FPT Education

FPT UNIVERSITY

FPT University

ESG Analysis and Scoring in Vietnamese Corporate Sustainability Reports

By
Tran Duc Trung, Nguyen Van Manh,
Nguyen Huu Duy Anh, Do Kien Lam

Supervisor: PhD. Tran Van Ha

Thursday 11th September, 2025

Project Specification

The authors confirm their contribution to the thesis as follows:

Tran Duc Trung:

- Methodology
- Implementation.
- Data preparation.
- Paper research.

Nguyen Van Manh:

- Methodology.
- Implementation.
- Paper research.
- Data preparation.

Nguyen Huu Duy Anh:

- Software design.
- Implementation.
- Paper research.
- Data preparation.

Do Kien Lam:

- Software design.
- Implementation.
- Paper research.
- Data preparation.

Abstract

Environmental, Social, and Governance (ESG) factors have become critical indicators for assessing corporate sustainability and ethical practices. However, existing ESG evaluation systems often face issues such as inconsistent criteria and limited linguistic coverage, particularly for Vietnamese data. To address these challenges, we introduce a comprehensive framework for ESG text classification and scoring. We developed the ViEn-ESG dataset, a high-quality bilingual Vietnamese-English dataset with 130,798 pairs of sentences - labels, which we release publicly to foster further research. We fine-tuned a BERT-based model for ESG classification specifically tailored to Vietnamese financial and sustainability reports. Our models exhibit strong classification performance with an accuracy of 94.66% on the Vietnamese-only model, in the bililingual model archive 94.83% F1 score on the English language, and 91.94% on the Vietnamese. In addition to classification, we also implemented a robust scoring mechanism to compute ESG scores by aggregating sentiment signals within each ESG dimension, providing a more interpretable and quantitative assessment of corporate sustainability. Our results highlight the effectiveness of pre-trained language models in low-resource settings, and an integrated scoring mechanism, collectively offering new benchmarks and practical tools for ESG analytics in the Vietnamese market. ESG, BERT, classification, machine learning

Keywords: ESG Classification, ESG Scoring, ESG Dataset.

Contents

List of Figures

List of Tables

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 [24]. The rise of ESG as a fundamental framework for evaluating corporate performance reflects a paradigm shift in how stakeholders—including investors, regulators, consumers, and communities—assess business value beyond traditional financial metrics. This transformation has been driven by mounting evidence that companies with strong ESG practices demonstrate superior long-term financial performance, reduced operational risks, and enhanced stakeholder trust [3].

Evaluating a company's ESG performance reveals insights into non-financial aspects of their operations and helps stakeholders understand their contributions to sustainable development [52]. Environmental factors encompass a company's impact on natural resources, climate change mitigation efforts, pollution control measures, and commitment to circular economy principles. Social considerations include labor practices, community engagement, product safety, data protection, and diversity and inclusion initiatives. Governance aspects focus on board composition, executive compensation, anti-corruption measures, transparency in reporting, and stakeholder rights protection [4]. The comprehensive assessment of these dimensions provides a holistic view of corporate responsibility and sustainable business practices.

The increasing need for objective and efficient evaluation of companies' ESG activities has highlighted significant limitations in traditional manual assessment methods. Manual analysis of extensive textual data like sustainability reports, annual filings, news articles, and regulatory documents is not only laborious and time-consuming but also susceptible to human error, subjective biases, and potential manipulation [33]. The sheer volume of ESG-related information generated by corporations has grown exponentially, with sustainability reports now spanning hundreds of pages and containing complex, nuanced language that requires domain expertise to interpret accurately. Furthermore, differing evaluation criteria across various ESG rating agencies—such as MSCI, Sustainalytics, and Refinitiv—often lead to inconsistent and sometimes contradictory assessments of the same company, complicating objective judgment for investors and other stakeholders [5, 28].

Recent advances in NLP offer promising solutions to automate and standardize ESG evaluations, addressing both scalability and consistency challenges. The application of NLP techniques to ESG analysis represents a rapidly evolving field that encompasses multiple sophisticated tasks [2]. Researchers are leveraging NLP for automated ESG scoring systems that can process vast amounts of textual data in real-time, classify text into specific ESG categories with high granularity, identify and quantify ESG-related risks and opportunities, extract structured information from unstructured reports and news sources, detect emerging ESG issues from financial communications, and perform sentiment analysis on ESG-related disclosures [32]. These applications extend beyond simple classification to include complex tasks such as ESG materiality assessment, impact measurement, and predictive modeling of ESG performance trends.

A significant focus in this research area involves leveraging powerful pre-trained language models (PLMs), particularly those based on the Transformer architecture such as BERT [13] and its variants including RoBERTa [30], DistilBERT [44], Sentence-BERT [43], and domain-specific models like FinBERT and FinBERT-ESG [19]. These models have demonstrated exceptional performance in understanding complex financial and sustainability language, capturing contextual nuances that are critical for accurate ESG assessment. However, despite their success, current approaches to ESG classification using these models typically rely on adding task-specific classification heads during fine-tuning, which introduces several methodological concerns that have not been adequately addressed in the literature.

However, existing ESG classification systems often struggle with multilingual applicability, particularly for emerging markets where ESG reporting is conducted in local languages. The majority of available ESG datasets and models are English-centric, creating significant barriers for comprehensive global ESG analysis. This limitation is particularly problematic for international investors and multinational corporations that need to assess ESG performance across diverse linguistic and cultural contexts. The lack of high-quality, annotated ESG datasets in languages other than English has been a persistent bottleneck in developing truly global ESG analysis capabilities.

In this thesis, we address this gap by introducing a comprehensive ESG classification pipeline tailored for the Vietnamese language. Our key contributions include:

1. We introduce **ViEn-ESG** dataset, a large-scale bilingual (Vietnamese–English) dataset comprising 130,798 sentence-level samples related to ESG, along with an additional subset of 6,430 sentiment-labeled sentences. To the best of our knowledge, this is the first publicly available resource of its kind, designed to advance research in Vietnamese ESG analysis.

2. We develop a state-of-the-art BERT-based model for ESG classification, fine-tuned on our dataset. This includes the first dedicated model for Vietnamese ESG classification, achieving strong performance in both English and Vietnamese, thereby demonstrating the effectiveness of pre-trained language models in low-resource, domain-specific settings.

3. We introduce a robust ESG scoring mechanism that aggregates sentiment signals across Environmental, Social, and Governance dimensions, offering an interpretable and quantitative approach to assessing corporate sustainability.

4. This thesis establishes a new benchmark for automated ESG classification in Vietnamese and provide practical tools to support ESG analysis for researchers, practitioners, and stakeholders.

2. Related Work

2.1. ESG Text Classification Approaches

The E-BERT model [52], specifically trained based on Google's BERT architecture for ESG rating tasks, achieved a notable 93% accuracy in automating the evaluation process, aiming for precise and consistent outcomes. Other BERT-based models have shown strong results in tasks like classifying sentences as sustainable or unsustainable or identifying ESG issues. Furthermore, SEC-BERT finetune [48] developed a domain-adapted SEC-BERT [32] 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 [27], 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 [23], an open-source NLP-based model built on MPNet [47] 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, recent developments have also explored ESG performance assessment through machine learning approaches [20], where researchers have developed frameworks for evaluating corporate sustainability performance using advanced analytical techniques and structured datasets from Fortune 500 companies. The emergence of domain-specific ESG datasets has extended to specialized applications such as ESG rating explainability [11], where researchers have developed annotated corpora specifically designed to understand and interpret ESG scoring methodologies through machine learning techniques.

2.2. ESG Scoring and Evaluation Frameworks

The Rasch model [46] proposed an innovative framework for ESG scoring by combining NLP with Item Response Theory (IRT). Using Portuguese news articles on Petrobras, the method extracted ESG-related sentiment and validated its psychometric robustness, offering more reliable temporal dynamics in ESG performance. Similarly, Patel et al. [39] developed a systematic ESG scoring system by leveraging Social Network Analysis (SNA) and machine learning, highlighting how stakeholder networks and textual sentiment can be jointly modeled to produce more holistic and sustainable ESG ratings. Another study [51] provided a survey of sentiment-based ESG scoring methods, emphasizing that probability-based scoring with advanced deep learning models such as FinBERT offers greater sensitivity to nuanced language compared to ratio-based sentiment methods.

Building on these foundations, ESG-KIBERT [27] represents a paradigm shift in sentiment-driven ESG evaluation. The model extends BERT through pre-training on ESG-specific corpora and performs joint ESG classification and sentiment analysis at the sentence level. ESG-KIBERT further integrates SASB's industry-specific materiality weights, ensuring that sentiment scores are not only linguistically accurate but also financially and sector-relevant. Empirical results show that it outperformed both ratio-based and evidence-based models, achieving 99.72% accuracy and strong alignment with MSCI benchmarks. Its transparent aggregation of sentiment polarity and sector-sensitive weights establishes it as a state-of-the-art method for ESG scoring.

2.3. ESG-Specific Datasets and Resources

The development of specialized datasets for domain-specific NLP and AI applications has become increasingly important as researchers seek to address the unique challenges and requirements of various professional fields. While general-purpose datasets like OpenWebText [18], Common Crawl [10], and The Pile [17] have provided broad linguistic coverage for pre-training large language models, domain-specific applications require more targeted resources. In the biomedical domain, well-known examples include MIMIC-III [21] for clinical note analysis, PubMed[41] for biomedical literature mining, and BioASQ-QA [25] for biomedical question answering. In the financial domain, datasets such as FiQA [34] for question answering and sentiment analysis in financial texts have enabled progress in specialized financial NLP. The computer vision community has benefited enormously from ImageNet [12], which has driven advances in image classification and transfer learning, while speech processing has advanced with corpora like LibriSpeech [38] for automatic speech recognition. These domain-specific resources have demonstrated the critical importance of targeted datasets in developing robust AI systems capable of understanding the nuanced language, visual information, and specialized terminology of professional and technical domains.

In the ESG domain, datasets are scarce. Early ESG datasets were predominantly English-centric and relatively small in scale. Schimanski et al. [45], developed three specialized 2,000-sentence datasets designed for precise ESG text classification. Another notable dataset was produced by [16], the authors gathered more than 450,000 headlines, with about 27,000 ESG-related headlines from the Twitter account of The Guardian. They then manually mapped relevant headlines to ESG aspects after examining the tags for ESG-relatedness.

Multilingual approaches have gained particular attention, such as ESG-Kor [29] represents a pioneering effort in multilingual ESG dataset development, providing a comprehensive Korean dataset comprising 118,946 sentences for ESG-related information extraction with manual annotations based on objective rules from ESG evaluation agencies. This dataset specifically targets the extraction of Environmental, Social, and Governance information from Korean companies' sustainability reports, addressing the critical gap in non-English ESG resources and demonstrating significant classification performance with Korean pre-trained language models. While multilingual ESG tasks are emerging, involving datasets in languages like Chinese [40], English [8] and French, Japanese [22], a significant portion of the existing research has focused primarily on English-language data. More recently, The ESG-Activities benchmark [6] introduces a novel approach to ESG text classification by focusing on fine-grained environmental activities comprising 1,325 labeled text segments as defined in the EU ESG taxonomy. This dataset uniquely combines manually curated data with synthetically generated examples to enhance model performance, demonstrating that fine-tuned smaller models can outperform larger proprietary solutions in specific ESG classification tasks.

The remainder of this thesis is organized as follows. In Section 3, we describe the methodology and present the selected language models used for ESG classification. Section 4 introduces the ViEn-ESG dataset, detailing the data collection process, annotation methodology, and data analysis. In Section 5, we outline the model configuration, followed by a comprehensive evaluation of model performance across Vietnamese and multilingual ESG classification tasks. Section 6 shows our demo for the ESG classification tool. Finally, Section 7 concludes the thesis and discusses directions for future work.

3. Methodology

Our approach leverages state-of-the-art Natural Language Processing (NLP) techniques, specifically fine-tuning a pre-trained transformer-based language model, to address the task of ESG classification in the Vietnamese language. ESG factors are considered crucial indicators for assessing corporate sustainability and value creation, and NLP has emerged as a valuable tool for analyzing related textual data. Given the identified challenge of limited comprehensive, labeled ESG datasets, particularly for low-resource languages like Vietnamese, our methodology focuses on building a robust classification system supported by a newly constructed large-scale dataset.

3.1. ESG Classification using BERT-based Models

BERT [13] is an advanced language model used in many NLP tasks, first introduced in the paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". The power of BERT comes from several factors. First, BERT has a compact size, with only 110 million parameters. Second, the neural network is purposefully designed to capture complex relationships between words and sentences. BERT is an encoder model based on the Transformer [50] architecture — with the self-attention mechanism at its core — designed to create contextual representations for words or sentences. With its bidirectional nature, BERT excels at understanding the overall semantics of a sentence, making it very suitable for tasks such as ESG information classification. Third, BERT was pretrained on large text datasets including BookCorpus and English Wikipedia, enabling the model to learn both word and sentence representations and facilitating transfer learning.

BERT is a transfer learning framework, and its use typically involves two stages: pretraining and fine-tuning. Many BERT models have been pretrained on different unlabeled text datasets. These models can then be applied directly to a variety of tasks: text classification, named entity recognition (NER), question answering, or sentiment classification. After the original BERT was introduced, several improved models, such as RoBERTa [30], ALBERT [26], DistilBERT [44], were developed based on BERT's architecture, achieving higher performance through slight modifications to model design or pretraining hyperparameters. These pretrained models are publicly available to the research community and can be further fine-tuned for specific language tasks.

We experiment with a suite of BERT-based models, each fine-tuned specifically for ESG classification. These models are based on established architectures known for their success in a range of NLP classification tasks. Our goal is to evaluate the effectiveness of different model bases in the Vietnamese ESG context.

The standard approach to classification tasks using BERT-based models involves a structured fine-tuning process that follows a multi-stage pipeline that transforms input text into classification predictions through dedicated task-specific components.

Consider an input sentence $X = \{x_1, x_2, \ldots, x_n\}$ of length n. The classification pipeline can be formally represented through the following sequential transformations.

Input Embedding: The input sentence undergoes an initial embedding transformation that converts discrete tokens into dense vector representations:

$$H_0 = \text{Embedding}(X) \tag{1}$$

Where $H_0 \in \mathbb{R}^{n \times d}$ represents the initial token embeddings with dimensionality d, including special tokens like [CLS] and [SEP] that provide structural information to the model.

Transformer Encoding: The embedded input passes through multiple transformer layers, where each layer applies self-attention and feed-forward transformations:

$$H_L = \text{Transformer}_L(H_{L-1}), \quad L = 1, 2, \ldots, M \tag{2}$$

Where M is the total number of transformer layers, and $H_L \in \mathbb{R}^{n \times d}$ represents the contextualized hidden states after the L-th layer. Each transformer layer captures increasingly complex linguistic patterns and dependencies.

Classification Head: The final classification is performed using a dedicated linear classification head that operates on the [CLS] token representation:

$$\hat{y} = \text{Softmax}(W_c \cdot H_M[0] + b_c) \tag{3}$$

Where $H_M[0] \in \mathbb{R}^d$ is the [CLS] token representation after the final transformer layer, $W_c \in \mathbb{R}^{|\mathcal{Y}| \times d}$ is the classification weight matrix, $b_c \in \mathbb{R}^{|\mathcal{Y}|}$ is the classification bias vector, $\mathcal{Y} = \{E, S, G, N\}$ is the set of ESG classification labels and $\hat{y} \in \mathbb{R}^{|\mathcal{Y}|}$ represents the predicted class probability distribution.

Training Objective: The model parameters are optimized to minimize the cross-entropy loss between predicted and true labels:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{|\mathcal{Y}|} y_i \log(\hat{y}_i) \tag{4}$$

Where $\mathcal{L}_{\text{CE}}$ is the Cross-entropy loss function, y_i is the true label, $y \in \{0, 1, 2, 3\}^{|\mathcal{Y}|}$ is the encoded true label vector.

3.2. Scoring System

The dataset was constructed to support the ESG sentiment classification in Vietnamese. Each sample in the dataset consists of a Vietnamese text segment and labels corresponding to that domain, representing negative, neutral, or positive sentiment (e.g., Environmental Negative, Environmental Neutral, Environmental Positive).

To enhance analysis effectiveness and ensure accuracy during training, we divided the initial dataset into three subsets, each representing a distinct thematic class: E (Environment), S (Social), and G (Governance). After the partitioning, we trained separate models for each class to clearly distinguish between sentiments (negative, neutral, and positive) within the corresponding data. This partitioning approach allows the models to focus on and learn the unique characteristics of each class, thereby improving the accuracy and effectiveness of the analysis process. We believe that this method not only enhances the performance of the models but also increases the reliability and precision in sentiment classification for each class.

3.2.1. Scoring Model Architecture

We fine-tuned a pre-trained transformer-based language model specialized for Vietnamese text. For each ESG domain, a separate multi-label classification model was constructed, taking the preprocessed input and predicting outputs corresponding to the negative, neutral, and positive sentiment classes. The model architecture includes a standard BERT backbone followed by a dense output layer for multi-label prediction. To address class imbalance, we employed the Binary Cross-Entropy loss with optional class weights. Similar to ESG classification, due to the characteristics of the Vietnamese language, spaces are not used to separate words in compound words. We continue to use ViTokenizer in Pyvi to analyze the input text string and accurately determine word boundaries.

The initial training phase involved fine-tuning the pre-trained model exclusively on the labeled dataset of the chosen ESG domain. Hyperparameters, including learning rate, batch size, and the number of epochs were optimized through experiments and practical tuning to balance convergence and overfitting risks. Early stopping mechanisms were implemented by monitoring validation loss and macro-F1 scores on the held-out validation set.

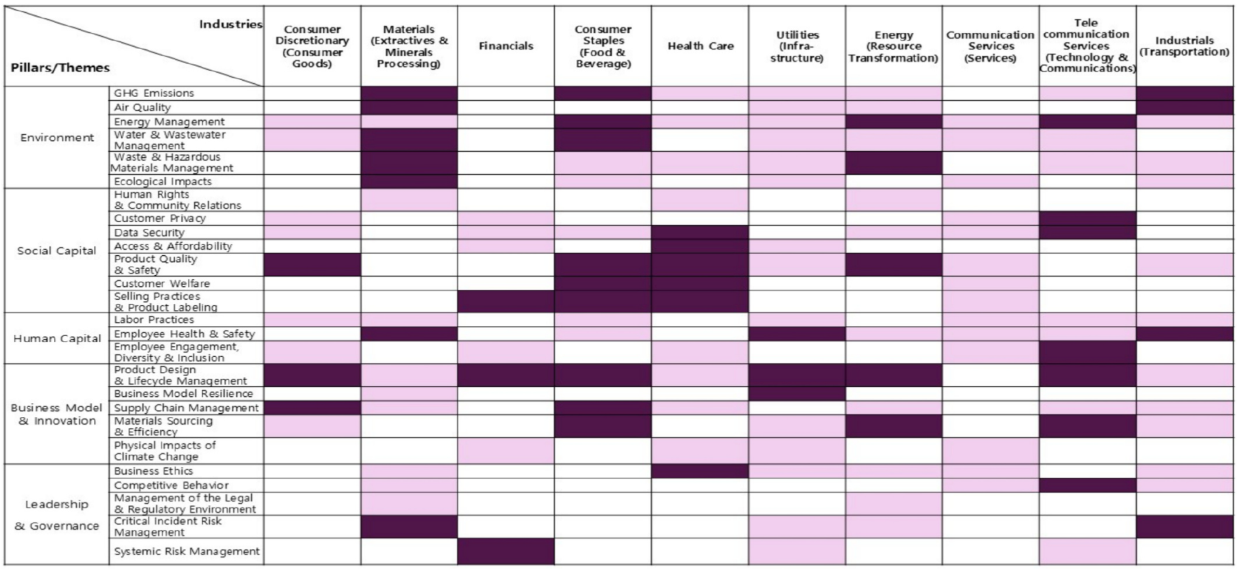
Performance metrics included precision, recall, and F1-score for each sentiment class, as well as macro, micro, and weighted averages to provide a comprehensive assessment. The impact of pseudo-labeled data was analyzed both quantitatively (through metrics) and qualitatively (via inspection of high-confidence pseudo predictions). All experiment metadata, including label splits, training logs, and detailed results, were archived to ensure reproducibility.

3.2.2. ESG Score Computation and Evaluation

Based on the method proposed in ESG-KIBERT [27], we have adopted the ESG scoring method based on industry-specific weights and sentiment analysis. The sentiment model classifies each sentence into three emotional categories: positive, neutral, or negative, according to E, S, and G. Sentiment scores are assigned as +1 (positive), 0 (neutral), and -1 (negative), creating a clear and balanced scale.

We apply this method because SASB was chosen over other frameworks like GRI and the Task Force on Climate-related Financial Disclosures (TCFD) due to its industry-specific approach, which allows for more precise customization of ESG weightings. Furthermore, the authors aligned SASB's industry classifications with MSCI's sector framework to establish a coherent analytical structure that meets financial market demands. This approach emphasizes the significance of ESG factors specific to each industry, thereby enhancing the reliability of the evaluation process.

The materiality map, illustrated in Figure 1, depicts industry sectors with varying ESG relevance. Dark shading represents sectors linked to over 50% of industries, while light



Industries Pillars/Themes		Consumer Discretionary (Consumer Goods)	Materials (Extractives & Minerals Processing)	Financials	Consumer Staples (Food & Beverage)	Health Care	Utilities (Infra-structure)	Energy (Resource Transformation)	Communication Services (Services)	Tele communication Services (Technology & Communications)	Industrials (Transportation)
Environment	GHG Emissions		■		■	▨	▨	▨		▨	■
	Air Quality		■				▨	▨			■
	Energy Management	▨	▨		■	▨	▨	■	▨	■	▨
	Water & Wastewater Management	▨	■		■		▨	▨	▨	▨	
	Waste & Hazardous Materials Management		■		▨	▨	▨	■		▨	▨
	Ecological Impacts		■		▨		▨		▨		▨
Social Capital	Human Rights & Community Relations		▨			▨		▨			
	Customer Privacy	▨		▨					▨	■	
	Data Security	▨		▨	▨	■		▨	▨	■	
	Access & Affordability			▨		■	▨				
	Product Quality & Safety	■			■	■	▨	■	▨		▨
	Customer Welfare				■	■			▨		
	Selling Practices & Product Labeling			■	■	■			▨		
Human Capital	Labor Practices	▨	▨		▨		▨		▨	▨	▨
	Employee Health & Safety		■		▨		■	▨	▨	▨	■
	Employee Engagement, Diversity & Inclusion	▨		▨		▨			▨	■	
Business Model & Innovation	Product Design & Lifecycle Management	■	▨	■	■	▨	■	■		■	▨
	Business Model Resilience		▨				■				
	Supply Chain Management	■	▨		■	▨		▨		▨	▨
	Materials Sourcing & Efficiency	▨			■		▨	■		■	▨
	Physical Impacts of Climate Change			▨		▨	▨		▨		
Leadership & Governance	Business Ethics		▨			■	▨	▨	▨		▨
	Competitive Behavior		▨						▨	■	▨
	Management of the Legal & Regulatory Environment		▨					▨			
	Critical Incident Risk Management		■				▨	▨			■
	Systemic Risk Management			■			▨			▨	

Figure 1: SASB-based materiality map for ESG factor weighting by industry.

shading indicates those linked to fewer than 50%. The authors derived quantitative values to enhance the consistency and intuitiveness of ESG evaluations. Weightings were assigned according to shading intensity: 0.25 for light purple, representing the median of the less than 50% importance range, and 0.75 for dark purple, representing the median of the greater than 50% importance range. The total weight for each industry was determined by summing the weighted counts within each pillar. The Environment pillar was categorized as "E", Social Capital and Human Capital were grouped under "S", and Business Model & Innovation along with Leadership & Governance were classified as "G".

$$ESG_{\text{company}} = W_E \cdot \sum_{i=1}^{N_E} S_{E,i} + W_S \cdot \sum_{j=1}^{N_S} S_{S,j} + W_G \cdot \sum_{k=1}^{N_G} S_{G,k} \tag{5}$$

Where W_E, W_S, W_G represent the industry-specific weight for E, S, G to which the specific company belongs and $S_{E,i}, S_{S,j}, S_{G,k}$ refer to sentiment scores assigned to each sentence labeled as E, S, G, respectively. Based on the formula provided by the authors in the paper, we have also applied the formula for calculating the ESG rating based on industry-specific weights and sentiment analysis.

3.3. Model Selection

We experiment with a suite of BERT-based models, each fine-tuned specifically for ESG classification. These models are based on established architectures known for their success in a range of NLP classification tasks. Our goal is to evaluate the effectiveness of different model bases in the Vietnamese ESG context.

The selection of BERT-based models as the foundation for our classifiers stems from their proven effectiveness across a wide spectrum of NLP tasks, including text classification, information extraction, and sentiment analysis. The Transformer architecture, upon which

BERT and its variants are built, excels at capturing long-range dependencies and contextual relationships within text, which is crucial for understanding the often subtle and complex language used in ESG disclosures.

We use a pretrained model trained on large-scale datasets to develop the ESG classifier for Vietnamese ESG model. We employed PhoBERT [35], along with experiments using other Vietnamese language models such as ViSoBERT [37], viBERT [7] and vELECTRA [7].

Table 1: Model selected in our experiment

Backbone	**# Param**	**Data Training Source**
PhoBERT (2020)	135M	Vietnamese Wikipedia (1GB), Vietnamese news (19GB)
ViSoBERT (2023)	97M	Vietnamese social networks comment (Facebook, TikTok, Youtube)
vELECTRA (2020)	110M	NewsCorpus, OscarCorpus (58.4 GB)
viBERT (2020)	115M	10GB subset data of vELECTRA
BERT-base-multilingual-uncased/cased (2017)	168/179M	BookCorpus, 104 languages of Wikipedia
DistilBERT-base-multilingual-cased (2019)	168M	BookCorpus, 104 languages of Wikipedia
RoBERTa-base XML-RoBERTa-base (2019)	124/279M	BookCorpus, Wikipedia, OpenWebText, Stories, CC-News(160GB)
DeBERTa-v3-small/base (2022)	141/184M	BookCorpus, Wikipedia, OpenWebText, Stories (78GB)
FinBERT (2022)	109M	Fine-tuning BERT on TRC2-financial

For multi-language model setting, we specifically chose to experiment with a suite of models, including BERT-base-cased, BERT-base-uncased, RoBERTa-base, XLM-RoBERTa-base, DeBERTa-v3-base, DeBERTa-v3-small, FinBERT and DistilBERT-base-cased, to provide a comparative analysis. This allows us to assess not only the general applicability of transformer models but also the relative strengths of different architectural variations and pre-training strategies when applied to the specific domain of Vietnamese ESG text. The decision to fine-tune these models, rather than training from scratch, is motivated by the need to leverage the extensive linguistic knowledge already encoded within them, thereby mitigating the data requirements and computational costs associated with training large models on a limited domain-specific dataset. Models detailed in Table 1.

4. Dataset

In this section, we present the datasets used for developing and evaluating our ESG classification models. Effective domain-specific text classification requires high-quality labeled data that captures the linguistic characteristics and thematic diversity of the target domain. We systematically collected ESG-related content from corporate sustainability reports and financial news sources, followed by comprehensive preprocessing, manual annotation and many annotation techniques to create reliable ground truth labels for training and evaluation purposes. Our dataset can access at https://huggingface.co/datasets/nguyen599/ViEn-ESG-100.

Table 2: Overview of datasets utilized for ESG classification and sentiment analysis.

Data type	Language	No. Sentence	Source	Labeling techique
ESG classification	English	45,942	Corporate Reports, News	Manual label
ESG classification	Vietnamese	60,222	Corporate Reports, News	Manual label
ESG classification	Vietnamese	14,634	Corporate Reports, News	Pseudo label
ESG classification	Vietnamese	10,000	English data	Translate
ESG sentiment	Vietnamese	6,430	Corporate Reports	Manual

4.1. Dataset Description

A cornerstone of this research is the development and utilization of the **ViEn-ESG dataset**, a novel, large-scale, bilingual corpus specifically curated for ESG classification tasks in the English and Vietnamese languages. The creation of this dataset was motivated by the significant scarcity of publicly available, high-quality labeled data for ESG analysis, particularly for low-resource languages such as Vietnamese. Addressing this gap is crucial for advancing NLP applications in sustainable finance and corporate responsibility assessment within emerging markets. ViEn-ESG comprises a total of 130,798 sentence-level samples, meticulously collected and annotated to serve as a robust foundation for training and evaluating ESG classification models. In addition, we selected a subset of 6,430 sentences in the ViEn-ESG dataset and created a segment label that includes Negative, Neutral, or Positive sentiment to capture sentiment signals associated with ESG-related statements. This sentiment-labeled subset plays a critical role in training the sentiment analysis component of our framework, which is later integrated into the ESG scoring mechanism. By leveraging these sentiment annotations, we enable the model to compute fine-grained ESG scores by aggregating sentiment distributions across the Environmental, Social, and Governance dimensions. This design ensures that the scoring process is both interpretable and quantitatively grounded, providing a practical tool for assessing corporate sustainability beyond categorical classification.

4.2. Data Collection

Corporate sustainability reports represent comprehensive disclosures of organizational performance across environmental, social, and governance dimensions, serving as primary vehicles for communicating non-financial achievements and ESG commitments to diverse stakeholder groups. These documents have gained prominence as regulatory frameworks increasingly mandate ESG transparency and investors demand greater accountability regarding sustainable business practices [1]. The growing emphasis on ESG disclosure has led companies across various industries to publish detailed sustainability reports that document their environmental stewardship, social responsibility initiatives, and governance structures.

The ViEn-ESG dataset draws from two primary data sources: ESG news articles and official corporate sustainability disclosures. For news articles content, we systematically gathered ESG-focused articles published between January 1, 2015, and August 14, 2024, sourcing

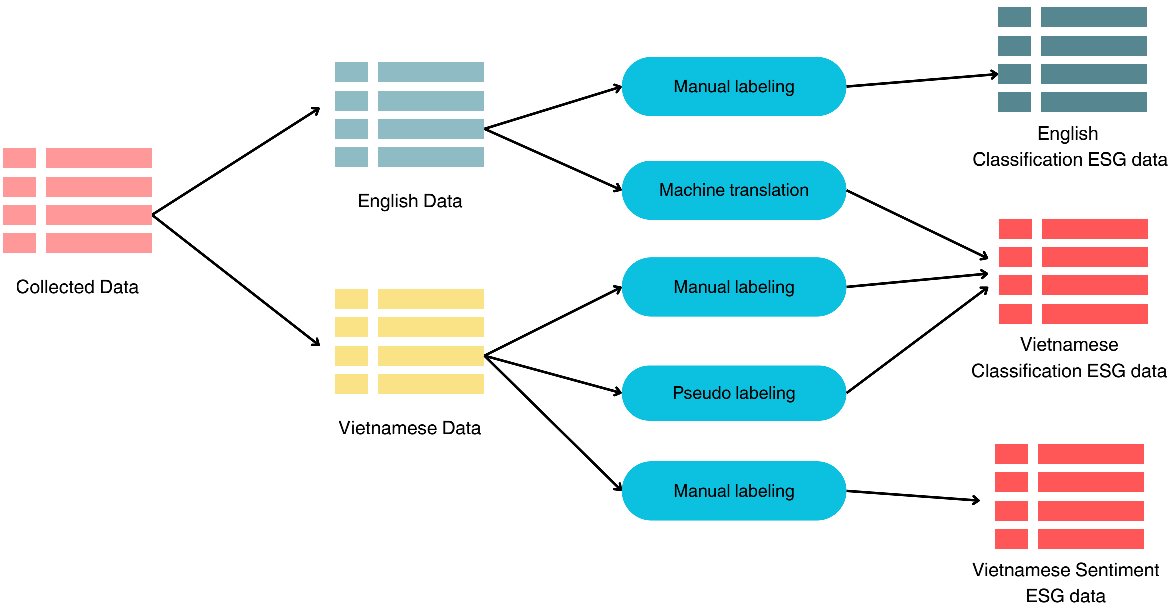


Figure 2: Overview labeling data pipeline

English-language content from ESGToday and Vietnamese articles from reputable news sites such as VnExpress and Vietnam News. To ensure thematic relevance, we selected articles that were explicitly tagged with "ESG" or categorized under ESG-related sections, guaranteeing that all collected content directly addressed environmental, social, or governance themes.

The corporate disclosure component involved a collection of sustainability reports from major Vietnamese enterprises listed in the Ho Chi Minh Stock Exchange (HOSE)[15] and Hanoi Stock Exchange (HNX)[14], with English reports collected from Standard and Poor's 500 (S&P 500) 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.

4.3. Data Preprocessing

We collected 348 sustainability reports, extracted 65,103 curated textual segments. With the news dataset, we collected a total of 68,553 sentences, the comprehensive dataset reached 130,798 textual entries, providing extensive coverage of ESG-related corporate communications across multiple formats and contexts. The collected sustainability reports, predominantly distributed in PDF format, contained much unstructured content. We decided to keep only text and drop tabular and visual elements such as images or charts. An overview of the distribution of sustainability reports by category and language is illustrated in Figure 3, highlighting the diversity of sectors and the bilingual nature of the collected data.

In the preprocessing step, we utilized the spaCy [49] natural language processing library to perform sentence-level segmentation of the extracted textual content, enabling precise

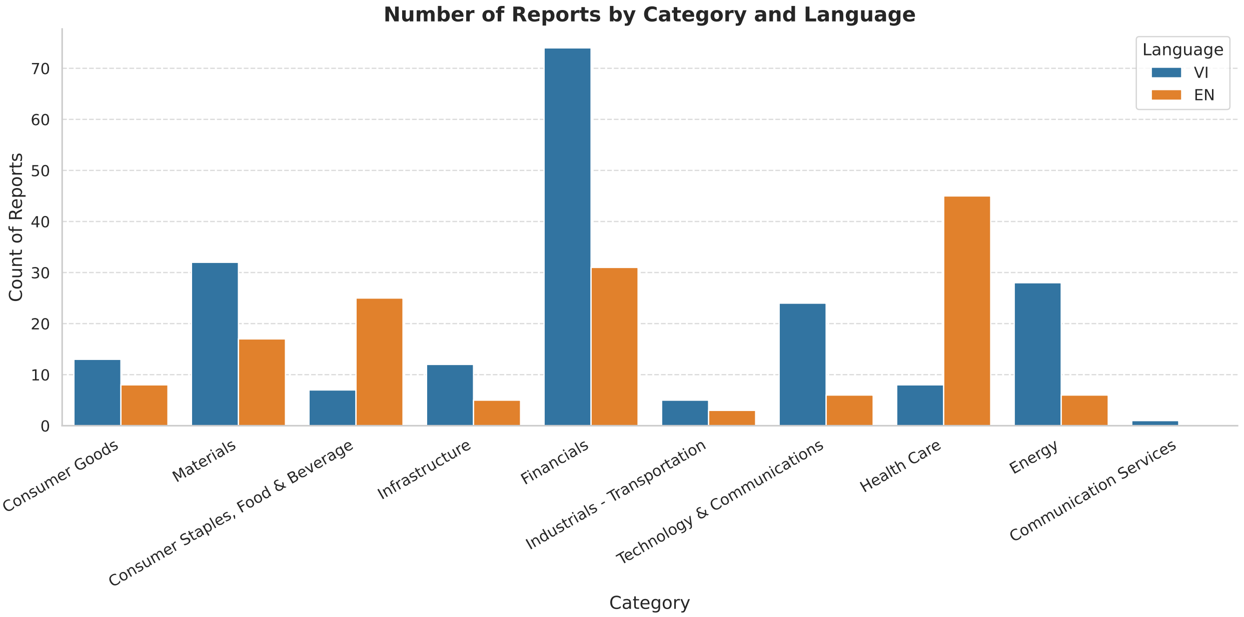


Figure 3: Distribution of Sustainability Reports by Category and Language

annotation at the sentence granularity level. The preprocessing pipeline incorporated several data normalization procedures including standardization of numerical formats, systematic removal of non-textual characters and symbols, and elimination of duplicate sentences. Post-processing measures included a comprehensive manual review to identify and correct automatic segmentation errors, inappropriate sentence boundaries, and content fragments that were either excessively brief or excessively long, with problematic segments being either manually corrected through proper segmentation or removed entirely from the dataset.

4.4. Annotation Methodology and Quality Assurance

4.4.1. Manual Labeling

The development of our ESG classification system required establishing a robust theoretical foundation grounded in internationally recognized ESG evaluation standards. We drew upon the methodological frameworks employed by prominent global ESG rating organizations, such as Morgan Stanley Capital International (MSCI), Sustainalytics, and the Carbon Disclosure Project (CDP). Labeling pipeline shown in the Figure 4, a team of four domain experts, each possessing specialized knowledge in ESG assessment methodologies and familiar with the evaluation standards of these institutions, collaboratively developed and refined our annotation guidelines through iterative discussion and consensus-building processes. The labeling rules for ESG classification are detailed in Table 3. In the Table 4, we show labeling rules for ESG sentiment.

To ensure the highest quality and reliability of our dataset annotations, we implemented a rigorous two-stage labeling methodology involving four annotators who were thoroughly trained on the ESG classification guidelines and possessed extensive domain knowledge. During the initial phase, annotators independently classified sentences that exhibited clear categorical distinctions, focusing on unambiguous examples that could be readily assigned

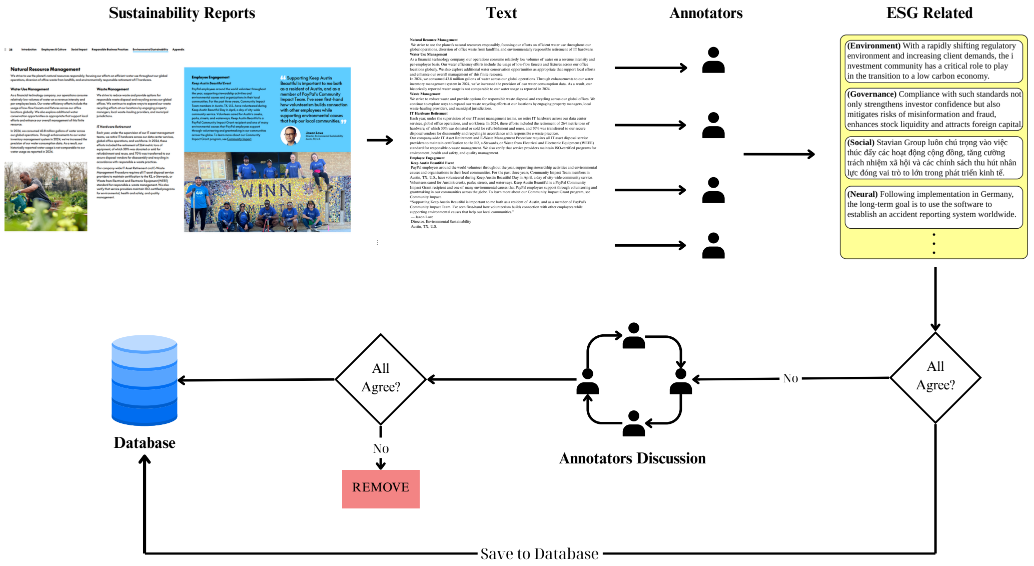

Sustainability Reports
Text
Annotators
ESG Related
(Environment) With a rapidly shifting regulatory environment and increasing client demands, the i nvestment community has a critical role to play in the transition to a low carbon economy.
(Governance) Compliance with such standards not only strengthens investor confidence but also mitigates risks of misinformation and fraud, enhances stock liquidity and attracts foreign capital.
(Social) Stavian Group luôn chú trọng vào việc thúc đẩy các hoạt động cộng đồng, tăng cường trách nhiệm xã hội và các chính sách thu hút nhân lực đóng vai trò to lớn trong phát triển kinh tế.
(Neural) Following implementation in Germany, the long-term goal is to use the software to establish an accident reporting system worldwide.
All Agree?
No
Annotators Discussion
All Agree?
No
REMOVE
Database
Save to Database

Figure 4: Manual labeling data pipeline

to Environmental, Social, Governance, or Non-ESG categories. The second phase involved collaborative deliberation sessions where annotators engaged in structured discussions to resolve challenging cases that presented classification ambiguities or overlapping characteristics across multiple ESG dimensions. To maintain stringent quality control standards, we established a consensus-based decision framework whereby sentences were included in the final dataset only when all four annotators reached unanimous agreement, thereby ensuring robust inter-annotator reliability while minimizing subjective bias.

Another cases were systematically excluded from the dataset to preserve annotation quality, resulting in the removal of less than 0.2% of the total sentences due to persistent disagreement among annotators. Our the inter-annotator agreement was quantitatively assessed using Cohen's kappa coefficient [9], which demonstrated exceptionally high consensus levels across all categories and languages, with scores ranging from 0.974 to 0.990 as detailed in Table 5, indicating almost perfect agreement according to established interpretation guidelines and validating the consistency and reliability of our annotation process.

The labeling schedule of four annotators involved two-hour daily sessions conducted five days per week, with annotators achieving a productivity rate of 600-700 sentences per hour, translating to approximately 7,000 sentences labeled per annotator per week. The comprehensive annotation timeline spanned approximately 3-4 months, with the initial three months period dedicated to processing clearly classifiable sentences, followed by an additional month of intensive phase focused on collaborative discussion and resolution of challenging cases, ensuring thorough coverage and meticulous attention to quality throughout the entire dataset construction process.

This methodologically rigorous manual labeling approach, which mirrors best practices established in other domain-specific dataset creation studies, guarantees high-quality supervised

Table 3: Labeling rules for the ViEn-ESG dataset

Class	Rule
E	Environmental factors include the reduction of hazardous substances, eco-friendly management, climate change, carbon emission, natural resources, pollution and waste, environmental performance, resource use, product innovation, and energy conservation.
S	Social factors include human capital, product responsibilities, workers, working environment, partners and competitors, consumers, community contribution, human rights, gender and diversity, labor standards, customer satisfaction, information protection, and privacy.
G	Governance factors include corporate behavior, shareholder rights, board of directors, auditing bodies, disclosures, stakeholders, CSR strategy, executive dividends, ethical management, laws and taxes.
N	Neutral contains ESG information as a neutral element, excluded from the environmental, social, and governance categories, or includes two categories of in one sentence, such as environmental, social, or social, governance.

Table 4: Labeling rules for the sentiment dataset.

Class	Rule
Environment	• Positive: Sustainable practices, green innovations, renewable energy, carbon reduction. • Neutral: Descriptive or factual reporting without clear evaluative tone. • Negative: Environmental violations, pollution, unsustainable activities, deforestation.
Social	• Positive: Initiatives improving social well-being, equity, diversity, community support. • Neutral: Objective descriptions or informational reporting on social issues. • Negative: Labor disputes, discrimination, social harm, human rights violations.
Governance	• Positive: Strong leadership, transparent policies, effective governance, compliance with regulations. • Neutral: Factual updates on governance practices without clear sentiment. • Negative: Scandals, corruption, fraud, governance failures, lack of transparency.

data that accurately reflects ESG principles and linguistic nuances across both English and Vietnamese contexts, ultimately providing a solid foundation for training robust and reliable ESG classification models, with representative examples of sentences from each class in the ViEn-ESG dataset presented in Table 6. In Table 7, we show the class-wise distribution of the ViEn-ESG dataset.

4.4.2. Pseudo-labeling for Data Enrichment

We also apply pseudo-labeling for data enrichment. The model initially trained on labeled data was used to generate probabilistic predictions for an external corpus of unlabeled texts. Only samples for which the model assigned a highest class probability exceeding a predefined confidence threshold of 0.94 were retained. For each retained text, pseudo-labels were assigned to the corresponding classes based on these confidence scores. This process optionally employed an iterative scheme, gradually adjusting the threshold across multiple rounds to maximize both the quantity and quality of automatically labeled samples.

Table 5: Cohen's Kappa coefficients measuring inter-annotator agreement across ESG categories and languages in the ViEn-ESG dataset.

Category	**English**	**Vietnamese**	**All**
Environment	0.987	0.979	0.983
Social	0.984	0.982	0.983
Governance	0.990	0.989	0.989
Neural	0.984	0.974	0.979
All	0.986	0.981	0.984

Table 6: Sample labeled data in the ViEn-ESG dataset.

Class	**Rule**
E	• Known reserves of oil, gas and coal can be extracted for some decades into the future, in a cost-efficient manner using current technologies, at current energy prices. • this notably good development consequence 25% decrease waste api processes fermion. • Greenhouse gas emissions were down meet the reduction target (15% below fiscal 2013 level).
S	• Economies need educated and skilled workforces to ensure their economic health. • One of the focuses of Servicio País is to revive public spaces near schools and neighborhoods in order to improve the quality of life of residents. • Moreover, onsite healthcare opportunities for employees are included in this program, including access to a blood pressure machine, a mammography van, blood drives, a flu shot clinic, biometric screening and an annual onsite visit from a wellness doctor.
G	• Senior management reviews engagement system results annually. • The Audit Committee provides oversight with respect to financial reporting and legal risk. • The system of internal accounting control is supported by written policies and guidelines, the selection and training of qualified employees, an organizational structure that provides an appropriate division of responsibility and a program of internal auditing.
N	• About Evercomm Evercomm is a software company supporting businesses to net zero. • As of the date of this filing, we are not aware of any matters that are required to be disclosed pursuant to this standard. • An increase of $due to industry-wide price escalation and normal-course salary increases for on-site property management personnel.

The resulting pseudo-labeled data was then combined with the original training set. The model was then retrained on this composite dataset, leveraging both human-annotated and high-confidence machine-labeled samples.

In traditional pseudo-labeling methods, using a single model to predict labels for unlabeled data often carries significant risks. Specifically, if the model has high bias or suffers from overfitting on the initial training set, the pseudo-labels generated may contain systematic errors. These errors can accumulate when incorrectly labeled data is fed back into the training set, leading to a phenomenon known as confirmation bias—where the model increasingly reinforces its incorrect predictions.

To minimize problems due to pseudo-labeling, we propose a method that integrates ensemble learning into the pseudo-labeling process. This semi-supervised labeling approach leverages an ensemble of multiple BERT models. First, the labeled training data is used to train K

Table 7: Class-wise distribution of ViEn-ESG Dataset.

Data	**No. of Sentences**		**Total**
	English	Vietnamese	
Environment	10,517	19,080	29,597
Social	11,112	25,201	36,313
Government	9,103	18,046	27,149
Neural	15,210	22,529	37,739
Total	45,942	84,856	130,798

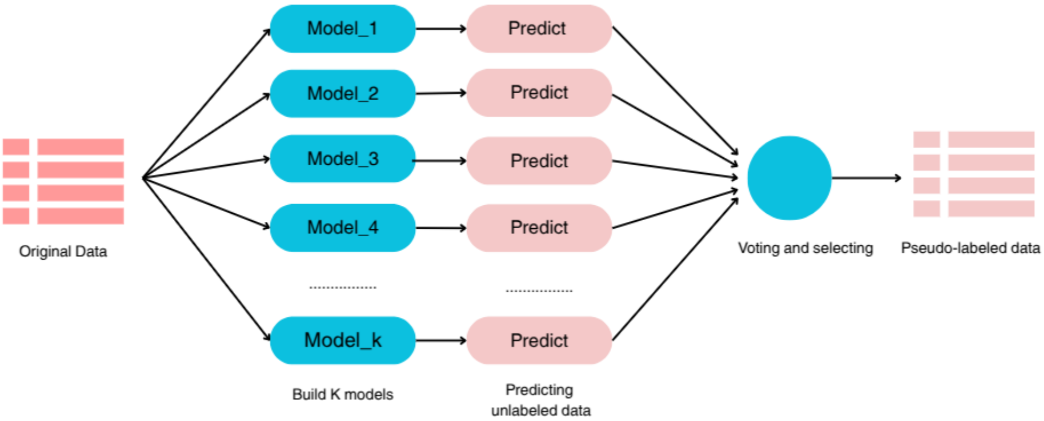


Figure 5: BERT Ensemble for Reliable Pseudo-Labels

independent BERT models, each trained on a different subset of the original data. After the training process is complete, all of these BERT models simultaneously predict labels for the unlabeled data. The predicted labels from these models are then aggregated using a voting mechanism, where the labels with the highest confidence are selected as the pseudo-labels. Specifically, labels that receive high consensus among the models are prioritized. Finally, the data with these pseudo-labels is added back into the training set, thereby improving the model's quality by effectively utilizing both labeled and unlabeled data. From the original train dataset of 27,366 sentences, we generated an additional 14,634 sentences using the unlabeled data by pseudo-labeling. This approach takes advantage of the power of BERT models to minimize the risk of mislabeling, reduce confirmation bias, and enhance the stability of the expanded label set.

The selected models include both multilingual and Vietnamese-optimized transformers: DistilBERT-base-multilingual, ViSoBERT, Vi-Electra, BERT-base-multilingual, PhoBERT, ViDeBERTa, RoBERTa-base. Each training iteration utilizes pseudo-data generated from the previous round to supplement the original training dataset, creating an iterative improvement process.

1) First Iteration: In the initial iteration, models were trained on the 7 subset dataset of the original data and demonstrated promising baseline performance. DistilBERT-base-multilingual achieved the highest performance with 87.36% accuracy, followed by ViSoBERT at 87.23% and Vi-Electra at 86.86%.

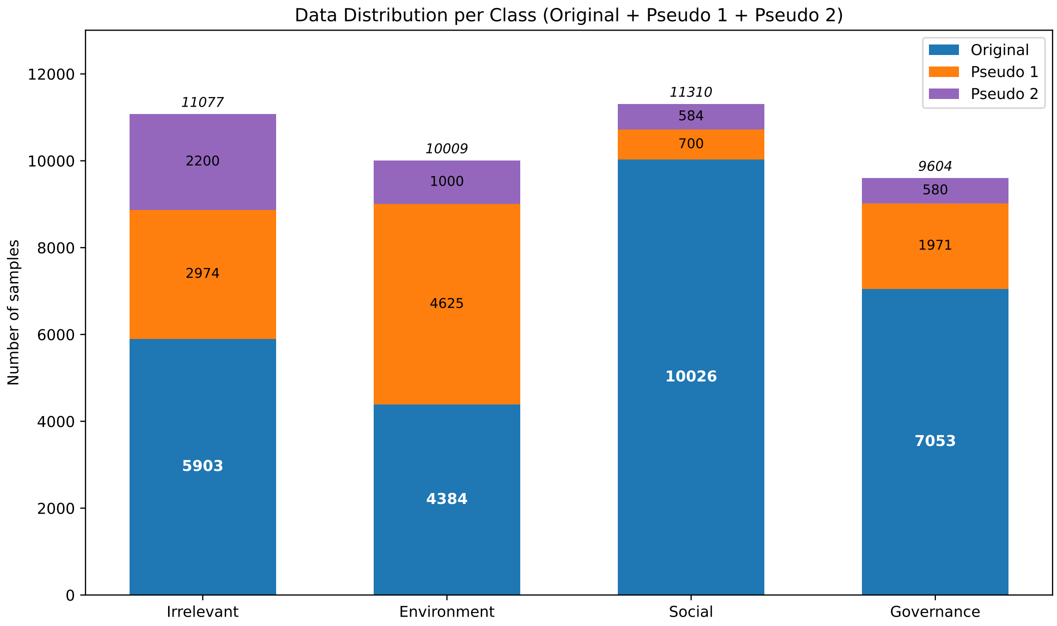


Figure 6: Data distribution by each pseudo labeling iteration

The first iteration generated substantial pseudo-data with 8,830 samples for Environment, 7,667 for Social, 6,300 for Governance, and 2,974 for Irrelevant classes. To ensure dataset balance, we adjusted the pseudo-data quantities used, resulting in a total of 37,636 samples for the subsequent iteration with the following distribution: Irrelevant (8,877 total: 5,903 original + 2,974 pseudo), Environment (9,009 total: 4,384 original + 4,625 pseudo), Social (10,726 total: 10,026 original + 700 pseudo), and Governance (9,024 total: 7,053 original + 1,971 pseudo).

2) Second Iteration: The second iteration showed significant performance improvements across all models. Vi-Electra achieved the best overall performance with 92.13% accuracy, accompanied by impressive F1 scores across all classes. ViSoBERT and ViDeBERTa also demonstrated strong performance with accuracies of 91.33% and 91.13%, respectively.

The second iteration generated 8,255 new pseudo-data samples, with Environment contributing 4,212 samples, Social 1,915 samples, Governance 1,548 samples, and Irrelevant 580 samples. After data balancing, the total sample count for the final iteration reached 42,000 samples with the following distribution: Irrelevant (11,077 total: 8,877 original + 2,200 pseudo), Environment (10,009 total: 9,009 original + 1,000 pseudo), Social (11,310 total: 10,726 original + 584 pseudo), and Governance (9,604 total: 9,024 original + 580 pseudo).

3) Third Iteration: The final iteration marked the peak performance of our research. Vi-Electra continued to lead with 94.66% accuracy, followed by BERT-base-multilingual at 94.36% and ViSoBERT at 93.83%. Notably, all models achieved accuracy above 91%, demonstrating the stability and effectiveness of the pseudo-labeling methodology.

The research results demonstrate a consistent improvement trend across pseudo-labeling iterations. The average accuracy of models increased from 85.75% in the first iteration to 90.82% in the second iteration, reaching a peak of 93.47% in the final iteration. Particularly,

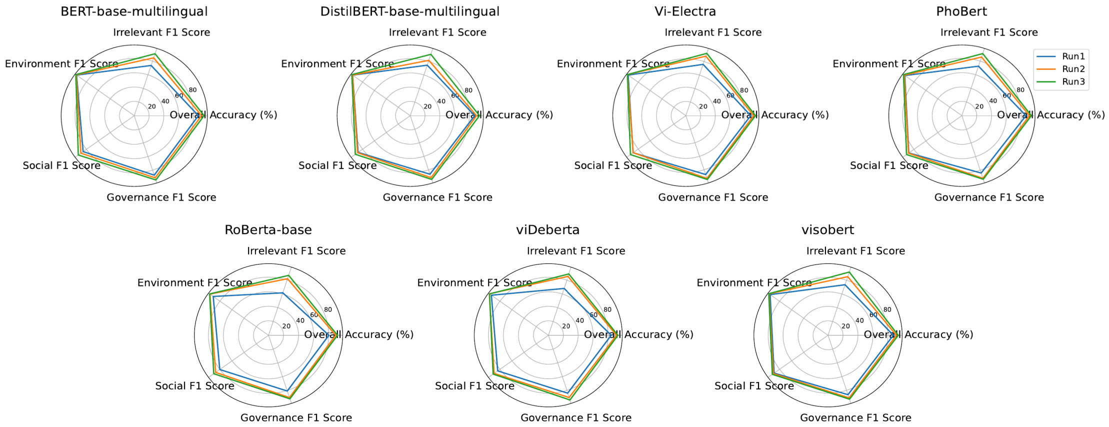


Figure 7: Training Results following each pseudo labeling stage

Vi-Electra showed consistent superiority across all three iterations, indicating excellent adaptability to Vietnamese text data.

The Environment class consistently achieved the highest F1 scores across all models, while the Irrelevant class presented the most classification challenges. This can be attributed to the diverse and ambiguous nature of the Irrelevant class compared to the more specialized domain-specific classes. The pseudo-labeling approach proved particularly effective in enriching data and addressing class imbalance issues. By strategically selecting pseudo-labeled samples, we were able to maintain better class distribution while progressively increasing the dataset size, contributing to the observed performance improvements.

4.4.3. Annotation through Machine Translation

To expand the coverage of our Vietnamese dataset, we adopted a translation-based labeling strategy that leverages existing high-quality English data. Specifically, we selected 10,000 English sentences that had already been annotated with gold-standard labels. These sentences were then translated into Vietnamese using the `erax-ai/EraX-Translator-V1.0` model [36], a neural machine translation system that has been shown to achieve strong performance on English-Vietnamese translation tasks. By preserving the semantic structure of the original sentences, this approach allows us to inherit the English labels and directly assign them to their translated Vietnamese counterparts.

The main advantage of this method is that it enables rapid and cost-efficient creation of labeled data in a low-resource language setting, without the need for extensive manual annotation. Furthermore, because the labels are derived from semantically equivalent English sentences, this process ensures alignment across languages and facilitates the development of cross-lingual benchmarks. To mitigate the risk of translation errors, we applied post-translation validation by sampling a subset of the Vietnamese outputs and verifying whether the translated sentence preserved the original meaning and label consistency. In practice, we observed that label preservation remained reliable for well-structured sentences, though some minor noise could occur due to idiomatic expressions or domain-specific terminology.

Despite these limitations, translation-based labeling serves as an effective intermediate solution between manual annotation and pseudo-labeling. It provides a scalable mechanism to enrich the dataset with diverse Vietnamese examples while maintaining semantic fidelity to the original English data. The resulting bilingual resource not only strengthens the robustness of our model training but also contributes toward the long-term goal of building more inclusive ESG datasets that support multilingual applications. The detailed prompt used to guide the translation process is provided in the Appendix A.

5. Experiments and Analysis

This section details the experimental design and results of our proposed ESG classification model. We outline the implementation settings, including hardware specifications, training configurations, and evaluation protocols. Subsequently, we present a comprehensive performance analysis on both English and Vietnamese subsets of the ViEn-ESG dataset, along with an assessment of the bilingual models and the ESG scoring mechanism.

5.1. Experimental Setup and Implementation Details

To comprehensively address the ESG classification task, we adopted a multi-faceted modeling strategy, training four distinct types of classifiers based on the fine-tuned BERT architecture. Three binary classifiers were developed, each dedicated to identifying the relevance of a sentence to one specific ESG pillar: Environmental (E), Social (S), or Governance (G). This binary approach allows for focused assessment within each dimension and aligns with methodologies used in prior literature for relevance or impact identification. Complementing these, we developed a multi-class classifier, termed ViBERT-ESG (and its variants based on other backbones), designed to categorize a sentence into one of four mutually exclusive classes: E, S, G, or Neutral (N). The Neutral class serves as a crucial catch-all category for sentences that are either irrelevant to ESG or potentially span multiple dimensions ambiguously. This four-class model provides a holistic classification across the primary ESG pillars.

1) Implementation Details: The experiments were conducted with the following specifications. For all work, the authors utilize an Intel Core i7–13700 processor, 48 GB of RAM, and 2 RTX 3060 GPUs for training, testing, and evaluation of the models. The dataset was split into 70% training, 15% validation, and 15% test data. Key hyperparameters were kept consistent across models to ensure fair comparison, based on common practices for fine-tuning BERT-like models. The training process employs full-parameter fine-tuning with DeepSpeed ZeRO-3 optimization [42]. All models are trained with BFloat16 precision for 3000 steps, including 200 warm-up steps. We use a learning rate of 5×10^{-5} with cosine learning rate scheduling, weight decay of 0.01, a batch size of 48 and a maximum gradient norm of 2.5. The optimizer is AdamW [31], and the maximum input sequence length is set to 512 tokens. For reproducibility, we set the random seed to 42 in all experiments.

2) Evaluation Metrics: In our experimental setup, we primarily report the macro-averaged F1 score ($F1_{\text{macro}}$) as it provides a balanced view of performance across classes with varying frequencies. We also include accuracy as a complementary metric to capture the overall correctness of the prediction. This combination ensures that both class-level balance and overall classification performance are properly assessed.

$$F1_{\text{macro}} = \frac{\sum_{i=0}^{N} F1_i}{N} \tag{6}$$

where $F1_i$ denotes the F1-score of class i, and N is the total number of classes. We report macro-F1 as the primary metric since it equally weights all classes, making it particularly suitable for imbalanced ESG datasets. In addition, we also provide accuracy to offer a more comprehensive view of model performance.

5.2. Performance on Vietnamese ESG Classification

We applied a fine-tuning method to the model using sentence and ESG label pairs, utilizing a pretrained language model. In the experiment, we used multilingual BERT models, the multilingual version of BERT, and Vietnamese models such as PhoBert-Base, PhoBert-Large, viBert, viELECTRA, and ViSoBERT.

The results of multi-class classification for ESG-related sentences with 4 labels in the ESG dataset are presented in Table 8. In the experiment, each model was trained 5 times, and the average accuracy and accuracy for each class were calculated by excluding the highest and lowest accuracy values.

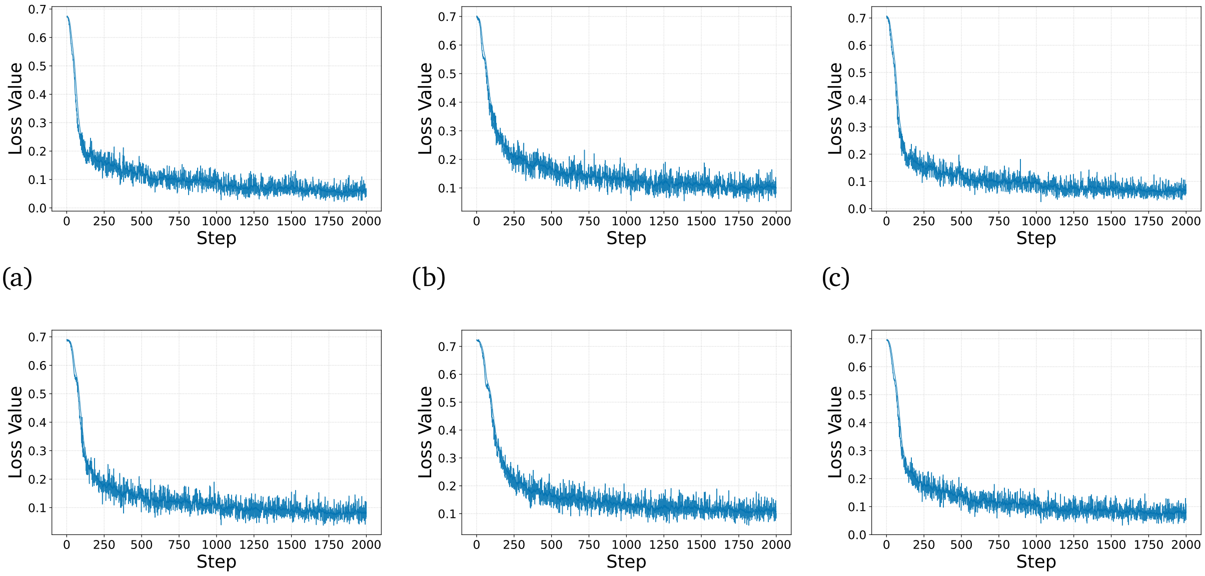
Table 8: Comparison of Transformer-based models for Vietnamese language

Model	Overall accuracy	Env F1 score	Gov F1 score	Soc F1 score	Neu F1 score
viDEBERTA	93.1	97.15	91.17	92.22	89.1
viELECTRA	94.66	98.30	93.13	94.36	92.09
PhoBERT-Base	93.7	97.42	92.29	93.25	91.04
ViSoBERT	93.83	97.83	91.94	92.94	91.93
BERT-Base-Multilingual	94.36	97.61	93.49	94.49	90.95
RoBERTa	91.8	96.92	90.12	92.08	86.79
DistilBERT-Multilingual	93.86	98.01	92.13	94.03	90.43

The comparative analysis of Transformer-based models for Vietnamese language processing reveals viELECTRA as the top-performing model, achieving superior results across all evaluation metrics. With an overall accuracy of 94.66%, viELECTRA demonstrates exceptional performance in all classification categories: Environment (98.30% F1), Governance (93.13% F1), Social (94.36% F1), and Irrelevant (92.09% F1). While other models show competitive results - ViSoBERT (93.83%) and DistilBERT-Multilingual (93.86%) - none surpass the balanced performance of PhoBERT-Base (93.7% accuracy). The evaluation also highlights the relatively weaker performance of viDEBERTA (93.1%) and BERT-Base-Multilingual (94.36%), with RoBERTa showing the lowest overall accuracy (91.8%), particularly struggling with Irrelevant content classification (86.79% F1). These findings establish viELECTRA as the most robust model for Vietnamese language processing tasks, especially for nuanced ESG (Environmental, Social, and Governance) text classification. viELECTRA is the optimal choice for classifying ESG information in Vietnamese, with outstanding performance across key classification categories.

5.3. Performance on Bilingual ESG Classification

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.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 8: Training loss of ESG classification results on bilingual dataset. That denote (a) bert-base-multilingual-cased (b) roberta-base (c) xlm-roberta-base (d) deberta-v3-base (e) deberta-v3-small (f) distilbert-base-multilingual-cased.

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 dataset. This bilingual training setup allows the models to generalize better across languages and benefit from the shared semantic space of ESG-related concepts. We evaluate our fine-tuned models on both English and Vietnamese test sets.

Table 9 summarizes the performance of existing English-based ESG models alongside our bilingual fine-tuned models on the English portion of the ViEn-ESG dataset. Among the baseline models, FinBERT-ESG and its 9-class variant exhibit strong performance, with overall F1 scores of 87.51% and 89.80%, respectively, reflecting their domain-specific pretraining on financial text. However, models such as SEC-BERT-ft and ESGify show significantly lower scores (68.12% and 48.33%, respectively), indicating that fine-grained classification schemes or excessive label granularity can reduce effectiveness when labels are mapped into a unified 4-class format (E/S/G/N). Similarly, the binary models (EnvironmentBERT, SocialBERT, and GovernanceBERT) perform well on their respective domains but lack comprehensive coverage across all ESG categories.

In contrast, our bilingual fine-tuned models consistently outperform existing baselines across all ESG dimensions. Notably, XLMRoBERTa-ESG-base achieves the highest overall F1 score

Table 9: F1 scores of models on each ESG category in the English ViEn-ESG dataset.

Model	**# Param**	**Env**	**Soc**	**Gov**	**Neu**	**Overall**
SEC-BERT-ft	109M	83.12	66.77	66.53	60.30	68.12
FinBERT-ESG	109M	92.67	84.90	86.25	87.26	87.51
FinBERT-ESG-9-class	109M	92.16	89.01	91.35	86.89	89.80
ESGify	109M	67.72	30.20	50.76	43.44	48.33
EnvironmentBERT	82M	92.15	-	-	92.76	-
SocialBERT	82M	-	76.81	-	81.23	-
GovernanceBERT	82M	-	-	64.46	80.06	-
BERT-base-multilingual-cased	168M	93.76	94.53	94.98	**94.15**	94.75
BERT-base-multilingual-uncased	168M	94.62	93.81	94.26	92.13	93.83
RoBERTa-base	124M	95.43	94.06	95.01	91.32	94.11
XLMRoBERTa-base	278M	95.00	95.00	**95.47**	92.19	**94.83**
DeBERTa-base	184M	**95.50**	94.49	94.81	91.48	94.70
DeBERTa-small	141M	94.55	94.85	94.58	90.19	93.72
DistilBERT-multilingual-cased	135M	95.15	**95.19**	94.33	91.75	94.60
FinBERT	109M	94.62	93.16	94.10	92.13	93.50

(94.83%), closely followed by `DeBERTa-ESG-base` (94.70%) and `BERT-multilingual-cased-ESG` (94.75%). These results underscore the advantages of leveraging multilingual architectures and joint bilingual fine-tuning, which enhances generalization and mitigates domain shift between English and Vietnamese data. Interestingly, while `DeBERTa-ESG-base` slightly outperforms others on the Environmental category (95.50%), `DistilBERT-multilingual-cased-ESG` achieves the best score on Social (95.19%), suggesting that certain architectures may capture different ESG dimensions more effectively.

Table 10: F1 scores of models on each ESG category in the Vietnamese ViEn-ESG dataset.

Model	**# Param**	**Env**	**Soc**	**Gov**	**Neu**	**Overall**
BERT-base-multilingual-cased	168M	93.50	89.73	91.77	**91.78**	91.80
BERT-base-multilingual-uncased	168M	80.18	58.36	68.66	57.44	66.54
RoBERTa	124M	93.41	91.49	89.93	84.32	89.96
XLMRoBERTa	278M	93.45	91.02	91.69	90.41	**91.94**
DeBERTa-base	184M	**95.24**	89.36	**93.18**	85.23	90.89
DeBERTa-small	141M	92.90	87.79	90.63	81.48	88.70
DistilBERT-multilingual-cased	135M	93.87	**91.98**	90.63	87.17	91.02
FinBERT	109M	75.28	54.02	68.21	56.91	63.70

Turning to the Vietnamese results in Table 10, we observe a slight performance drop compared to English, which is expected given the lower availability of pretraining resources for Vietnamese. Nevertheless, all bilingual models maintain high overall F1 scores above 88%, demonstrating robust cross-lingual transfer. `BERT-base-multilingual-cased` and `XLMRoBERTa` achieve the highest overall scores (91.80% and 91.94%, respectively), confirming their strong multilingual capability. `DeBERTa-base` shows superior performance on the Environmental (95.24%) and Governance (93.18%) categories, while `DistilBERT-multilingual-cased` excels on Social (91.98%), again highlighting the variation in backbone strengths across ESG dimensions. We have a significant performance drop in `BERT-base-multilingual-uncased` and `FinBERT` that will be discussed in subsection 5.4.

Overall, these results validate the effectiveness of joint bilingual training and the use of large multilingual pre-trained models for ESG classification in low-resource languages. The relatively small gap between English and Vietnamese performance further demonstrates that our approach successfully narrows the linguistic disparity in ESG analytics. These models establish a new multilingual benchmark for ESG classification in both English and Vietnamese contexts.

5.4. The Impact of ‘uncased‘ Models on Vietnamese Performance

The results presented in Table 10 reveal a notable performance degradation for uncased models such as `BERT-base-multilingual-uncased` and `FinBERT` when applied to the Vietnamese dataset. This observation is further illustrated in Figure 9, which shows the training loss curves of three representative models. Both uncased models exhibit significantly higher loss and slower convergence compared to their cased counterparts, underscoring their inability to effectively capture Vietnamese linguistic characteristics.

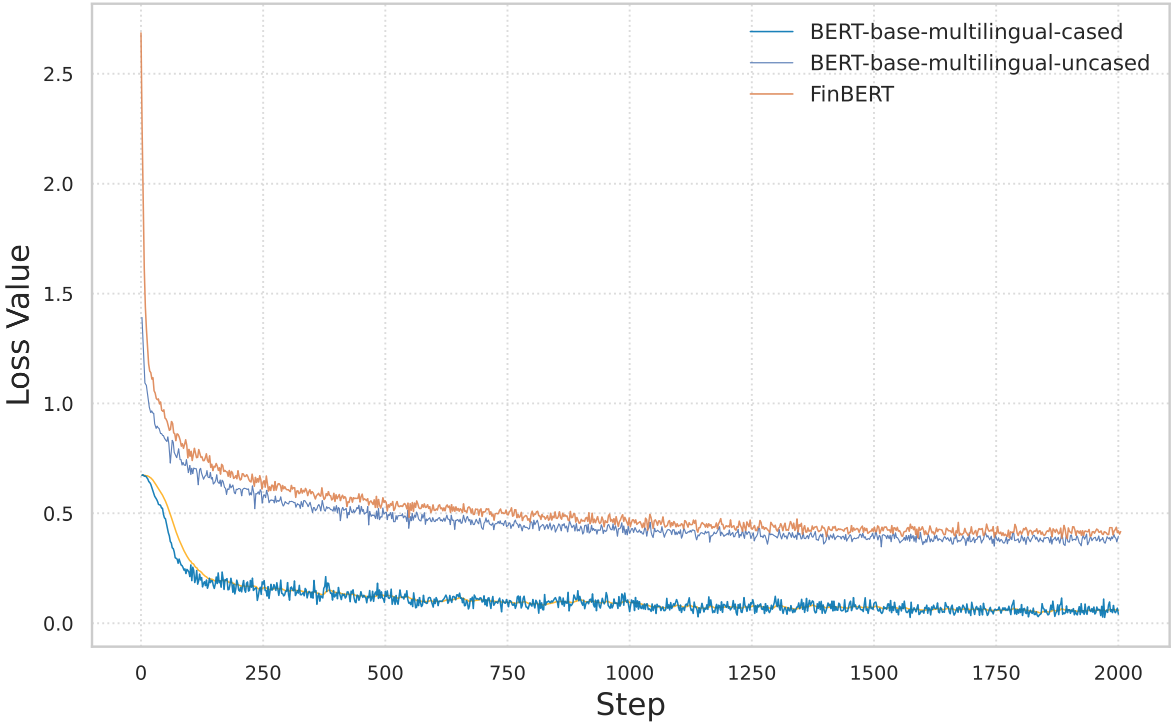


Figure 9: Training loss comparison between ‘cased‘ and ‘uncased‘ model

This performance gap can be attributed to the preprocessing strategy adopted by uncased models. Specifically, these models convert all text to lowercase and, more critically for tonal languages like Vietnamese, remove diacritical markers (accents). While lowercasing may have minimal semantic impact in English, it introduces severe distortions in Vietnamese, where diacritics are essential to distinguish between words and convey meaning. Stripping these markers effectively introduces systematic noise, resulting in a substantial loss of semantic information.

To illustrate this effect, Figure **??** presents examples of Vietnamese words before and after uncasing, highlighting the ambiguity and meaning collapse caused by diacritic removal. For instance, the Vietnamese word can lead to incorrect tokenization and misclassification in ESG contexts when accents are removed. Consequently, these findings emphasize the importance of preserving case and diacritics in multilingual models for low-resource, morphologically rich, or tonal languages such as Vietnamese.

This analysis demonstrates that model architecture alone is insufficient to guarantee strong cross-lingual performance; preprocessing strategies must account for language-specific orthographic and phonological features. These findings underscore the need for careful consideration when selecting multilingual models for languages such as Vietnamese. Practitioners should prioritize cased models or those that preserve diacritical markers, as uncased models can introduce severe semantic distortions and significantly degrade performance.

5.5. ESG Score Evaluate

Table 11: Sentiment Classification Performance for Environment

Model	**Overall accuracy**	**Negative F1 Score**	**Neutral F1 Score**	**Positive F1 Score**
PhoBERT	93.31	95.00	92.56	93.31
Vi-Electra	**93.87**	93.16	94.60	93.82
Visobert	92.47	92.36	92.92	92.18
DistilBERT-multilingual	92.20	91.86	92.64	92.11
DeBERTa	91.36	90.53	92.30	91.28
ViBERT	91.08	91.66	90.98	90.61
BERT-base	89.13	88.42	88.69	90.24
RoBERTa	87.46	87.28	88.98	86.17

From an environmental perspective, pre-trained language models such as Vi-Electra, PhoBERT, and Visobert demonstrate outstanding performance. Vi-Electra achieves the highest accuracy at 93.87%, along with superior F1 scores across all three classes (Negative, Neutral, Positive), indicating a well-balanced classification capability. PhoBERT also records an impressive accuracy (93.31%) and attains the highest F1 score for the “Negative” class (95.00%), showcasing strong ability in detecting environmentally negative content. DistilBERT-multilingual and Visobert follow closely with accuracies of 92.2% and 92.47%, respectively, while maintaining stable F1 scores across classes. Although DeBERTa achieves a very high F1 score for the Neutral class (92.30%), it shows a noticeable drop in accuracy (91.36%), placing it only in the moderately performing group. Meanwhile, BERT-base and especially RoBERTa, which is not multilingual-trained, demonstrate relatively low performance, making them unsuitable for complex environmental classification tasks.

In the governance domain, PhoBERT continues to lead with an accuracy of 92.04% and very high F1 scores across all three classes (Negative: 95.32, Neutral: 92.56, Positive: 93.31). This

Table 12: Sentiment Classification Performance for Governance

Model	**Overall accuracy**	**Negative F1 Score**	**Neutral F1 Score**	**Positive F1 Score**
PhoBERT	**92.04**	95.32	92.56	93.31
Vi-Electra	88.99	91.89	85.10	89.34
Visobert	88.68	93.21	84.65	87.70
DistilBERT-multilingual	86.85	90.90	83.41	85.95
DeBERTa	86.23	91.07	83.87	86.99
ViBERT	85.93	90.74	82.23	84.64
BERT-base	79.20	88.11	73.63	76.72
RoBERTa	65.74	75.12	61.75	61.20

model proves to be highly effective in capturing governance-related semantics and accurately classifying opinions. Vi-Electra and Visobert also show strong performance, with accuracies of 88.99% and 88.68%, respectively, particularly excelling in the "Negative" class with F1 scores above 91. DistilBERT-multilingual and DeBERTa yield fairly good results (accuracy around 86%) and remain viable options when a lightweight or multilingual-supporting model is needed. ViBERT performs at an average level (85.93%) and is slightly less competitive compared to newer models. Although BERT-base achieves a high F1 score for the Negative class (88.11), it suffers significant drops in the Neutral (73.63) and Positive (76.72) classes, resulting in an overall accuracy of just 79.2%. Lastly, RoBERTa continues to have the lowest performance with an accuracy of only 65.74%, clearly reflecting its unsuitability for governance classification tasks.

Table 13: Sentiment Classification Performance for Social

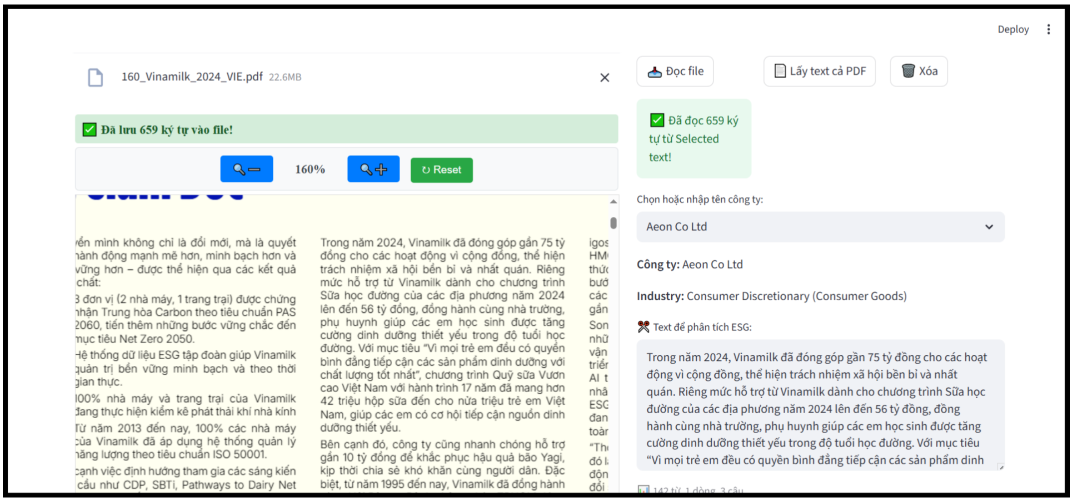
Model	**Overall accuracy**	**Negative F1 Score**	**Neutral F1 Score**	**Positive F1 Score**
PhoBERT	**92.75**	95.23	90.21	92.75
Vi-Electra	91.72	96.25	88.77	90.29
Visobert	88.58	92.39	83.23	86.99
DistilBERT-multilingual	86.55	90.05	83.06	86.40
DeBERTa	86.55	91.97	82.16	85.57
ViBERT	83.79	89.36	77.45	84.02
BERT-base	75.86	84.20	69.10	74.10
RoBERTa	73.79	78.16	69.84	73.91

In the social domain, which often involves more diverse content and subtler nuances. PhoBERT continues to affirm its position by achieving the highest accuracy (92.75%) along with outstanding F1 scores across all classes, especially the Negative class (95.23). Vi-Electra closely follows with an accuracy of 91.72%, standing out for its ability to detect negative content (96.25 F1), the highest among all models across all domains. Visobert, DeBERTa, and DistilBERT-multilingual show stable performance (accuracy above 86%) and demonstrate strong classification capabilities in the highly flexible linguistic context of social discourse.

ViBERT maintains an average position with an accuracy of 83.79%, lower than the leading group. Older models like BERT-base and RoBERTa, which were not pretrained in Vietnamese or in a multilingual setting, continue to perform poorly, with accuracies of only 75.86% and 73.79%, respectively. Their F1 scores for the Neutral class are particularly low, indicating difficulty in distinguishing between neutral and positive social nuances.

6. Software

The process of generating ESG analysis results begins with uploading a PDF document through the file uploader located in the left column of the interface. Once the PDF is uploaded, users can proceed with text extraction using one of two available methods. The first method involves manually selecting specific text sections within the PDF viewer by highlighting relevant content, which automatically saves the selected text to temporary files. The second method utilizes the "Get all PDF text" button to automatically extract the entire document content and normalize it for analysis. After text extraction, users must select the appropriate company from the dropdown menu or input a new company name, as this selection determines the industry-specific ESG weight factors that will be applied during scoring calculations according to the SASB Materiality Map standards. Our tool interface illustrator in Figure 10.



☑ Đã lưu 659 ký tự vào file!

ển mình không chỉ là đổi mới, mà là quyết hành động mạnh mẽ hơn, minh bạch hơn và vững hơn – được thể hiện qua các kết quả chất:

3 đơn vị (2 nhà máy, 1 trang trại) được chứng nhận Trung hòa Carbon theo tiêu chuẩn PAS 2060, tiến thêm những bước vững chắc đến mục tiêu Net Zero 2050.

Hệ thống dữ liệu ESG tập đoàn giúp Vinamilk quản trị bền vững minh bạch và theo thời gian thực.

100% nhà máy và trang trại của Vinamilk đang thực hiện kiểm kê phát thải khí nhà kính

Từ năm 2013 đến nay, 100% các nhà máy của Vinamilk đã áp dụng hệ thống quản lý năng lượng theo tiêu chuẩn ISO 50001.

cạnh việc định hướng tham gia các sáng kiến cầu như CDP, SBTi, Pathways to Dairy Net

Trong năm 2024, Vinamilk đã đóng góp gần 75 tỷ đồng cho các hoạt động vì cộng đồng, thể hiện trách nhiệm xã hội bền bỉ và nhất quán. Riêng mức hỗ trợ từ Vinamilk dành cho chương trình Sữa học đường của các địa phương năm 2024 lên đến 56 tỷ đồng, đồng hành cùng nhà trường, phụ huynh giúp các em học sinh được tăng cường dinh dưỡng thiết yếu trong độ tuổi học đường. Với mục tiêu "Vì mọi trẻ em đều có quyền bình đẳng tiếp cận các sản phẩm dinh dưỡng với chất lượng tốt nhất", chương trình Quỹ sữa Vươn cao Việt Nam với hành trình 17 năm đã mang hơn 42 triệu hộp sữa đến cho nửa triệu trẻ em Việt Nam, giúp các em có cơ hội tiếp cận nguồn dinh dưỡng thiết yếu.

Bên cạnh đó, công ty cũng nhanh chóng hỗ trợ gần 10 tỷ đồng để khắc phục hậu quả bão Yagi, kịp thời chia sẻ khó khăn cùng người dân. Đặc biệt, từ năm 1995 đến nay, Vinamilk đã đồng hành

☑ Đã đọc 659 ký tự từ Selected text!

Chọn hoặc nhập tên công ty:

Aeon Co Ltd

Công ty: Aeon Co Ltd

Industry: Consumer Discretionary (Consumer Goods)

Text để phân tích ESG:

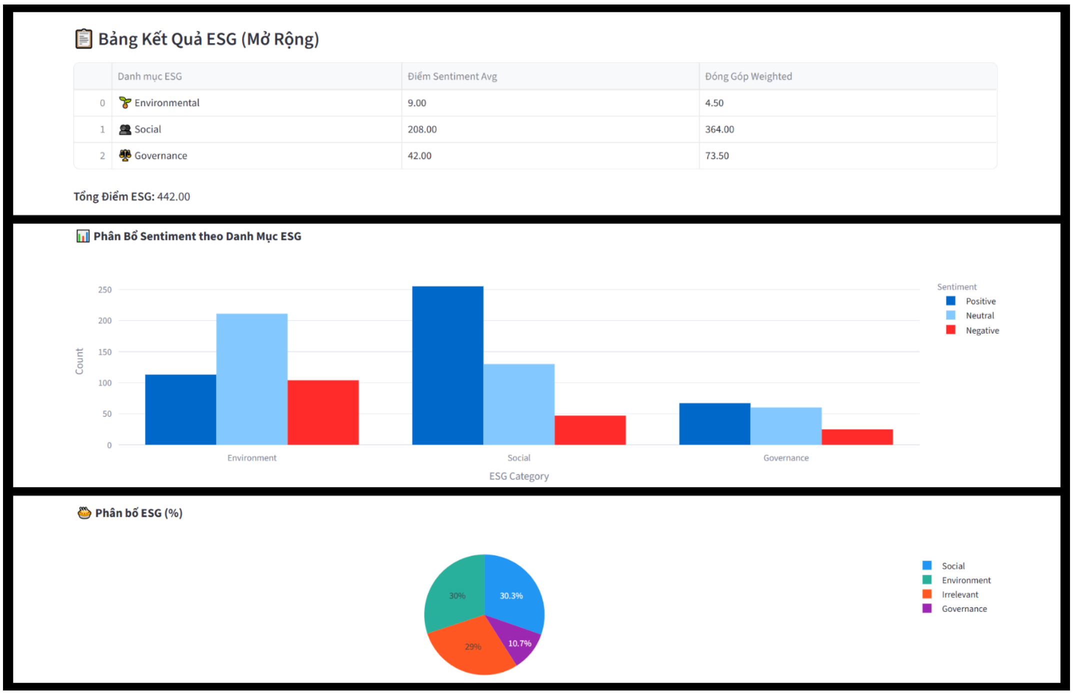
Trong năm 2024, Vinamilk đã đóng góp gần 75 tỷ đồng cho các hoạt động vì cộng đồng, thể hiện trách nhiệm xã hội bền bỉ và nhất quán. Riêng mức hỗ trợ từ Vinamilk dành cho chương trình Sữa học đường của các địa phương năm 2024 lên đến 56 tỷ đồng, đồng hành cùng nhà trường, phụ huynh giúp các em học sinh được tăng cường dinh dưỡng thiết yếu trong độ tuổi học đường. Với mục tiêu "Vì mọi trẻ em đều có quyền bình đẳng tiếp cận các sản phẩm dinh

142 từ, 1 dòng, 3 câu

Figure 10: Interface for ESG classification and score system

Following text preparation, users review the extracted content in the analysis area to ensure it meets the minimum requirement of 10 characters for analysis. The system automatically displays text statistics including word count, line count, and sentence count, and provides a status indicator showing when the text is ready for analysis. Once validation is complete, users initiate the ESG analysis by clicking the "Analyze ESG" button, which triggers the automated processing system to categorize text into Environmental, Social, Governance, and Irrelevant categories while performing sentiment analysis for each ESG dimension.

The final step involves reviewing the comprehensive analysis results, which are presented through multiple visualization formats including bar charts showing sentence counts per category, pie charts displaying percentage distributions, and sentiment distribution charts



📋 Bảng Kết Quả ESG (Mở Rộng)

	Danh mục ESG	Điểm Sentiment Avg	Đóng Góp Weighted
0	Environmental	9.00	4.50
1	Social	208.00	364.00
2	Governance	42.00	73.50

Tổng Điểm ESG: 442.00

Figure 11: Output for ESG classification and score system

illustrating the Positive/Neutral/Negative breakdown across ESG categories. The system automatically calculates weighted ESG scores based on sentiment analysis and industry-specific weight factors, presenting the results in a detailed table format that shows sentiment averages and weighted contributions for each ESG dimension, ultimately providing a complete ESG performance assessment. All input analysis is shown in Figure 11.

7. Conclusion and Future Work

This thesis addresses the critical gap in NLP resources and tools for Environmental, Social, and Governance (ESG) analysis in Vietnamese, a low-resource language that has been largely overlooked in prior research. To bridge this gap, we introduced the **ViEn-ESG** dataset, a large-scale bilingual (Vietnamese–English) resource comprising 130,798 sentence-level samples labeled across four classes (Environmental, Social, Governance, and Neutral). To the best of our knowledge, this is the first publicly available dataset designed for ESG text classification in Vietnamese, providing a valuable foundation for future research in sustainability analytics for emerging markets.

Building on this dataset, we fine-tuned a series of BERT-based models for ESG classification, demonstrating the effectiveness of transfer learning for domain-specific and low-resource settings. Our best-performing bilingual model achieved an F1 score of 94.83% on English and 91.94% on Vietnamese, while the Vietnamese-only model attained an accuracy of 94.66%. These results establish a strong multilingual benchmark for ESG classification and underscore the capability of large pre-trained language models to deliver robust cross-lingual performance.

In addition to classification, we introduced an ESG scoring mechanism that aggregates sentiment signals across Environmental, Social, and Governance dimensions. This approach provides interpretable, quantitative assessments of corporate sustainability, offering practical utility for researchers, rating agencies, and industry stakeholders. Collectively, our contributions advance both the methodological and practical landscape for ESG analysis in low-resource languages.

This thesis addresses the critical gap in Natural Language Processing (NLP) resources and tools for Environmental, Social, and Governance (ESG) analysis in Vietnamese, a low-resource language that has been largely overlooked in prior research.

Building on the findings and contributions of this thesis, we identify several directions for future research to advance ESG classification in Vietnamese and multilingual contexts:

- **Dataset Enrichment and Extension:** With ViEn-ESG released in this work, future efforts can expand the dataset with additional ESG sources such as CSR disclosures and sustainability reports. Augmentation techniques and integration of structured data (e.g., emissions, workforce metrics) can further enhance model performance and applicability.
- **Extending to Broader ESG Tasks:** Beyond classification, future research could explore tasks such as ESG entity recognition, impact type and duration detection, and hierarchical topic modeling for deeper insight extraction from sustainability reports.
- **Cross-Lingual and Multilingual Extensions:** Expanding to additional languages and leveraging advanced cross-lingual transfer learning techniques can enable scalable ESG analytics in diverse linguistic contexts, supporting global sustainability assessments.
- **Real-World Applications:** The classified ESG information could be used to track companies' non-financial performance in near real-time or to enhance ESG rating systems with timely and explainable insights.

By addressing these directions, future research can further strengthen scalable, multilingual ESG analytics and contribute to more transparent, data-driven sustainability evaluation frameworks.

References

[1] Victoria Agbakwuru, Obinna Barnabas Onyenahazi, Bernard Owusu Antwi, and Kabir Oyewale. The impact of environmental, social, and governance (esg) reporting on corporate financial performance. *International Journal of Research Publication and Reviews*, 5(9):3629–3644, September 2024. ISSN 2582-7421. doi: 10.55248/gengpi.5.0924.2710.

[2] Doron Avramov, Si Cheng, Abraham Lioui, and Andrea Tarelli. Sustainable investing with esg rating uncertainty. *Journal of Financial Economics*, 145(2):642–664, August 2022. ISSN 0304-405X. doi: 10.1016/j.jfineco.2021.09.009.

[3] Mahmut Aydoğmuş, Güzhan Gülay, and Korkmaz Ergun. Impact of esg performance on firm value and profitability. *Borsa Istanbul Review*, 22:S119–S127, December 2022. ISSN 2214-8450. doi: 10.1016/j.bir.2022.11.006.

[4] Florian Berg, Julian Kouml;lbel, and Roberto Rigobon. Aggregate confusion: The divergence of esg ratings. *SSRN Electronic Journal*, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3438533.

[5] Monica Billio, Michele Costola, Iva Hristova, Carmelo Latino, and Loriana Pelizzon. Inside the esg ratings: (dis)agreement and performance. *SSRN Electronic Journal*, 2020. ISSN 1556-5068.

[6] Mattia Birti, Francesco Osborne, and Andrea Maurino. Optimizing large language models for esg activity detection in financial texts, 2025.

[7] The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. Improving sequence tagging for Vietnamese text using transformer-based neural models. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 13–20, Hanoi, Vietnam, October 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.paclic-1.2/.

[8] Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 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 111–115, Macao, 20 August 2023. -. URL https://aclanthology.org/2023.finnlp-1.11/.

[9] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. ISSN 1552-3888. doi: 10.1177/001316446002000104.

[10] Common Crawl. Common crawl, 2025. URL https://commoncrawl.org/. Accessed: 2025-09-10.

[11] Alessandro Del Vitto, Daniele Marazzina, and Davide Stocco. Esg ratings explainability through machine learning techniques. *Annals of Operations Research*, July 2023. ISSN 1572-9338. doi: 10.1007/s10479-023-05514-z.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009.

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

[14] Ha Noi Stock Exchange, 2025. URL https://hnx.vn/en-gb/. Accessed: 2025-09-10.

[15] Ha Noi Stock Exchange, 2025. URL https://www.hsx.vn/. Accessed: 2025-09-10.

[16] Jannik Fischbach, Max Adam, Victor Dzhagatspanyan, Daniel Mendez, Julian Frattini, Oleksandr Kosenkov, and Parisa Elahidoost. Automatic esg assessment of companies by mining and evaluating media coverage data: Nlp approach and tool, 2022.

[17] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2021.

[18] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[19] Allen H. Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, January 2023. ISSN 1911-3846. doi: 10.1111/1911-3846.12832.

[20] Xuran Jiang. Predicting corporate esg scores using machine learning: A comparative study. *Advances in Economics, Management and Political Sciences*, 118(1):141–147, December 2024. ISSN 2754-1177. doi: 10.54254/2754-1169/2024.18728.

[21] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1), May 2016. ISSN 2052-4463.

[22] Yu Han Kao, Vidhya Nataraj, Ting-Chi Wang, Yu-Jyun Zheng, Hsiao-Chuan Liu, Wen-Hsuan Liao, Chia-Tung Tsai, and Min-Yuh Day. IMNTPU at ML-ESG-3: Transformer language models for multi-lingual ESG impact type and duration classification. pages 228–233, May 2024. URL https://aclanthology.org/2024.finnlp-1.23/.

[23] A. Kazakov, S. Denisova, I. Barsola, E. Kalugina, I. Molchanova, I. Egorov, A. Kosterina, E. Tereshchenko, L. Shutikhina, I. Doroshchenko, N. Sotiriadi, and S. Budennyy. Esgify: Automated classification of environmental, social, and corporate governance risks. *Doklady Mathematics*, 108(S2):S529–S540, December 2023. ISSN 1531-8362. doi: 10.1134/s1064562423701673.

[24] Takuya Kiriu and Masatoshi Nozaki. A text mining model to evaluate firms' esg activities: An application for japanese firms. *Asia-Pacific Financial Markets*, 27(4):621–632, 2020. ISSN 1573-6946. doi: 10.1007/s10690-020-09309-1.

[25] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1), March 2023. ISSN 2052-4463.

[26] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.

[27] Haein Lee, Jang Hyun Kim, and Hae Sun Jung. Esg-kibert: A new paradigm in esg evaluation using nlp and industry-specific customization. *Decision Support Systems*, 193:114440, June 2025. ISSN 0167-9236. doi: 10.1016/j.dss.2025.114440.

[28] Jaeyoung Lee and Misuk Kim. Esg information extraction with cross-sectoral and multi-source adaptation based on domain-tuned language models. *Expert Systems with Applications*, 221:119726, 2023. ISSN 0957-4174.

[29] Jaeyoung Lee, Geonyeong Son, and Misuk Kim. Esg-kor: A korean dataset for esg-related information extraction and practical use cases. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 6627–6643. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.387.

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017.

[32] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. Finer: Financial numeric entity recognition for xbrl tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.303.

[33] Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. Analyzing sustainability reports using natural language processing. 2020. doi: 10.48550/arXiv.2011.08073.

[34] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, WWW '18, page 1941–1942. ACM Press, 2018. doi: 10.1145/3184558.3192301.

[35] Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese, 2020.

[36] Hatto EraX Team Nguyen. Erax-translatoe-v1.0: A compact and capable multilingual translation model. 2025. URL https://huggingface.co/erax-ai/EraX-Translator-V1.0.

[37] Nam Nguyen, Thang Phan, Duc-Vu Nguyen, and Kiet Nguyen. Visobert: A pre-trained language model for vietnamese social media text processing. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing, page 5191–5207. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.315.

[38] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 5206–5210. IEEE, April 2015. doi: 10.1109/icassp.2015.7178964.

[39] Aarav Patel and Peter Gloor. Creating a systematic esg (environmental social governance) scoring system using social network analysis and machine learning for more sustainable company practices, 2023.

[40] Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. Leveraging bert language models for multi-lingual esg issue identification, 2023.

[41] PubMed. Pubmed, 2025. URL https://pubmed.ncbi.nlm.nih.gov/. Accessed: 2025-09-10.

[42] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2019.

[43] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[44] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[45] Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979, March 2024. ISSN 1544-6123. doi: 10.1016/j.frl.2024.104979.

[46] César Pedrosa Soares. Leveraging natural language and item response theory models for esg scoring, 2024.

[47] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020.

[48] Priyank Soni, Sohom Ghosh, and Sudip Kumar Naskar. *Detecting Issues Related to Environmental, Social, and Corporate Governance Using SEC-BERT*, page 329–338. Springer Nature Singapore, 2024. ISBN 9789819978205. doi: 10.1007/978-981-99-7820-5_27.

[49] Spacy. Spacy, 2025. URL https://pypi.org/project/spacy/. Accessed: 2025-09-10.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[51] Abby Yaqing Zhang and Joseph H. Zhang. Renovation in environmental, social and governance (esg) research: the application of machine learning. *Asian Review of Accounting*, 32(4):554–572, November 2023. ISSN 1321-7348. doi: 10.1108/ara-07-2023-0201.

[52] Mengdi Zhang, Qiao Shen, Zhiheng Zhao, Shuaian Wang, and George Q. Huang. Optimizing esg reporting: Innovating with e-bert models in nature language processing. *Expert Systems with Applications*, 265:125931, March 2025. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.125931.

A. Translate Prompt for EraX-Translator-V1.0

To enrich the Vietnamese portion of our dataset, we applied a translation-based labeling strategy that converts high-quality English sentences with gold-standard ESG labels into Vietnamese. This approach provides a scalable method for generating labeled data in a low-resource language, ensuring semantic alignment across both languages.

For this process, we used the official system prompt recommended by the authors of the `erax-ai/EraX-Translator-V1.0` model to ensure accurate and domain-preserving translations:

> Bạn là Trợ lý AI xuất sắc về dịch thuật nhiều ngôn ngữ, đặc biệt tiếng Anh, tiếng Trung Hoa, tiếng Việt.
>
> Bạn cũng là 1 Hoà thượng Phật giáo uyên thâm về dịch thuật Cổ văn Trung Quốc. Người dùng sẽ giao nhiệm vụ dịch thuật cho bạn từ ngôn ngữ bất kỳ sang một ngôn ngữ được chỉ định.
>
> Nhiệm vụ của bạn là dịch thật sát nghĩa, thể hiện đúng ý của bài gốc và không chế tác hay bịa đặt gì thêm. Đặc biệt lưu ý danh xưng phải giữ nguyên vẹn, dịch đúng tên người, tên địa danh phải tuyệt đối chính xác. Không được bình luận, không được cung cấp lời giới thiệu hay mở bài hay kết luận gì, chỉ dịch thật sát nghĩa và không bỏ qua bất kỳ ý hay từ nào.
>
> {sentence}. Dịch sang tiếng Việt

Figure 12: Prompt for EraX-Translator-V1.0 models.