Multilingual ESG News Impact Identification using an Augmented Ensemble Approach

Harika Abburi[1], Ajay Kumar[2], Edward Bowen[2], Balaji Veeramani[2]

[1]Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited India
[2]Deloitte & Touche LLP, USA

{abharika, ajaykumar35, edbowen, bveeramani}@deloitte.com

Abstract

Determining the duration and length of a news event's impact on a company's performance remains elusive for financial analysts. The complexity arises from the fact that the effects of these news articles are influenced by various extraneous factors and can change over time. As a result, in this work, we investigate our ability to predict 1) the duration (length) of a news event's impact, and 2) level of impact on companies. The datasets used in this study are provided as part of the Multi-Lingual ESG Impact Duration Inference (ML-ESG-3) shared task. To handle the data scarcity, we explored data augmentation techniques to augment our training data. To address each of the research objectives stated above, we employ an ensemble approach combining transformer model, a variant of Convolutional Neural Networks (CNNs), specifically the KimCNN model and contextual embeddings. The model's performance is assessed across a multilingual dataset encompassing English, French, Japanese, and Korean news articles. For the first task of determining impact duration, our model ranked in first, fifth, seventh, and eight place for Japanese, French, Korean and English texts respectively (with respective macro F1 scores of 0.256, 0.458, 0.552, 0.441). For the second task of assessing impact level, our model ranked in sixth, and eight place for French and English texts, respectively (with respective macro F1 scores of 0.488 and 0.550).

Keywords: impact, data augmentation, transformers, CNN

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). This rising awareness, and importance of the analyzing large volumes of ESG related documents has necessitated the use of language technologies in this area.

The rapid advancements in deep learning, and Natural Language Processing (NLP) technologies have enabled the research in development of systems designed to extract relevant information from ESG reports. Language models have been used for various financial tasks such as sentiment analysis, named-entity recognition, and document classification (Araci, 2019; Wu et al., 2023). However, their application to ESG-specific tasks remains relatively limited. Existing works have begun to explore this area, demonstrating the potential of language models for ESG analysis. (Raman et al., 2020) evaluate the impact of language model embeddings on the classification of sentences concerning their relevance to the ESG domain. Similarly, (Mehra et al., 2022) pre-train a Bidirectional Encoder Representations from Transformers (BERT) model on ESG-related text to show improvement on classification tasks. Furthermore, (Wang et al., 2023) explore the potential of combining contrastive learning with BERT language model for the task of identifying environmental, social, and governance issues in news articles.

Despite these promising initial explorations, the scarcity of publicly available ESG data, particularly for low-resource languages, remains a significant challenge that hinders further advancements in this field. To address this issue, various data augmentation techniques have been explored to expand and enrich the training data, including Easy Data Augmentation, translation, zero-shot classification, contextual augmentation (Lee et al., 2023; Kobayashi, 2018). (Nugent et al., 2021) leverage back-translation technique to generate additional training data to perform ESG document classification. The generated data is then used to fine-tune the BERT model to further enhance its performance. Furthermore (Glenn et al., 2023), generated synthetic data with LLMs in zero-shot and few-shot settings effectively bridging the gaps in data availability for low-resource languages.

These efforts have paved the way for the development of advanced multilingual solutions. (Mashkin and Chersoni, 2023) highlights the usefulness of using Transformer-based representations and cross-lingual models for multilingual ESG analysis. (Jør-

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 197–202

gensen et al., 2023) extend the concept of pre-training on financial text to multilingual data in seven languages. Additionally, (Pontes et al., 2023) investigates the use of BERT and its variants for classifying news articles into different ESG categories. They also explores the effectiveness of these models in multiple languages, offering insights into the potential of this approach for expanding the scope of ESG issue identification.

Motivated by these developments, our team participated in the Financial Technology and Natural Language Processing, the Knowledge Discovery from Unstructured Data in Financial Services, and Economics and Natural Language Processing (FinNLP-KDF-ECONLP-2024) shared task on ML-ESG-3 (Chen et al., 2024). This task aimed to predict the duration and level of a news article's impact on a company. Towards this task, we adopted data augmentation techniques such as translation, paraphrasing, and Generative Pre-training Transformer (GPT) mix to augment the training data. Furthermore, we trained an ensemble model combining transformers, KimCNN architecture (Kim, 2014), and Voyage AI embeddings[1] and assessed their performance across various languages and subtasks. Our model achieved top rankings (ranging from 1st to 8th) across different subtasks, demonstrating the effectiveness of our approach in furthering the capabilities of NLP for identifying ESG impact level and duration.

2. Dataset

This section describes the dataset used for exploring the ML-ESG-3 shared task (Chen et al., 2024). The dataset consists of ESG new articles annotated with one or more annotations to each news article. The data was provided by the task organizers and the task is slightly different across the language subsets.

- English and French: This dataset includes two annotations: "Impact Level" and "Impact Length". Impact Level qualifies the opportunity or risk as being of "low", "medium" or "high." Impact Length annotations of "Less than 2 years", "2 to 5 years", and "More than 5 years".
- Japanese: For this language, only the annotations of 'Impact Length' are provided which are similar to the English and French datasets.
- Korean: In this dataset, there are two annotations: 'Impact Length' and 'Impact Type'. Impact Length annotations are same as the English and French datasets where as 'Impact Type' is categorized as 'opportunity,' 'risk,' or 'cannot distinguish' (Tseng et al., 2023). In Korean language, we participated only in Impact length.

Table 1 shows the data statistics for the languages and subtasks. More detailed information about the dataset can be found in the shared task overview paper (Chen et al., 2024).

Language	Impact Length		Impact Level	
	$Train$	$Test$	$Train$	$Test$
English	545	136	545	136
French	661	146	661	146
Japanese	52	1500	—	—
Korean	800	200	—	—

Table 1: Dataset statistics (number of samples) of Original data

Data Type	Impact Length	Impact Level
Original	545	545
Original + tr	1835	1054
Original + tr +pp	3670	2108
Original + tr+ pp + GPT-mix	6670	5108

Table 2: Original and augmented training data statistics (number of samples) for English subtasks. Augmentations were performed using translation (tr), paraphrasing (pp) and GPT-mix.

2.1. Synthetic Data Generation

Due to the limited data available in each language task, we employed various data augmentation techniques to enrich the training set: we used translation (tr), paraphrasing (pp), and GPT-mix.

Translation: To augment the training data, we translated the French, Japanese, and Korean datasets into English using the widely recognized DeepL[2] translation service. We used English data as it is and converted the other languages to English.

Paraphrase: After translation, we employed the Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) transformer model (Marceau et al., 2022) for paraphrasing the text. While this model was originally designed for abstractive summarization, its ability to leverage large amounts of text and understand semantic relationships between words makes it suitable for paraphrasing tasks.

GPT-mix: We further augmented the data using

[1] https://www.voyageai.com/

[2] https://www.deepl.com/translator

GPT-mix (Yoo et al., 2021), a technique that leverages large language models to generate summary of text samples. GPT-mix effectively captures human language nuances by blending two real samples. We selected two samples with identical labels from the original dataset and used GPT-mix to generate summary of these pairs, creating new data points. This process yielded 3,000 additional samples.

Table 2 shows the number of samples augmented for the English subtasks using the aforementioned augmentation techniques. We then translated this augmented data into other languages using DeepL translator, standardizing the number of training samples across language subtasks.

Our use of translation of data across languages, and the use of transformation (pp) and mixing samples for augmentations helps create better datasets for data sparse tasks.

3. Proposed approach

In this section, we describe our approach for detecting the duration and length of a news event's impact on a company's performance. Our text classification architecture builds upon a modified KimCNN framework (Kim, 2014), with carefully incorporated transformer-based representations and Voyage AI embeddings. The proposed framework comprises five specific layers: embedding layer, CNN layer, pooling layer, enriched representation layer and output layer. The detailed description of each layer is given as follows.

Embedding layer: Our approach begins with a pre-trained transformer model, which has been trained on a massive corpus of text data. This allows our model to capture rich contextual information about the meaning of words and their relationships within the text. Instead of using the standard output of the transformer model, we specifically focus on the final four hidden layers of the transformer model as these layers effectively captures the relevant information from input data.

Convolutional layer: Following the embedding layer, the extracted representations then undergo a series of convolutional operations. We build upon the KimCNN architecture, which is known for its effectiveness in text classification tasks. This architecture utilizes multiple convolutional layers with varying filter sizes (specifically 3, 4, and 5) with padding enabling the model to learn patterns from different n-gram combinations within the text. This allows the model to focus on different n-gram lengths, potentially capturing both short and long-range dependencies that contribute to the overall meaning. To improve efficiency, we use depthwise separable convolutions. After each convolutional layer, we apply a Rectified Linear activation Unit (ReLU) activation function to introduce non-linearity. This allows the model to learn more complex relationships between features. Additionally, we use dropout as a regularization technique to prevent overfitting to the training data.

Pooling layer: Following the convolutional layer, max-over-time pooling is applied to each convolutional layer's output. This operation extracts the prominent feature from each sequence captured by the convolution, focusing on the relevant information within each n-gram length. The features from the convolutional layers are then concatenated into a single representation, effectively combining the information learned from different n-gram lengths.

Enriched representation layer: To further enhance the model's understanding, the single representation is again concatenated with Voyage AI embeddings. These state-of-the-art pre-trained text embedding models capture semantic meaning from text data, effectively injecting external knowledge into the model.

Output layer: The final concatenated representation is then fed into a fully connected layer. This layer performs the final classification task, assigning probabilities to each possible class the text belongs to, enabling the model to predict the impact duration of the news event.

3.1. Different transformer based models

We explored various state-of-the-art large language models (Kalyan et al., 2021) to extract the features from the embedding layer. These include prominent models like such as; Bidirectional Encoder Representations from Transformers (BERT),Robustly optimized BERT approach (RoBERTa), and its cross-lingual language model RoBERTa (XLM-RoBERTa) along with their variants. However, we recognized that a single set of models might not perform equally well across diverse datasets and languages within the task. Therefore, we fine-tuned different LLM variants for each subtask and language and pick the top performing models based on heldout data. Table 3 lists the different LLMs that we explored for embedding layer of each subtask: English impact length (English-len), English impact level (English-lev),French impact length (French-len), French im-

Subtask	Transformer model
English-len	EnvRoBERTa-base[3] (Schimanski et al., 2023)
English-lev	ESG-BERT[4]
French-len	xlm-roberta[5] (Reimers and Gurevych, 2019)
French-lev	bert-base-multilingual[6]
Japanese-len	xlm-roberta (Reimers and Gurevych, 2019)
Korean-len	multilingual-mpnet-base[7]

Table 3: Transformer models used for different subtasks

Data Type	Impact Length				Impact Level			
	Acc	F_{macro}	$Prec$	Rec	Acc	F_{macro}	$Prec$	Rec
Original	0.476	0.345	0.317	0.378	0.524	0.462	0.546	0.468
Original + tr	0.451	0.376	0.405	0.383	0.500	0.438	0.466	0.445
Original + tr +pp	0.451	0.386	0.390	0.388	0.561	0.525	0.578	0.521
Original + tr+pp + GPT-mix	**0.524**	**0.477**	**0.478**	**0.476**	**0.585**	**0.530**	**0.583**	**0.533**

Table 4: Results on held out English dataset using various data augmentations

pact level (French-lev), Japanese impact length (Japanese-len), Korean impact length (Korean-len).

4. Experiments

This section provides the experimental evaluation of our proposed methods. For each of the tasks we report accuracy (Acc), macro F1 score (F_{macro}), precision ($Prec$) and recall (Rec).

4.1. Implementation details

We set aside 15% from the training data for performance evaluation. For the testing phase, the held-out set is merged with the training set. The experiments were conducted using the same hyperparameters: batchsize of 64, learning rate of 2e-5, epoch of 10, and optimizer of AdamW. The experiments were run on two A100 GPUs.

4.2. Results

For each task and language, we submitted three runs to the leaderboard (team name *Drocks*). These runs correspond to the different approaches on the heldout data. In this paper, we show only the top run results for each of the tasks. The full leaderboard is available at *https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-kdf-2024/shared-task-ml-esg-3*.

4.3. Effect of data augmentation techniques

To evaluate the effectiveness of data augmentation, we conducted experiments with incremental addition of data through translation, paraphrasing, and GPT-mix on the English dataset using the ESG-BERT model. Table 4 shows the performance metrics for models trained on augmented data, as detailed in Table 2. Notably, the results demonstrate that combining data augmentation with translation, paraphrasing and GPT-mix techniques improves the model's performance on both English subtasks with good margin of F_{macro} score. Therefore, for subsequent experiments, we utilize the "Original+tr+pp+GPT-mix" training data for training the model, and report the results on the held-out data for various sub-tasks.

Subtask	Acc	F_{macro}	$Prec$	Rec
English-len	0.596	0.441	0.439	0.451
English-lev	0.574	0.550	0.583	0.535
French-len	0.500	0.458	0.469	0.470
French-lev	0.486	0.488	0.508	0.483
Japanese-len	0.363	0.256	0.220	0.370
Korean-len	0.625	0.552	0.580	0.549

Table 5: Results on the subtasks on testing data

4.4. Results of proposed approach

Table 5 presents the performance of our proposed architecture across the subtasks using the augmented datasets. On English subtasks, the model achieved F_{macro} scores of 0.441 and 0.550 for the English-len and English-lev subtasks, respectively.

[3]https://huggingface.co/ESGBERT/EnvRoBERTa-base

[4]https://huggingface.co/nbroad/ESG-BERT

[5]https://huggingface.co/sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens

[6]https://huggingface.co/Tiamz/bert-base-multilingual-uncased-finetuned-news

[7]https://huggingface.co/sosoai/multilingual-mpnet-base-v2-embedding-all-safetensor

Utilizing the augmented French data, the model achieved F_{macro} scores of 0.458 and 0.488 for the French-len and French-lev subtasks respectively. Similar to English, the slightly lower score for French-len as compared to French-lev suggests identifying impact length might be more challenging as compared to impact level. For the Japanese subtask, the model only achieved an F_{macro} score of 0.256. While this score is lower than other subtasks, it secured the first rank in the competition, highlighting the potential of the approach in this specific task. However, the model achieved a higher F_{macro} score of 0.552 on the Korean subtask.These variations showcase the complexities of applying the model to diverse languages with varying volumes of data, potentially pointing towards areas of future investigation that will be of interest.

5. Conclusion

In this paper, we described our submission to the FinNLP-KDF shared task which consists of multiple sub tasks in determining the duration and level of the impact an event in the news article might have on the company. Our experiments demonstrated that data augmentation techniques effectively improved model performance. Furthermore, the proposed approach ranked within the top 10 for several languages (English, French, Korean) and securing first place based on F_{macro} score for the Japanese language subtask. These findings highlight the potential of the approach for multilingual ESG impact duration inference. However, the variations in performance across languages and subtasks underscore the inherent challenge of this domain. Future work will focus on further enhancing and adapting the model to address these complexities and improve performance across languages and tasks.

6. Bibliographical References

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

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

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*.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the finnlp-2023: Using synthetic data and transfer learning for multilingual esg issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP.

Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. 2021. mDAPT: Multilingual domain adaptive pretraining in a single model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3404–3418, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

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.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457.

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*.

Louis Marceau, Raouf Belbahar, Marc Queudot, Nada Naji, Eric Charton, and Marie-Jean Meurs. 2022. Quick starting dialog systems with paraphrase generation. *arXiv preprint arXiv:2204.02546*.

Ivan Mashkin and Emmanuele Chersoni. 2023. Hkesg at the ml-esg task: Exploring transformer representations for multilingual 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 140–145.

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 *CS & IT Conference Proceedings*, volume 12. CS & IT Conference Proceedings.

Tim Nugent, Nicole Stelea, and Jochen L Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings 14*, pages 157–169. Springer.

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.

Natraj Raman, Grace Bang, and Armineh Nourbakhsh. 2020. Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction*, 2(4):453–468.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Bridiging the Gap in ESG Measurement: Using NLP to Quantify Environmental, Social, and Governance Communication. *Available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract$_id = 4622514$*.

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.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335.

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*, pages 5412–5416.

Weiwei Wang, Wenyang Wei, Qingyuan Song, and Yansong Wang. 2023. Leveraging contrastive learning with bert for 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 116–120.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.