ViBERT-ESG: Advancing ESG Classification in Vietnamese with ViEnESG-100 Dataset

Vanmanh Nguyen, Ductrung Tran, Kienlam Do, Thiloan Bui, and Vanha Tran(✉)

FPT University, Hanoi 155514, Vietnam
{manhnvhe170552,trungtdhe173146,lamdkhe172132}@fpt.edu.vn,
{loanbt7,hatv14}@fe.edu.vn

Abstract. Environmental, Social, and Governance (ESG) factors play a critical role in assessing corporate sustainability and ethical impact. While recent advances in natural language processing (NLP) have leveraged large pre-trained models for ESG classification, most prior work focuses on English-language data, limiting applicability in multilingual contexts. In this work, we present a BERT-based approach for ESG classification tailored to Vietnamese financial and sustainability reports. A key contribution of our work is the construction of the ViEn-ESG-100 dataset, a high-quality, manually labeled bilingual Vietnamese–English ESG dataset containing 103,656 sentences, which we plan to release publicly to foster further research. We fine-tune BERT-based models on this dataset and obtain strong performance across multiple ESG dimensions. Our results highlight the effectiveness of pre-trained language models for low-resource, domain-specific ESG classification and provide a useful evaluation resource for Vietnamese ESG text classification. Our dataset, model and code are available at https://github.com/nguyen599/ViBERT-ESG.

Keywords: ESG · Classification · BERT · Multilingual Dataset

1 Introduction

Environmental, Social, and Governance (ESG) factors have gained significant attention as crucial indicators for assessing corporate sustainability, ethical impact, and long-term value creation [8]. Evaluating a company's ESG performance reveals insights into non-financial aspects of their operations and helps stakeholders understand their contributions to sustainable development [20]. The increasing need for objective and efficient evaluation of companies' ESG activities has highlighted limitations in traditional manual assessment methods. Manual analysis of large volumes of ESG-related text, such as sustainability reports and news articles, is time-consuming and prone to human error and subjective judgment. In addition, differences in evaluation criteria across ESG rating agencies make consistent assessment difficult [10].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16766, pp. 226–235, 2027.
https://doi.org/10.1007/978-3-032-30539-8_16

Recent advances in NLP provide promising tools for automating ESG analysis. Existing studies have applied NLP to tasks such as ESG scoring, text classification, risk detection, and information extraction from reports and financial news [13].

A significant focus in this research area involves leveraging powerful pretrained language models (PLMs), particularly those based on the Transformer architecture like BERT [3] and its variants like RoBERTa [12], DistilBERT [17], Sentence-BERT [16], FinBERT [5]. When it comes to identifying language and retrieving pertinent data for ESG evaluation, these models have proven to perform quite well. For instance, the E-BERT model [20], specifically trained based on Google's BERT architecture for ESG rating tasks, achieved a notable 93% accuracy in automating the evaluation process, aiming for accurate and reliable results. In tasks such as recognizing ESG concerns or classifying words as sustainable or unsustainable, several BERT-based models have demonstrated great performance. Furthermore, SEC-BERT finetune [19] developed a domain-adapted SEC-BERT [13] model for detecting ESG-related issues from financial news, achieving improved performance through in-domain pre-fine-tuning. Their system classifies news into 33 ESG issue types and powers the ESG Issue Detector (EID) tool, enabling investors to identify ESG risks efficiently. The study also explored multilingual augmentation and LLM-based zero-shot baselines, finding domain-specific models to be significantly more effective.

ESG-KIBERT [9], a BERT-based model further pre-trained on ESG-specific corpora, which incorporates a hard attention mechanism to enhance ESG classification performance. The model achieved 99.72% accuracy on a four-class (E, S, G, None) task and was combined with sentiment analysis and industry-specific weighting using SASB's materiality map to generate transparent, domain-aware ESG ratings. ESGify [7], an open-source NLP-based model built on MPNet [18] for multi-label ESG risk classification. Leveraging a manually annotated dataset of 14,000 sentences from ESG reports, the model achieved superior performance to GPT-3.5 using tailored data augmentation strategies such as back translation and LLM-generated examples.

Despite these advances, one of the main challenges is the lack of comprehensive, labeled ESG datasets—especially for low-resource languages. While multilingual ESG tasks are emerging, involving datasets in languages like Chinese [14], English [1], French, Japanese [6], and Korean [11], a significant portion of the existing research has focused primarily on English-language data. This limits the applicability of existing ESG models in other languages, especially in low-resource settings where labeled data is limited.

To address this gap, we introduce a Vietnamese–English ESG classification pipeline based on a new bilingual dataset and a comparative study of transformer backbones. Beyond ESG classification, this work also contributes to multilingual NLP through bilingual dataset construction, annotation quality control, and cross-lingual evaluation. Our main contributions are as follows:

1. We introduce the **ViEn-ESG-100** dataset, a new large-scale, bilingual (Vietnamese–English) dataset comprising 103,656 sentence-level ESG samples.

Recognizing the general lack of such resources, we created this dataset to support and accelerate future research in Vietnamese ESG analysis and multilingual ESG text classification.

2. We present BERT-based ESG classification models tailored to Vietnamese financial and sustainability reports, and we provide a comparative evaluation across several transformer backbones in both Vietnamese and English settings.

3. This work provides strong baseline results for automated ESG classification in Vietnamese and offers practical resources to support ESG analysis for stakeholders and researchers.

2 Methodology

Our methodology is designed for the low resource setting of Vietnamese ESG analysis while taking advantage of recent progress in NLP. We fine tuning pre-trained BERT-based models for the ESG classification task, since these models have shown strong performance across many NLP applications. This approach is suitable for Vietnamese because large annotated ESG datasets are still limited. By adapting pre-trained models to our ESG dataset, we aim to build a classification system that is accurate, efficient, and able to capture the characteristics of ESG reporting in Vietnamese documents. The overall ESG classification pipeline is shown in Fig. 1.

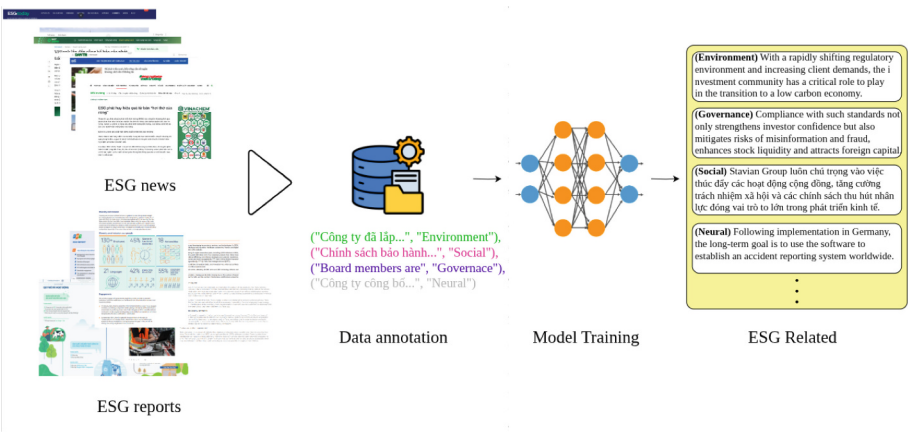

Fig. 1. Overview of the proposed ESG classification pipeline, including data preparation, bilingual annotation, model fine-tuning, and evaluation.

2.1 Model Selection

We experiment with several BERT-based backbones for ESG classification to compare how different pre-training strategies perform in our bilingual setting.

Our goal is not to propose a new architecture, but to examine which transformer models adapt most effectively to Vietnamese ESG text and whether similar trends also appear in English.

We focus on transformer models because they have become a standard choice for sentence classification, information extraction, and other NLP tasks where context matters. ESG statements often contain indirect wording, long dependencies, and domain-specific expressions, so models that can capture contextual meaning are a natural fit for this problem.

Our comparison includes BERT-base-cased [3], RoBERTa-base [12], XLM-RoBERTa-base [2], DeBERTa-v3-base and DeBERTa-v3-small [4], and DistilBERT-base-cased [17]. These models differ in size, pre-training strategy, and multilingual capacity, which makes them useful reference points for our experiments. BERT serves as a strong baseline, RoBERTa and DeBERTa represent alternative pre-training and encoding strategies, XLM-RoBERTa is included for its multilingual design, and DistilBERT provides a lighter-weight alternative with lower computational cost.

We fine-tune these models instead of training from scratch because our dataset is still much smaller than the large corpora typically used for language model pre-training. Using pre-trained checkpoints allows the models to reuse general language knowledge and focus on learning ESG-specific patterns. This approach is also more practical, since adapting existing models is more realistic than training a new model from scratch.

3 Dataset

The **ViEn-ESG-100** dataset is the main empirical resource used in this study. We created it to support ESG sentence classification in a bilingual Vietnamese–English setting, with particular attention to the lack of public Vietnamese data for this task. The final dataset contains 103,656 sentence-level samples and is intended to support both model training and controlled evaluation across ESG categories.

3.1 Data Collection

The sentences constituting the ViEn-ESG-100 dataset were sourced from a diverse range of publicly available documents relevant to corporate activities. The authors used two main data sources: ESG articles and reports. Articles were collected with the tag “ESG” from January 1, 2015 to January 14, 2025, for English articles, data were collected at ESGToday. Total of 78,291 sentences collected from the news. Annual ESG earnings report data from Vietnamese and S&P500 companies across ten key economic sectors, including technology, retail, consumer goods, health care, automotive, aviation, electronics, communication services, finance, and heavy industry, with publication dates spanning from 2012 to 2024 to capture evolving ESG practices and reporting standards over more than a decade. We collected 448 reports and a total of 45,103 text

data were manually collected. Table 1 shows the class-wise distribution of the ViEn-ESG-100 dataset.

Table 1. Class-wise distribution of the ViEn-ESG-100 Dataset.

Data		No. of Sentences
English		Vietnamese
Environment	10,517	14,712
Social	11,112	15,201
Government	9,103	11,030
Neural	15,210	16,771
Total	45,942	57,714

3.2 Annotation Methodology and Quality Assurance

To improve label quality, we relied on expert manual annotation rather than weak supervision or automatic labeling. Five trained annotators with backgrounds in finance and sustainability followed a shared guideline that defined the Environmental, Social, Governance, and Neutral categories. The work was carried out in two rounds: an initial annotation pass and a cross-check pass. Sentences that still lacked unanimous agreement after review were removed from the final dataset.

The Vietnamese portion contains 57,714 sentences and was labeled with attention to local reporting patterns and language use. The English portion contains 45,942 sentences drawn from public ESG disclosures and annual reports and was independently reviewed by three domain experts before adjudication. We measured agreement with Fleiss' Kappa, and the scores are reported in Table 2. Overall, the agreement values indicate that the annotation protocol was stable across both languages and suitable for downstream evaluation.

Table 2. Fleiss's Kappa coefficients measuring inter-annotator agreement across ESG categories and languages.

Category	English	Vietnamese
Environment	0.94	0.92
Social	0.91	0.90
Governance	0.93	0.89
Neutral	0.95	0.93
Overall	**0.93**	**0.91**

The agreement scores support the reliability of the annotation process and suggest that the label definitions are sufficiently clear for bilingual ESG classification. This manually curated dataset is therefore used as the core training

and evaluation resource throughout the paper. Figure 2 shows the token-length distribution obtained with the BERT-base-multilingual-cased tokenizer.

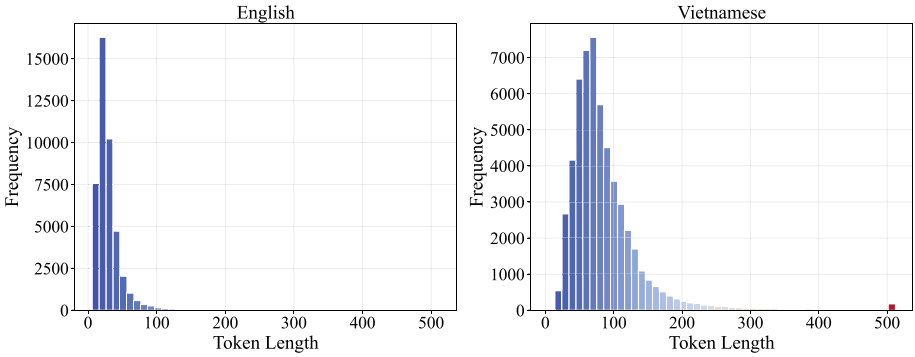

Fig. 2. Token-length distribution of the ViEn-ESG-100 dataset under the BERT-base-multilingual-cased tokenizer.

4 Results and Evaluation

This section describes the training setup and summarizes the main empirical results. We report performance for both English and Vietnamese and compare our bilingual models with several existing ESG baselines under binary and four-class settings.

4.1 Model Configuration and Training Strategy

We trained two types of models. The first group consists of three binary classifiers, one for each ESG pillar: Environmental, Social, and Governance. The second group consists of four-class models that assign each sentence to one of the labels E, S, G, or Neutral. The four-class setting is our main evaluation setup because it reflects the full label space of the dataset, while the binary models provide an additional view of category-specific performance.

Each experiment was conducted on a computer equipped with four RTX 3060 GPUs, 96 GB of RAM, and an Intel Core i7-13700 CPU. Torch 2.6.0, Datasets 3.6.0, Accelerate 1.7.0, and Transformers 4.51.0 are required for the implementation.

We used DeepSpeed ZeRO-3 optimization for full-parameter fine-tuning during training [15], 2,000 steps including 100 warm-up steps, were used to train the models in BFloat16. We used AdamW with a learning rate of 2×10^{-5}, cosine scheduling, weight decay of 0.01, a batch size of 48, a maximum gradient norm of 1, and a maximum input length of 512 tokens.

4.2 Experimental Result

Standard classification measures were used to assess the performance of our trained models, with an emphasis on the F1-score. The F1-score, which represents the harmonic mean of precision and recall, provides a balanced measure of a model's accuracy, particularly in scenarios with potential class imbalance, which can occur in ESG datasets where certain categories might be more prevalent than others. For the multi-class models, we offer F1-scores for each individual class (E, S, G, and N) as well as overall performance measures when applicable.

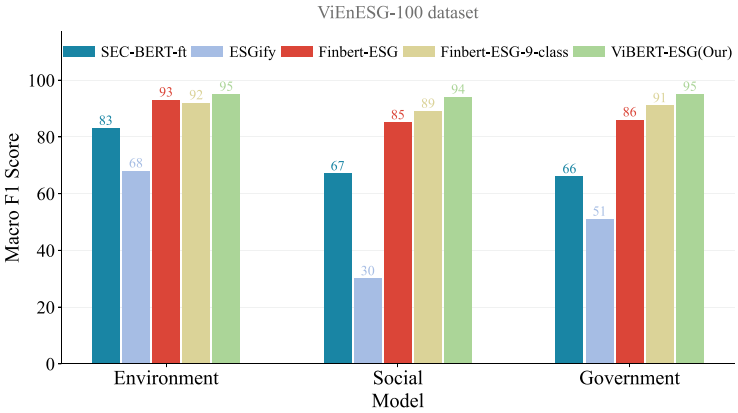

Fig. 3. Model Performance on ViEn-ESG-100 dataset.

For the English portion of the dataset, we benchmarked several publicly available ESG models, including FinBERT-ESG, SEC-BERT-ft, ESGify, and three binary classification models. These models differ in label granularity: FinBERT-ESG-9-class predicts 9 ESG-related categories, SEC-BERT-ft supports 33 fine-grained issue types, and ESGify handles 47 labels. To enable consistent comparison, we mapped all predicted labels into a unified 4-class format E/S/G/N. Figure 3 illustrates the performance of these models on the ViEn-ESG-100 dataset.

For English, ViDeBERTa-ESG-base achieved the highest F1-score on the Environment category (95.50%), while ViXLMRoBERTa-ESG-base slightly outperformed others on Governance (95.47%). Across all models, our multilingual fine-tuned versions consistently surpassed existing English-only ESG models, demonstrating the benefit of cross-lingual semantic transfer from bilingual training. In particular, lightweight models such as ViDistilBERT-ESG-base maintained competitive results (average F1 of 94.6%) while reducing computational cost.

For the Vietnamese subset, ViDeBERTa-ESG achieved the strongest performance overall, with F1-scores of 95.24% (Environment) and 93.18% (Governance), indicating robust understanding of domain-specific ESG contexts. Mean-

Table 3. F1 scores of models on each ESG category in the English ViEn-ESG-100 dataset.

Model	Backbone	# Param	Env	Soc	Gov	Neu
SEC-BERT-ft	SEC-BERT-base	109M	83.12	66.77	66.53	60.30
FinBERT-ESG	FinBERT	109M	92.67	84.90	86.25	87.26
FinBERT-ESG-9-class*	FinBERT	109M	92.16	89.01	91.35	86.89
ESGify	MPNet-base	109M	67.72	30.20	50.76	43.44
EnvironmentBERT	DistilRoBERTa	82M	92.15	–	–	92.76
SocialBERT	DistilRoBERTa	82M	–	76.81	–	81.23
GovernanceBERT	DistilRoBERTa	82M	–	–	64.46	80.06
ViBERT-ESG	BERT-base-multilingual-cased	168M	93.76	94.53	94.98	**94.15**
ViRoBERTa-ESG-base	RoBERTa-base	124M	95.43	94.06	95.01	91.32
ViXLMRoBERTa-ESG-base	XLM-RoBERTa-base	278M	95.00	95.00	**95.47**	92.19
ViDeBERTa-ESG-base	DeBERTa-v3-base	184M	**95.50**	94.49	94.81	91.48
ViDeBERTa-ESG-small	DeBERTa-v3-small	141M	94.55	94.85	94.58	90.19
ViDistilBERT-ESG-base	DistilBERT-base-multilingual-cased	135M	95.15	**95.19**	94.33	91.75
ViBERT-Env**	BERT-base-multilingual-cased	168M	94.62	–	–	92.13
ViBERT-Soc**	BERT-base-multilingual-cased	168M	–	94.86	–	92.22
ViBERT-Gov**	BERT-base-multilingual-cased	168M	–	–	93.47	93.82

Table 4. F1 scores of models on each ESG category in the Vietnamese ViEn-ESG-100 dataset.

Model	Backbone	# Param	Env	Soc	Gov	Neu
ViBERT-ESG	BERT-base-multilingual-cased	168M	93.50	89.73	91.77	**91.78**
ViRoBERTa-ESG	RoBERTa-base	124M	93.41	91.49	89.93	84.32
ViXLMRoBERTa-ESG	XLM-RoBERTa-base	278M	93.45	91.02	91.69	90.41
ViDeBERTa-ESG	DeBERTa-v3-base	184M	**95.24**	89.36	**93.18**	85.23
ViDeBERTa-small-ESG	DeBERTa-v3-small	141M	92.90	87.79	90.63	81.48
ViDistilBERT-ESG	DistilBERT-base-multilingual-cased	135M	93.87	**91.98**	90.63	87.17
ViBERT-Env*	BERT-base-multilingual-cased	168M	94.87	–	–	91.15
ViBERT-Soc*	BERT-base-multilingual-cased	168M	–	91.07	–	90.29
ViBERT-Gov*	BERT-base-multilingual-cased	168M	–	–	92.62	90.11

while, ViDistilBERT-ESG demonstrated the best performance in the Social category (91.98%), highlighting its efficiency and adaptability despite fewer parameters. The consistency of results across both languages shows that bilingual fine-tuning effectively preserves semantic alignment while enhancing model generalization.

Since existing ESG models are primarily designed for English and do not support Vietnamese input, our models are fine-tuned jointly on both English and Vietnamese samples from the ViEn-ESG-100 dataset. The models are better able to generalize across languages and take advantage of the shared semantic

field of ESG-related concepts thanks to this multilingual training arrangement. We assess our improved models using test sets in both Vietnamese and English.

This evaluation provides a clear comparison of different model backbones and configurations across ESG categories. The results in Tables 3 and 4 show strong performance across all ESG dimensions in both English and Vietnamese. Overall, these findings establish a strong multilingual benchmark for ESG classification in English and Vietnamese settings.

5 Conclusion and Future Work

This study focuses on building a practical ESG text classification system for Vietnamese, where public datasets and evaluation resources remain limited. To address this gap, we introduced the **ViEn-ESG-100** dataset and evaluated several transformer-based models in both Vietnamese and English. The results show that pre-trained language models can achieve strong ESG classification performance after fine-tuning on a carefully constructed bilingual dataset.

The main contribution of this work is the creation of a bilingual ESG resource together with baseline benchmarks for Vietnamese ESG analysis. We hope the dataset and experimental results can support future research on multilingual and low-resource ESG NLP tasks.

Future work may extend the dataset with additional ESG documents, explore more advanced tasks such as ESG entity recognition and issue extraction, and evaluate these models in real-world ESG monitoring systems. Overall, we hope this work provides a useful foundation for future Vietnamese ESG NLP research.

References

1. Chen, C., et al.: 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, pp. 111–115. Association for Computational Linguistics (2023). https://aclanthology.org/2023.finnlp-2.18/
2. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019). https://doi.org/10.48550/arXiv.1911.02116
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). https://doi.org/10.48550/arXiv.1810.04805
4. He, P., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention. arXiv preprint arXiv:2006.03654 (2020). https://doi.org/10.48550/arXiv.2006.03654
5. Huang, A., Wang, H., Yang, Y.: FinBERT: a large language model for extracting information from financial text. Contemp. Account. Res. **40**(2), 806–841 (2023). https://doi.org/10.1111/1911-3846.12832
6. Kao, Y., et al.: IMNTPU at ML-ESG-3: transformer language models for multi-lingual ESG impact type and duration classification. In: Proceedings of FinNLP@LREC-COLING 2024, pp. 228–233. Association for Computational Linguistics (2024). https://aclanthology.org/2024.finnlp-1.38/

7. Kazakov, A., et al.: ESGify: automated classification of environmental, social, and corporate governance risks. Dokl. Math. **108**(S2), S529–S540 (2023). https://doi.org/10.1134/S1064562423701673
8. Kiriu, T., Nozaki, M.: A text mining model to evaluate firms' ESG activities: an application for Japanese firms. Asia-Pacific Finan. Markets. **27**(4), 621–632 (2020). https://doi.org/10.1007/s10690-020-09309-1
9. Lee, H., Kim, J., Jung, H.: ESG-KIBERT: a new paradigm in ESG evaluation using NLP and industry-specific customization. Decis. Support Syst. **193**, 114440 (2025). https://doi.org/10.1016/j.dss.2025.114440
10. Lee, J., Kim, M.: ESG information extraction with cross-sectoral and multi-source adaptation based on domain-tuned language models. Expert Syst. Appl. **221**, 119726 (2023). https://doi.org/10.1016/j.eswa.2023.119726
11. Lee, J., Son, G., Kim, M.: ESG-KOR: a Korean dataset for ESG-related information extraction and practical use cases. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 6627–6643. (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.387
12. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019). https://doi.org/10.48550/arXiv.1907.11692
13. Loukas, L., et al.: FiNER: financial numeric entity recognition for XBRL tagging. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.acl-long.303
14. Pontes, E., Benjannet, M., Ming, L.: Leveraging BERT language models for multilingual ESG issue identification. arXiv preprint arXiv:2309.02189 (2023). https://doi.org/10.48550/arXiv.2309.02189
15. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: memory optimizations toward training trillion parameter models. arXiv preprint arXiv:1910.02054 (2019). https://doi.org/10.48550/arXiv.1910.02054
16. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084 (2019). https://doi.org/10.48550/arXiv.1908.10084
17. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019). https://doi.org/10.48550/arXiv.1910.01108
18. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.: MPNet: masked and permuted pre-training for language understanding. arXiv preprint arXiv:2004.09297 (2020). https://doi.org/10.48550/arXiv.2004.09297
19. Soni, P., Ghosh, S., Naskar, S.: Detecting issues related to environmental, social, and corporate governance using SEC-BERT. In: Data Science and Applications, pp. 329–338. Springer, Singapore (2024). https://doi.org/10.1007/978-981-99-7820-5_27
20. Zhang, M., Shen, Q., Zhao, Z., Wang, S., Huang, G.: Optimizing ESG reporting: Innovating with E-BERT models in natural language processing. Expert Syst. Appl. **265**, 125931 (2025). https://doi.org/10.1016/j.eswa.2024.125931