esg-impact

[3] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-kdf-2024/shared-task-ml-esg-3

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 188–196
20 May, 2024.

sues as per MSCI ESG rating guidelines. ML-ESG-2 further explores the determination of whether the news signifies an opportunity or a risk from an ESG perspective.

3. Dataset & Task Setting

Multilingual ESG Impact Duration Inference (ML-ESG-3) is the latest task which seeks to evaluate the duration or length of impact that an event reported in a news article might have on a company. Specifically, the duration of the impact is classified into three groups: short-term (under 2 years), medium-term (2 to 5 years), and long-term (over 5 years).

The dataset includes 545 English articles from ESGToday[4], 352 Chinese articles from ESG Sustainable Taiwan[5], 661 French articles from Novethic[6], and 800 Korean articles from ESGEconomy[7], 2,358 news articles in total. Each article is provided with a title and the content of the news. As shown in Table 1, the distribution of labels in the training set is fairly even. However, the distribution of these labels under each language is not uniform and varies significantly across the different languages. Generally, news articles categorized with an impact duration of 2 to 5 years form the smallest group, which might require a more detailed examination.

Label	$<$ **2 yr**	**2 - 5 yr**	$>$ **5 yr**	**Total**
Chinese	87 (24.7%)	62 (17.6%)	203 (57.7%)	352
English	82 (15%)	198 (36.3%)	265 (48.6%)	545
French	131 (19.8%)	231 (34.9%)	299 (45.2%)	661
Korean	446 (55.8%)	142 (17.8%)	212 (26.5%)	800
Total	746 (31.6%)	633 (26.9%)	979 (41.5%)	2,358

Table 1: Validation Set Label Distribution

The test set provided later comprises datasets in English, French, and Korean, with a total of 482 news articles. The test set shows a varied distribution of impact duration in English from the training set, as shown in Table 2.

Label	$<$ **2 yr**	**2 - 5 yr**	$>$ **5 yr**	**Total**
English	6 (4.4%)	47 (34.6%)	83 (61%)	136
French	31 (21.2%)	32 (21.9%)	83 (56.8%)	146
Korean	96 (48%)	40 (20%)	64 (32%)	200
Total	133 (27.6%)	119 (24.7%)	230 (47.7%)	482

Table 2: Test Set Label Distribution

4. Methodology

This section describes our approach, covering data pre-processing, model selection, and the application of ensemble learning techniques. We detail the different strategies considered at each stage and provide the reasoning for our choice of methodologies.

4.1. Data Pre-processing

4.1.1. Translation

Initially, the bert-base-multilingual-cased model (Devlin et al., 2018) was used as a baseline for our multilingual dataset but achieved a low accuracy of 0.31. Pivoted away from further multilingual adaptations, we adopted machine translation to convert the dataset into English, following a strategy noted in previous research (Lee et al., 2023).

For the translation task, we experimented with the Facebook M2M model (Fan et al., 2020), Google Translation API [8], and DeepL API [9] to translate titles and content from Chinese, French and Korean into English, including a preliminary step of converting traditional Chinese to simplified Chinese to enhance API compatibility. To assess the translation quality and stability, the BLEU metric was introduced, aiming to compare the machine-generated translations with original content by evaluating the precision of n-gram matches (Papineni et al., 2002). This process involved selecting samples from the Chinese, Korean, and French datasets for translation to English, and then back-translating these English texts into the original languages. Table 3 displays the average BLEU scores for all translators under all language translation tasks.

Our sampling review found Google Translate occasionally repeated sentences, and Facebook's M2M underperformed significantly, especially with long articles. DeepL API, however, showed consistent quality without these issues, achieving the high-

[4] https://www.esgtoday.com/
[5] https://esg.businesstoday.com.tw/
[6] https://www.novethic.fr/
[7] https://www.esgeconomy.com/
[8] https://cloud.google.com/translate
[9] https://www.deepl.com/en/docs-api

Method	Chinese	French	Korean
DeepL API	0.48	0.82	0.62
Google API	0.15	0.56	0.37
FB M2M	0	0	0

Table 3: BLEU Score Across Translation Methods

est BLEU scores across all tested languages. In addition, it showed a strong capability in accurately translating both simplified and traditional Chinese characters. Therefore, DeepL API was selected for all non-English translations.

4.1.2. Article Segmentation

After exploratory data analysis of the translated dataset, we observed that English and French news articles were much shorter, averaging word counts of 73 and 96, respectively. These samples typically highlight the most pertinent sentences—often 2 to 3—from a given news piece, as selected by their annotators. In contrast, articles in Chinese and Korean exhibit substantially higher word counts, averaging 922 and 555 words respectively. Notably, the Chinese articles were extracted and cleaned from their original HTML by our team.

Since a more uniform distribution of content length is generally preferred for model training, it is necessary to employ segmentation techniques, which divide the longer articles into smaller, more manageable paragraphs. Specifically, each article was divided at intervals of every five sentences. Following this, each segment was fed into the FinBERT-ESG classification model, where it was assigned to one of four categories: Environmental, Social, Governance, or None, accompanied by a respective probability score. Paragraphs categorized as None with a probability of 0.9 or higher were excluded. This step not only prevents the datasets in Chinese and Korean from disproportionately influencing the overall training set, but also ensures the integrity and ESG-related quality of the new samples, balancing both the quantity and the quality of data across languages.

Following this segmentation and filtering process, our dataset expanded from 2,358 to 6,115 samples, while reducing the Chinese and Korean average word count to 163 and 119 respectively. A group shuffle split is applied to separate the training and validation sets, ensuring that samples with identical titles are not present in both sets to prevent data leakage. With a training-validation split of 0.2, our dataset was divided into 4,887 samples for training and 1,238 for validation.

4.1.3. Data Augmentation

The step of segmenting articles helps to standardize the length of the content but also introduces the problem of class imbalances, particularly noticeable in the category of impacts lasting 2 to 5 years, which constitutes only 22.3% of the training data. This imbalance makes it challenging for the model to accurately predict medium-term impacts, resulting in a prediction accuracy of less than 0.2 in our baseline.

To address this issue, we leveraged the widely recognized Reuters dataset[10] for news to augment the existing dataset. From the initial pool of 17,712 unique news articles, we adopted a similar methodology as the training set, using FinBERT-ESG for classification and segmentation, which narrowed down the dataset to 2,741 samples. Given the general nature of Reuters news, the selection process was much more stringent, filtering out news to only include those with an E/S/G label probability of 0.5 or higher.

Prior to annotating the Reuter dataset, a preliminary evaluation was conducted on chat-based models, including both commercially available models like GPT-4 (OpenAI, 2024) and Gemini-Pro (Team, 2023), and open-source alternatives such as GPT-NeoXT-Chat-Base-20B (Black et al., 2022) and Pythia-Chat-Base-7B-v0.16 (Biderman et al., 2023), as referenced by Lee et al. (2023). Our evaluation involved a random selection of 100 samples from the train set. GPT-4 and Gemini-Pro recorded accuracies of 50% and 48%, respectively, showing comparable outcomes albeit with noticeable differences in their label distribution. Specifically, GPT-4 categorizes 65% of its labels as more than five years of impact duration, whereas Gemini-Pro identified 53% of impact duration with a two to five-year range. Conversely, the other two models displayed inferior performance, yielding predictions that lacked generalizability. According to Table 4, GPT-NeoXT-Chat-Base-20B frequently predicted an impact duration of more than five years in 90% of cases, while Pythia-7B achieved a mere 29% accuracy rate.

Model	$<$ 2 yr	2 - 5yr	$>$ 5 yr	Acc
GPT-4	5	30	65	0.50
Gemini-Pro	10	53	37	0.48
NeoXT-20B	9	1	90	0.46
Pythia-7B	63	26	0	0.29

Table 4: Label Distribution and Prediction Accuracy across Language Models

From the preliminary study, GPT-4 and Gemini-pro were selected, with GPT-4 acting as the base

[10] https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

model and its classifications being cross-verified with those from Gemini-pro. Despite the significant biases and variances in label distribution across these two models, each offers distinct insights into the classification process. When both models agree on a label, the combined accuracy of their predictions can reach 56%. To further refine the selection of ESG-related news and reduce noise in the augmentation set, we also introduced a new classification, "Not have an ESG impact," into the prompt, detailed in the appendix. After GPT-4 filtered the unrelated news, any article receiving the same classification was included in the augmentation set.

The final augmented dataset thus comprised 1,221 samples, with 858 labeled as having an impact duration of 2 to 5 years. Following augmentation, the training set expanded to 6,098 samples, achieving a more balanced distribution across categories, as shown in Table 5.

Model	$<$ **2 yr**	**2 - 5 yr**	$>$ **5 yr**
Non-Augmented	1,681 (34.5%)	1,092 (22.4%)	2,104 (43.1%)
Augmented	1,867 (30.6%)	1,950 (32.0%)	2,281 (37.4%)

Table 5: Label Distribution after Augmentation

4.2. Model Selection

Considering the diverse topics and professional terminology found in ESG news, often presented in long articles, we selected ESG-BERT (Mehra et al., 2022), RoBERTa (Zhuang et al., 2021), DeBERTa (He et al., 2021), and Flan-T5 (et. al, 2022) as our pre-trained models to perform finetuning, for their advanced generalization capabilities, semantic understanding, and popularity in the prior ESG task series. Here is a brief overview of each model's strengths:

- ESG-BERT is highly effective in extracting and classifying information pertinent to sustainable investing and ESG themes. This effectiveness is largely due to its tailored training on ESG-specific text, enhancing its capacity for ESG task performance and semantic extraction, as highlighted by Lacoste et al. (2019).

- RoBERTa is an enhanced variant of BERT with a dynamic masking mechanism, 10x training corpus, and an improved training strategy, resulting in superior text comprehension and model generalization. Notably, a study conducted by Pontes et al. (2023) underscores RoBERTa's proficiency and accuracy in classifying news documents into specific ESG issue labels within an English-language dataset.

- DeBERTa is designed for processing lengthy articles, as its disentangled attention mechanism is key for analyzing long-distance sentence dependencies, essential for understanding context and handling complex sentence structures. In the realm of ESG, DeBERTa has demonstrated commendable efficacy and precision in identifying fraudulent ESG news, attributed to its advanced attention mechanism (Suryavardan et al., 2023).

- Flan-T5 is designed to generalize better to new tasks with minimal examples. Its few-shot learning capability allows it to understand and perform tasks that it might not have been explicitly trained for, using only a few examples to guide its predictions. This versatility makes Flan-T5 an excellent choice for ESG impact duration inference. Specifically, its encoder part is used for extracting semantic meanings, aiding in the inference.

During this stage, each model underwent finetuning on the training dataset, which was either the original or augmented version, through adjustments to its structure, such as selecting layers to unfreeze and incorporating extra layers before the softmax layer, as well as tweaking hyperparameters including batch size, dropout rate, and learning rate. The goal was to determine the optimal settings for each model under consideration.

4.3. Ensemble Learning

Ensemble learning can effectively reduce overfitting by averaging out biases and variances across diverse models (Opitz and Maclin, 1999) and correcting errors of weak learners, further improving model performance (Schapire, 1990). Several experiments were conducted to integrate classification results of the four distinct models by taking their softmax layers of probabilities as the inputs.

Prior to applying ensemble learning, to manage scenarios where multiple segments from a single news article yield divergent predictions, we calculated the mean softmax value for segments sharing the same Group ID. For example, Figure 1 demonstrates how a news article is partitioned into three segments, with each outputting a softmax layer configured as of 1 x 3 (corresponding to the probabilities of the 3 classes), and these segments are collectively averaged.

Figure 2 displays the entire model architecture for ensemble learning. After aggregating softmax results for each article, we applied averaging for each class across all four finetuned models as one of the

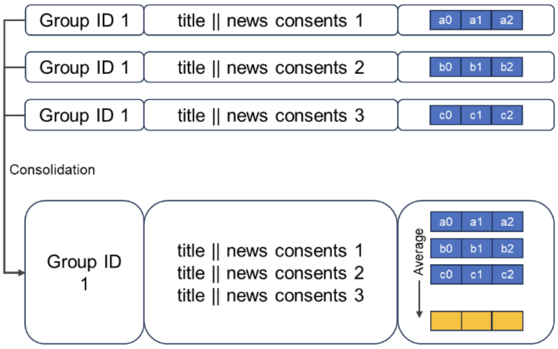

Figure 1: Consolidated prediction via averaging softmax outputs with same Group ID.

ensemble approaches. For the rest of the experiments, correlation analysis was conducted first to eliminate pairs exhibiting a correlation higher than 0.7. The remaining probabilities were input into various classifiers, including K-Nearest Neighbor, Decision Tree, Random Forest, and Multiple Linear Perceptrons, to produce the predicted labels.

5. Experiment Results

The experimental process is divided into three key stages: 1) evaluating the performance of the augmentation set, 2) tuning the hyperparameters for all models, and 3) identifying the most effective ensemble techniques to determine the optimal model. Results for each stage are presented and analyzed in this section.

5.1. Augmentation Effectiveness

During the first phase of the experiments, a learning rate of 1e-5, a batch size of 32, and the strategy of unfreezing the last three layers for each model were adopted. For the experiments that utilized non-augmented data, resampling techniques were employed to address issues of data imbalance.

As shown in Table 6 the comparison between models using non-augmented and augmented data reveals varying degrees of performance improvement. ESG-BERT and DeBERTa-base saw improvements in both Micro and Macro F1 score; RoBERTa-base experienced mixed results as Macro F1 dropped slightly.

The analysis suggests that while augmentation can lead to a more balanced dataset and potentially better model performance, the effectiveness of these techniques can differ between models. We use Macro F1 as the criteria for determining whether the augmented dataset is used for model training in subsequent training phases.

5.2. Hyperparameter Tuning Results

In the second phase of the experiments, we focused on hyperparameter tuning for each model, mainly on the number of un-freezed layers, batch size and learning rate. Table 7 illustrates the parameters we searched, best parameters for each model, and their Micro F1 and Macro F1 scores. Among all the models, RoBERTa-large, ESG-BERT, DeBERTa-base, and Flan-T5-large emerged as the top performers, making them prime candidates for integration into ensemble models due to their relatively high prediction accuracy.

5.3. Ensemble Learning Results

In the previous step, RoBERTa-base recorded a Micro F1 score of 0.5998 and a Macro F1 score of 0.4899, setting a benchmark for the ensemble learning phase. Figure 3 illustrates the Macro F1 scores for various ensemble learning methods and their performance across different languages. Techniques such as MLP, K-Nearest Neighbors, and Averaging all surpassed the RoBERTa-base baseline in terms of score.

5.4. Test Results

Finally, we presented the test outcomes for four individual models and five ensemble models, including the three submitted in the ML-ESG-3 task, as displayed in Table 8. The Flan-T5 model exhibited superior results on the English dataset, while the Averaging ensemble model outperformed others on the French dataset, and Random Forest emerged as the top-performing model for the Korean dataset.

6. Discussions

In this section, we dive into prediction results and discuss potential future work to improve model robustness. This entails analyzing the effect on word count, augmentation set and validation set.

6.1. Effect of Word Count

We observed a notable trade-off in model performance between English/French datasets and Korean datasets, likely due to the large disparity in article word count and label distribution. To understand the impact of word count, we performed a logistic regression analysis, revealing a significant positive correlation between word count and model performance, significant at a 10% level (p-value: 0.074).

Given the analysis, two principal methodologies are applicable to achieve this aim in the future: extending the text length per input and adopting ad-

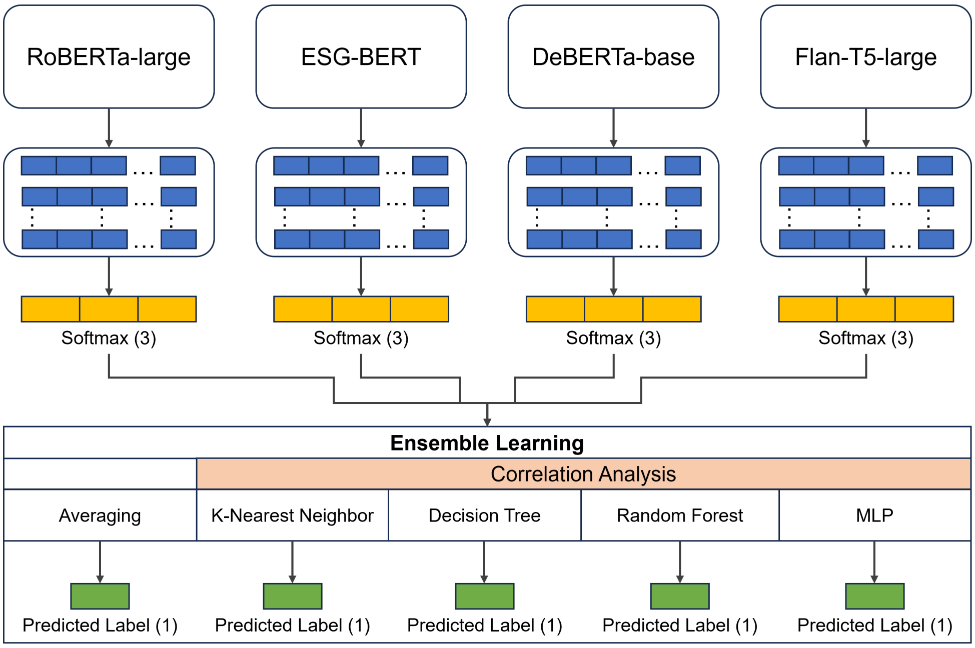

Figure 2: The Complete Ensemble Model Architecture

Model	Non-augmented		Augmented	
	Micro F1	Macro F1	Micro F1	Macro F1
ESG-BERT	0.5521	0.4823	0.5645	0.5050
RoBERTa-base	0.5721	0.5147	0.5807	0.4869
DeBERTa-base	0.5654	0.4823	0.5922	0.4881

Table 6: Augmentation Set Performance across Models

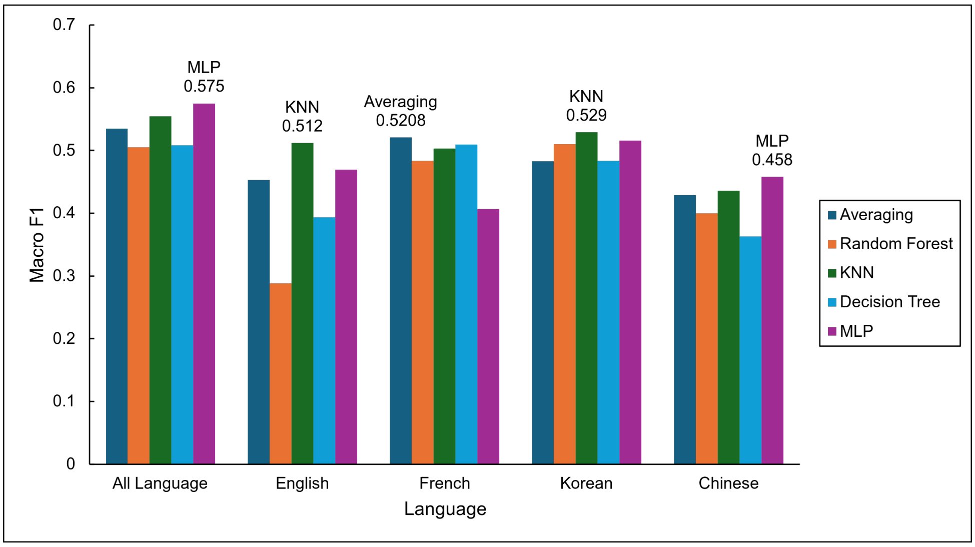

Figure 3: Ensemble Learning Macro F1 Score across Languages

vanced segmentation techniques beyond simple segmentation.

Regarding input length, directly extracting English and French news from original URLs could be beneficial, mirroring the success seen with Korean articles, which typically contains more information. In addition, adjusting segmentation sizes could be advantageous, especially considering the current

Model	Parameter Search Value	Best Parameter	Micro F1	Macro F1

ESG-BERT	num_layers = 1, 3, 6
	batch_size = [16, 32, 64]
	lr = [1e-4, 5e-4, 1e-5]
		num_layers = 3
		batch_size = 32
		lr = 1e-5
			0.5645	0.5050

RoBERTa-base	num_layers = 1, 3, 6
	batch_size = [16, 32]
	lr = [1e-5, 1.5e-5, 5e–5]
		num_layers = 6
		batch_size = 32
		lr = 1.5e-5
			0.5998	0.4899

RoBERTa-large	num_layers = 1, 3, 6
	batch_size = [16, 32]
	lr = [1e-5, 1.5e-5, 5e–5]
		num_layers = 6
		batch_size = 16
		lr = 1.5e-5
			0.5874	0.5057

DeBERTa-v3-base	num_layers = 1, 3, 6
	batch_size = [16, 32]
	lr = [1e-5, 5e-5]
		num_layers = 3
		batch_size = 16
		lr = 1.5e-5
			0.5702	0.4596

DeBERTa-v3-large	num_layers = 1, 3, 6
	batch_size = [16, 32]
	lr = [1e-5, 5e-5]
		num_layers = 3
		batch_size = 16
		lr = 1.5e-5
			0.5683	0.4549

Flan-T5-base	num_layers = 1, 3
	batch_size = [16, 32]
	lr = [1e-5, 2e-5, 5e-5]
		num_layers = 3
		lr = 5e-05
			0.5711	0.4050

Flan-T5-large	num_layers = 1, 3
	batch_size = [16, 32]
	lr = [1e-5, 2e-5, 5e-5]
		num_layers = 3
		lr = 2e-5
			0.6050	0.4293

Table 7: Hyperparameters Tuning Results on Validation Set

	English		French		Korean	
Model	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
ESG-BERT	0.5735	0.4120	0.4315	0.3844	0.6450	0.5148
RoBERTa	0.4559	0.3705	0.4521	0.4357	0.6500	0.5637
DeBERTa	0.5000	0.4063	0.5000	0.4654	0.6150	0.4630
Flan-T5-Large*	**0.6912**	**0.4689**	0.5753	0.4335	0.6450	0.4730
MLP (ES)*	0.6176	0.4035	0.4863	0.4488	0.6550	0.5126
Averaging (ES)*	0.5882	0.3983	**0.5616**	**0.4622**	0.6500	0.4732
KNN (ES)	0.4706	0.3820	0.4041	0.3965	0.6450	0.5881
RF (ES)	0.4632	0.3789	0.4247	0.4206	**0.6600**	**0.6214**
DT (ES)	0.4412	0.3546	0.3562	0.3312	0.6100	0.5378

Table 8: Performance Metrics by Language and Model on Testset. Model marked with asterisk(*) were submitted to the committee.

average segmentation size is about 100 tokens, while the capacity of all models extends to 512 tokens.

In terms of segmentation techniques, the implementation of sliding window segmentation could improve contextual flow and semantic continuity. Cross-segment attention integrates full-article context, potentially improving the model's ability to understand long-distance dependencies and intricate relationships (Lukasik et al., 2020). Hierarchical BERT serves to bridge local and global contexts within an article, amalgamating both detailed and overarching semantic information (Lu et al., 2021).

6.2. Effect of Augmentation Set

Despite our effort in handling imbalanced dataset and low prediction accuracy in impact duration between 2 and 5 years, our model still struggles with the out-of-sample distribution issue in the test set, indicating a potential over-fitting to the training data. A potential enhancement in our process could involve adopting Gemini-pro as the primary model for labeling augmentation dataset, given its superior F1 score of 0.5144 in the medium-term duration inference, in contrast to GPT4's F1 score of 0.4179.

In addition, the inclusion of ESG-related news rather generic business news for the augmentation set would likely boost model performance. For instance, utilizing the Global Database of Events, Language, and Tone (GDELT) project enabled Aue et al. to collect 8,000 ESG-related ratings derived from 3 million articles pertaining to 3,000 US corporations throughout the period of 2018 to 2020. We could use this database rather than the filtered Reuters dataset for data augmentation.

7. Conclusion

This study undertakes the ML-ESG-3 shared task, with the goal of predicting the ESG impacts duration across datasets in English, French, Korean, and Chinese. We finetuned BERT-based and T5-based classifiers in conjunction with techniques such as machine translation, text segmentation, data augmentation, and ensemble learning. Our findings indicate the performance enhancement from data augmentation and strategic segmentation while mitigate issues of class imbalance. Through experimentation, we identified the optimal model configurations that significantly enhanced our predictions' precision and reliability. Our research contributes valuable insights and methodologies to the FinNLP community, providing a robust framework for assessing the temporal effects of ESG-related news on corporations. Future directions include enhancing our approach by extending text inputs length, employing advanced segmentation techniques for better contextual understanding, and concentrating on ESG-related news to enrich our data augmentation process.

8. Availability

The code is available at https://github.com/roxy rong/ml-esg-3.

9. Bibliographical References & Language Resources

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Hyung Won Chung et. al. 2022. Scaling instruction-finetuned language models.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR 2021*.

Allen H. Huang, Hui Wang, and Yi Yang. 2022. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for esg scores. In *Proceedings of The 5th Workshop on Financial Technology and Natural Language Processing (FinNLP)*.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. EaSyGuide: ESG Issue Identification Framework leveraging Abilities of Generative Large Language Models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 127–132, Macao.

Jinghui Lu, Maeve Henchion, Ivan Bacher, and Brian Mac Namee. 2021. A sentence-level hierarchical bert model for document classification with limited labelled data.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Sim oes. 2020. Text segmentation by cross segment attention. In *Proceedings of* *the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. Association for Computational Linguistics.

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*, EMSA 2022. Academy and Industry Research Collaboration Center (AIRCC).

OpenAI. 2024. Gpt-4 technical report.

D. Opitz and R. Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

3DS Outscale. 2024. Ml-esg 2024 for social good (esg) - 3rd edition guidelines. Guidelines by 3DS Outscale.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging bert language models for multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 121–126, Macao.

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*, 21:140:1–140:67.

R.E. Schapire. 1990. The strength of weak learnability. *Machine Learning*, 5(2):197–227.

S. Suryavardan, S. Mishra, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A.N. Reganti, A. Das, A.P. Sheth, M.K. Chinnakotla, A. Ekbal, and S. Kumar. 2023. Findings of factify 2: Multimodal fake news detection. *ArXiv*, abs/2307.10475.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

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 and Knowledge Management (CIKM'23)*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

10. Appendix

10.1. Template for Chat-based Language Model for Labeling

Below is our one shot learning template for labeling augmented using various chat-based language models.

TEMPLATE =
"""
Label the ESG impact duration for the following news:
Options:

- 0 - below 2 years
- 1 - between 2 and 5 years
- 2 - more than 5 years

You should only output the number and have no explanations.

Example: The ways to practice self-care with a fitness watch are almost limitless, but here are six easy-to-implement tips to start today | Dubai-based airline Emirates announced plans to conduct its first experimental flight using 100% sustainable aviation fuel (SAF) in one engine this week, in a test aimed at supporting expanded use of SAF for commercial flights.
Output: 0

News: {$news}
Output:
"""