IMNTPU at ML-ESG-3: Transformer Language Models for Multi-Lingual ESG Impact Type and Duration Classification

Yu-Han Kao[a], Vidhya Nataraj[b], Ting-Chi Wang[c], Yu-Chun Cheng[c], Hsiao-Chuan Liu[a], Wen-Hsuan Liao[a], Chia-Tung Tsai[a], Min-Yuh Day[a]*

Graduate Institute of Information Management, National Taipei University[a],
Smart Healthcare Management, National Taipei University[b],
Department of Accountancy, National Taipei University[c]
Taiwan
{s711236116, s411177056, s411177047, s711136108, s711136109, s711136106, myday}@gm.ntpu.edu.tw,
vidhyanataraj99@gmail.com

Abstract

Team IMNTPU participated in the multi-lingual Environmental, Social, and Governance (ESG) classification task, focusing on datasets in three languages: English, French, and Japanese. This study leverages Pre-trained Language Models (PLMs), with a particular emphasis on the Bidirectional Encoder Representations from Transformers (BERT) framework, to analyze sentence and document structures across these varied linguistic datasets. The team's experimentation with diverse PLM-based network designs facilitated a nuanced comparative analysis within this multi-lingual context. For each language-specific dataset, different BERT-based transformer models were trained and evaluated. Notably, in the experimental results, the RoBERTa-Base model emerged as the most effective in official evaluation, particularly in the English dataset, achieving a micro-F1 score of 58.82 %, thereby demonstrating superior performance in classifying ESG impact levels. The major contribution of this paper highlights the adaptability and effectiveness of PLMs in tackling the complexities of multi-lingual ESG classification tasks. The practitioner implications of this paper provide ESG analysts with more reliable tools for assessing the impact duration and level of sustainability initiatives.

Keywords: Multi-Lingual ESG, Data Augmentation, ESG impact analysis, Classification, Pre-trained Language Models (PLMs)

1. Introduction

In recent years, the global investment and corporate governance community has increasingly recognized the pivotal role of ESG factors as essential perspectives for driving a company's long-term growth and informing investment decision-making. Evaluating the sustainability and ethical impact of investment opportunities, ESG considerations have underscored the necessity for robust tools and methodologies to address related issues. Meanwhile, the escalating risk associated with non-financial factors highlights ESG elements as a primary threat to the stability of financial systems (Ziolo et al., 2019). As a response to this imperative, initiatives have emerged to tackle the challenges of automatically identifying and categorizing ESG-related themes in textual data.

Responding to the imperative of incorporating ESG considerations, initiatives leveraging Natural Language Processing (NLP) technologies have been developed to automate the identification and categorization of ESG-related themes in textual data and revolutionizing the approach within the financial services sector. NLP serves as a potent instrument for extracting profound semantic insights from vast pools of unstructured data, ranging from financial reports to chat transcripts and news articles. Through such analysis, NLP has the potential to bolster scenario recognition and risk assessment across various financial contexts. Given the prevalence of individual opinions on financial matters, conveyed through diverse channels such as news outlets and social media platforms, strategic analysis of these sentiments offers invaluable insights, shaping decision-making processes and influencing both user and organizational perspectives within the financial domain.

In the progression from ML-ESG-2 to ML-ESG-3, the domain of ESG analysis has seen the introduction of sophisticated tasks aimed at enhancing the precision of ESG rating systems. In ML-ESG-2, a novel challenge was introduced, focused on ESG impact type identification, requiring models to discern whether a piece of news represents an opportunity or risk from an ESG perspective. Advancing further, ML-ESG-3 expanded the scope to include the classification of news articles based on impact duration and impact level, utilizing a multilingual dataset to reflect the global nature of ESG considerations. In this context, our team, IMNTPU, has employed in ML-ESG-3 utilizing the PLMs to adeptly classify sentences that describe a company's ESG efforts, assigning them to distinct labels for both impact duration and impact level, thereby showcasing the evolving complexity and understanding required in contemporary ESG analysis. Building upon the foundation laid by the tasks of ML-ESG-3, and the employment of advanced pre-trained language models by team IMNTPU for precise classification, this methodology facilitates the extraction of textual evidence for ESG impact duration and impact level from the often-noisy environment of news article reports. Consequently, this approach supports more informed investment decisions by leveraging the refined insights gained from the automated analysis

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 228–233

of ESG-related textual data.The contributions of this work can be summarized as follows:

- **Implementing Data Augmentation**: To combat class imbalance within the datasets, enhancing model robustness and ensuring a balanced representation for more accurate ESG impact duration and level classification.
- **Training with BERT-based Transformer Models**: Leveraging the sophisticated capabilities of BERT-based models across multilingual datasets to significantly improve the precision and comprehensiveness of ESG impact duration and level classifications.

Our research revealed a good correspondence in classifying the ESG impact duration and level in textual evidence. This finding will be helpful in future work on automatic estimation of ESG scores from textual resources.

The remaining part of the paper proceeds as follows: The second chapter introduces the related work related to the ML-ESG-3 shared task. Chapter three presents our approaches for each of the datasets. Chapter four provides a comprehensive account of the official experiment results and includes a detailed analysis. Finally, chapter five outlines the conclusions obtained from this study.

2. Related Work

In light of the heightened attention toward ESG issues, machine learning (ML) and NLP techniques have increasingly been leveraged in recent years to conduct sophisticated analyses of ESG ratings and predict impacts. By harnessing the predictive capabilities of Artificial Intelligence (AI) models have been created not only to assess current ESG ratings, classify them into various categories, but also to forecast future trajectories pertaining to both financial and societal impacts.(Tseng et al., 2023; Wang et al., 2023)

2.1 ESG in NLP

Lee et al. (2022)highlight the growing trend of companies disclosing their sustainability practices through various forms of unstructured text, such as reports and transcripts. They point out that NLP plays a crucial role in automating the classification and measurement of ESG-related news articles, enabling the parsing of extensive datasets to identify pertinent information efficiently. Furthermore, Zhuang et al. (2020) underscore the significance of transfer learning techniques within NLP, utilizing large language models to facilitate the transfer of knowledge across different sustainability domains and languages. These advancements underscore the pivotal role of NLP in enhancing the accessibility and analysis of sustainability information, contributing significantly to the field of ESG research.

2.2 Previous approach in multi-lingual ESG issues classification

In the realm of multi-lingual classification, the identification ESG issues across varied disclosure mediums presents a complex challenge. Recent efforts have explored numerous solutions, predominantly harnessing advanced NLP techniques to navigate this multifaceted landscape. A significant milestone in this ongoing journey was the 5th Workshop on Financial Technology and NLP (Kannan & Seki, 2023)which organized a shared task dedicated to ESG issue detection, attracting participation from 26 teams. Within this competitive context, a diversity of innovative approaches emerged, targeting a dataset encompassing 44 distinct ESG issues.

Armbhurst, Schäfer, and Klinger (2020) analyzed the impact of a company's environmental performance, derived from MD&A sections in financial filings, on its financial outcomes. They concluded that, while the MD&A text does not predict financial performance, environmental performance can be effectively identified using NLP techniques.

Wang et al. (2023) introduced the application of the MacBERT model(Cui et al., 2020), enhancing its capabilities with additional pre-training and contrastive learning strategies for the meticulous examination of ESG issues within the Chinese language track. In a similar vein, Pontes et al. (2023) employed a combination of models, including a Support Vector Machine (SVM) model (Platt, 1999) integrated with Sentence BERT (SBERT) embeddings (Reimers & Gurevych, 2019) and RoBERTa-based models (Liu et al., 2019), to classify multi-lingual ESG issues. Glenn et al. (2023) and Devlin et al. (2018) leveraged the potential of open-source large language models(LLM), notably gpt, for data augmentation purposes, thereby enhancing the performance of model. Mehra et al. (2022) made a notable contribution by developing ESGBERT, a tool specifically fine-tuned on a BERT model for sequence classification and conducting a Masked Language Model (MLM) task on an ESG-focused corpus, showcasing ESGBERT's efficacy in capturing the nuanced context of ESG for specialized text classification tasks.

Drawing inspiration from these pioneering contributions, our study leverages PLMs to classify the impact duration and level of ESG issues within news articles. By integrating the insights and methodologies from these notable works, our approach seeks to further refine the accuracy and applicability of NLP technologies in dissecting and understanding the complex domain of ESG disclosures, illustrating the interconnected progress within the field.

3. Proposed Methods

3.1 Dataset

Figure 1 shows the architecture used in this study.In the multi-lingual ESG-3 shared task, the organizers provided datasets in five languages, which were divided into different subtasks as outlined in Appendix 1. The training datasets included English,

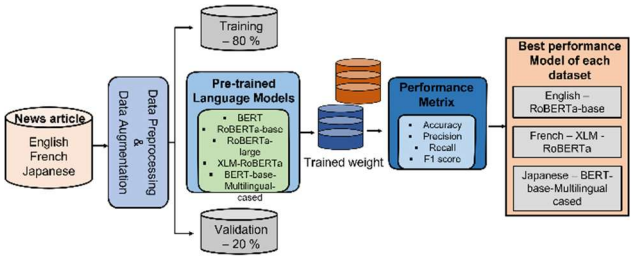


Figure 1: System architecture of Multi-Lingual ESG Impact type and Impact duration classification

French, Japanese, Korean, and Chinese. English, Korean and French datasets were associated with two subtasks, impact level and impact length, while Japanese, and Chinese had only one subtask, impact length. Team IMNTPU participated in three languages: English, French, and Japanese which consists of 545 news articles in the English dataset, 664 in the French dataset, and 50 news articles in the Japanese dataset, related to ESG issues.The English and French datasets include the following columns: "URL," "news_title," "news_content," "impact_level," and "impact_length." In contrast, the Japanese dataset contains "ID," "Text," "Relevancy," "ESG_type," "impact_type," and "impact_duration."

3.2 Data Augmentation

We observe that the dataset presents two primary challenges: a constrained overall size and uneven label distribution across different languages. To tackle these issues, we employed gpt-3.5-turbo in view of cost effective, an open-source large language model, for data augmentation purposes. This strategy not only expanded our dataset but also aimed at rectifying the imbalance in label distribution. Data augmentation, in this context, is crucial for enhancing the diversity and representativeness of our dataset, thereby improving model training outcomes and ensuring a more robust and accurate classification performance across the multilingual ESG classification task. Appendix 2 illustrates the prompt used to generate additional text, showcasing our methodology for augmenting the dataset effectively.

Figure 2 presents the dataset before and after augmentation. The English dataset has expanded from 545 to 11,556 news articles, the French dataset from 661 to 10,104 articles, and the Japanese dataset from 50 to 1,430 articles. Additionally, the label distribution for impact level and impact length is more balanced compared to the original dataset.

3.3 Pretrained language model

The surge in leveraging PLMs such as BERT and its transformer-based counterparts has marked a significant stride in the field of NLP, extending its impact to domain-specific applications. This growing fascination with large-scale language models, underscored by their remarkable efficacy across diverse NLP applications, is well-documented in recent scholarly discourse (Liu et al., 2023). In the ambit of our current endeavor, we strategically deployed an array of PLMs — namely, BERT, RoBERTa (both Base and Large variants), XLM-RoBERTa, and BERT-base-multilingual-cased.

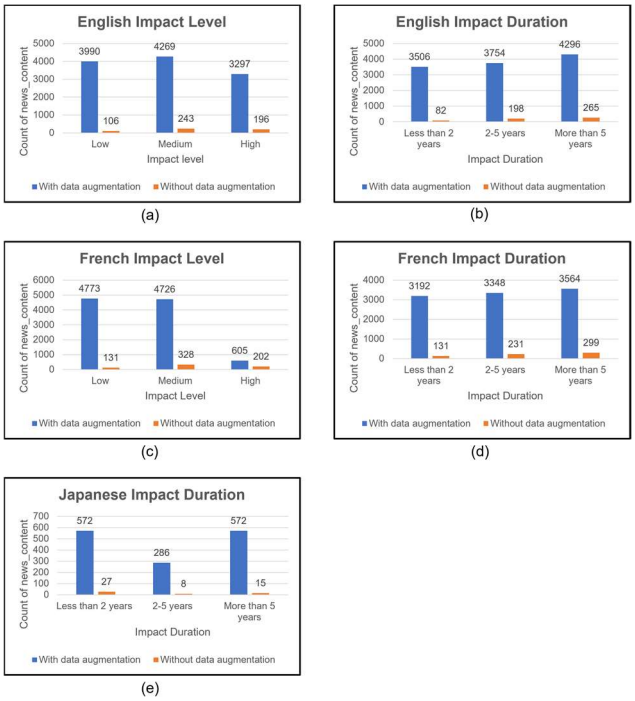


Figure 2: Comparative Analysis of Multilingual dataset before and after data augmentation (a) English Impact Level, (b) English Impact Duration, (c) French Impact Level, (d) French Impact Duration and (e) Japanese Impact Duration.

While these models share the foundational BERT architecture, they diverge in their pre-training approaches and the scale of parameters, which are pivotal in learning comprehensive language representations.

Our project aims to harness these models for the nuanced task of classifying the impact level and duration from the textual content of news articles. To this end, we utilize the transformative capabilities of Hugging Face's transformer models, meticulously chosen for their proficiency in comprehending and analyzing text. Given the multilingual nature of our dataset, we allocated the BERT and RoBERTa (Base and Large) for the English dataset, and the XLM-RoBERTa and BERT-base-multilingual-cased models for French and Japanese datasets, aligning with their inherent language processing strengths. Our methodology concentrates on parsing the news content, excluding titles, to derive predictions for designated labels.

To evaluate the models' performance, we first divided each original training dataset into an 80% slice for training purposes and a 20% segment for validation and we compared the models performance with data augmentation dataset. This structured approach not only amplifies the precision of our classification task but also underscores the adaptability of these PLMs in dissecting and understanding multilingual news narratives, setting a precedent for future research in

domain-specific NLP applications.The hyperparameters of each model is mentioned in Appendix 3.

4. Experimental Results

4.1 Submitted runs

In our comprehensive experimental setup, we evaluated the efficacy of five distinct models across English, French, and Japanese datasets, with our findings meticulously documented in Table 1. This table encapsulates the culmination of our official submissions and their corresponding performance metrics. Within the English dataset evaluations spanning three submission rounds, the RoBERTa model stood out, securing the premier position with an impressive Micro-F1 score of 58.82% and a Macro-F1 score of 55.03%. This achievement underscores RoBERTa's nuanced understanding and processing capabilities of the English language.

Transitioning to the French dataset, our exploration across two submission rounds revealed the XLM-RoBERTa model as the frontrunner, achieving a notable Micro-F1 score of 47.26% and a Macro-F1 score closely aligned at 47.16%. This result highlights XLM-RoBERTa's adeptness at navigating the linguistic intricacies of the French language, cementing its status as a potent tool for multilingual analysis.

Further delving into the Japanese dataset, again over two rounds of submissions, the Bert-base-multilingual-cased model emerged as the victor, albeit with a Micro-F1 score of 11.90% and a Macro-F1 score of 7.10%. Despite the lower scores relative to the other languages, this outcome signals the model's capacity to grapple with the Japanese language, albeit indicating potential areas for improvement and refinement.

Table 1 not only serves as a testament to the comparative strengths and areas for enhancement across the models but also illuminates the path forward for optimizing multilingual ESG classification tasks. The distinguished performance of the RoBERTa model in English, in particular, delineates a benchmark for excellence, suggesting a fertile ground for future investigations to build upon and extend its application across diverse linguistic landscapes. Appendix 4 and Appendix 5 shows the comparison report of performance metrics report obtained before and after data augmentation in the development dataset .

5. Conclusion

In conclusion, team IMNTPU engaged in the multi-lingual ESG classification task, with the aim of discerning impact levels and durations from ESG-related news articles across English, French, and Japanese datasets. Leveraging transformer models, notably the RoBERTa-base model, we focused on optimizing our approach to accurately classify the given information. The RoBERTa-base model, in particular, demonstrated superior performance in the English dataset, achieving a commendable Micro-F1 score of 58.82%, which stands as our best result. This was followed by the French dataset with a score of 47.26%, and the Japanese dataset at 11.90%, highlighting a significant opportunity for improvement in handling Japanese language data, potentially through parameter adjustments such as learning rate and epochs.

Dataset	Subtask	Runs	Team ID	Model	Micro-F1	Macro-F1
English	Impact Level	Run 1	English_IMNTPU_1	BERT	18.38%	15.54%
		Run 2	English_IMNTPU_2	RoBERTa-base	**58.82%**	**55.03%**
		Run 3	English_IMNTPU_3	RoBERTa-large	19.12%	17.22%
French	Impact Level	Run 1	French_IMNTPU_1	XLM-RoBERTa	**47.26%**	**47.16%**
		Run2	French_IMNTPU_2	BERT-base-multilingual cased	37.67%	34.46%
Japanese	Impact Length	Run1	Japanese_IMNTPU_1	XLM-RoBERTa	11.10%	5.00%
		Run 2	Japanese_IMNTPU_2	BERT-base-multilingual cased	**11.90%**	**7.10%**

Table 1: Official evaluation results submitted to ML-ESG 3

Additionally, our application of data augmentation techniques played a critical role in enhancing our model's performance, particularly by addressing issues of data scarcity and label imbalance. However the data augmentation has not improved the performance in all models but it improved the performance in RoBERTa-base model which outperformed in official run. The results underscore the effectiveness of the RoBERTa-base model and data augmentation in advancing our understanding and classification capabilities within the multi-lingual ESG domain.

5.1 Research Contributions

This study advances the field of ESG impact assessment by training with BERT-based Transformer Models across multilingual datasets to significantly improve the precision and comprehensiveness of ESG impact duration and level level. Furthermore, implementing the Data Augmentation to balance the class within the datasets, enhanced the model robustness and ensuring a

balanced representation for more accurate ESG impact duration and level classification.

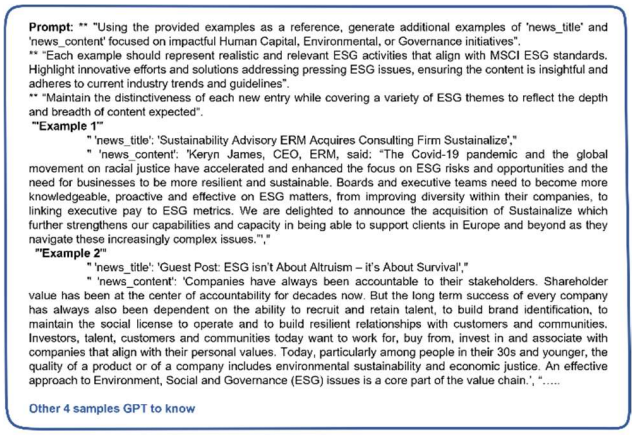
5.2 Managerial Implications

These advancements offer substantial benefits. First the enhanced models provide ESG analysts with more reliable tools for assessing the impact duration and level of sustainability initiatives, thus supporting more informed and strategic decision-making. And organizations can better align their operations with sustainable practices, accurately track their ESG performance, catering to a globally diverse audience.

6. Appendices

Dataset	Subtasks	Labels
English	Impact Level	low, medium, high
	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years
French	Impact Level	low, medium, high
	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years
Japanese	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years
Korean	Impact Level	Opportunity,risk,cannot distinguish
	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years
Chinese	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years

Appendix 1: Different classification subtasks for each language.



Prompt: ** "Using the provided examples as a reference, generate additional examples of 'news_title' and 'news_content' focused on impactful Human Capital, Environmental, or Governance initiatives".
** "Each example should represent realistic and relevant ESG activities that align with MSCI ESG standards. Highlight innovative efforts and solutions addressing pressing ESG issues, ensuring the content is insightful and adheres to current industry trends and guidelines".
** "Maintain the distinctiveness of each new entry while covering a variety of ESG themes to reflect the depth and breadth of content expected".
"Example 1"
" 'news_title': 'Sustainability Advisory ERM Acquires Consulting Firm Sustainalize',"
" 'news_content': 'Keryn James, CEO, ERM, said: "The Covid-19 pandemic and the global movement on racial justice have accelerated and enhanced the focus on ESG risks and opportunities and the need for businesses to be more resilient and sustainable. Boards and executive teams need to become more knowledgeable, proactive and effective on ESG matters, from improving diversity within their companies, to linking executive pay to ESG metrics. We are delighted to announce the acquisition of Sustainalize which further strengthens our capabilities and capacity in being able to support clients in Europe and beyond as they navigate these increasingly complex issues."',"
"Example 2"
" 'news_title': 'Guest Post: ESG isn't About Altruism – it's About Survival',"
" 'news_content': 'Companies have always been accountable to their stakeholders. Shareholder value has been at the center of accountability for decades now. But the long term success of every company has always also been dependent on the ability to recruit and retain talent, to build brand identification, to maintain the social license to operate and to build resilient relationships with customers and communities. Investors, talent, customers and communities today want to work for, buy from, invest in and associate with companies that align with their personal values. Today, particularly among people in their 30s and younger, the quality of a product or of a company includes environmental sustainability and economic justice. An effective approach to Environment, Social and Governance (ESG) issues is a core part of the value chain.', ".....

Other 4 samples GPT to know

Appendix 2: Prompt generated using GPT 3.5 Turbo

Model	Batch size	Epoch	Optimizer	Learning rate
BERT	16	10	Adam	5e-5
RoBERTa-base	16	10	Adam	5e-5
RoBERTa-Large	16	10	Adam	5e-5
XLM-RoBERTa	16	3	Adam	5e-5
BERT-base-multilingual-cased	16	3	Adam	5e-5

Appendix 3: The main hyperparameters used in this study

Dataset	Models	Accuracy	F1 score	Precision	Recall
English	BERT	0.82	0.82	0.82	0.82
	RoBERTa-base	0.79	0.79	0.79	0.79
	RoBERTa-large	0.34	0.17	0.12	0.34
French	XLM-RoBERTa	0.87	0.86	0.86	0.87
	BERT base-multilingual-cased	0.88	0.87	0.88	0.88
Japanese	XLM-RoBERTa	0.40	0.23	0.16	0.40
	BERT-base-multilingual-cased	0.36	0.19	0.13	0.36

Appendix 4: Performance matrix of development dataset after data augmentation

Dataset	Models	Accuracy	F1 score	Precision	Recall
English	BERT	0.64	0.61	0.61	0.64
	RoBERTa-base	0.53	0.50	0.50	0.53
	RoBERTa-large	0.49	0.22	0.16	0.33
French	XLM-RoBERTa	0.42	0.25	0.17	0.42
	BERT base-multilingual-cased	0.45	0.31	0.44	0.42
Japanese	XLM-RoBERTa	0.60	0.45	0.36	0.60
	BERT-base-multilingual-cased	0.40	0.34	0.30	0.40

Appendix 5: Performance matrix of development dataset without data augmentation

7. Acknowledgments

This research was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grants NSTC 112-2425-H-305-002-, and NSTC 112-2627-M-038-001-, and National Taipei University (NTPU), Taiwan under grants 113-NTPU_ORDA-F-003, 113- NTPU-ORDA-F-004, USTP-NTPU-TMU-113-03, and NTPU-112A413E01.

8. Bibliographical References

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Glenn, P., Gon, A., Kohli, N., Zha, S., Dakle, P. P., & Raghavan, P. (2023). Jetsons at the finnlp-2023: Using synthetic data and transfer learning for multilingual esg issue classification. Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting,

Kannan, N., & Seki, Y. (2023). Textual Evidence Extraction for ESG Scores. Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting,

Lee, O., Joo, H., Choi, H., & Cheon, M. (2022). Proposing an integrated approach to analyzing ESG data via machine learning and deep learning algorithms. Sustainability, 14(14), 8745.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., & Liu, Z. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology, 100017.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Mehra, S., Louka, R., & Zhang, Y. (2022). ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices. arXiv preprint arXiv:2203.16788.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3), 61-74.

Pontes, E. L., Benjannet, M., & Ming, L. K. (2023). Leveraging bert language models for multi-lingual esg issue identification. arXiv preprint arXiv:2309.02189.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Tseng, Y.-M., Chen, C.-C., Huang, H.-H., & Chen, H.-H. (2023). DynamicESG: A Dataset for Dynamically Unearthing ESG Ratings from News Articles. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management,

Wang, W., Wei, W., Song, Q., & Wang, Y. (2023). Leveraging contrastive learning with bert for esg issue identification. Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting,

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76.