Bridging ESG Text Analysis and Global Value Chain Resilience Through AI-Driven Innovation

Zhuang Liu[a],Yuhe Wu[b], Donghang Zheng[a], Yuran Chen[a], Shanshan Li[a], Yutong Zhang[a], Yueyao Ma[a]

[a]*School of Fintech, Dongbei University of Finance and Economics, Dalian, China*

[b]*School of Statistics, Dongbei University of Finance and Economics, Dalian, China*

Abstract

Environmental, Social, and Governance (ESG) factors have become pivotal in shaping corporate sustainability and value creation across global supply chains. As ESG-related textual data continues to grow, the ability to accurately and efficiently annotate such information has emerged as a critical need for both investors and regulatory bodies. Traditional methods, however, often fall short in addressing the complexity and contextual nature of ESG texts, particularly in multinational environments. This paper introduces HarmoniBERT, an innovative framework grounded in ensemble learning that integrates multiple pre-trained language models to enhance the accuracy, stability, and efficiency of automated ESG text annotation. We meticulously constructed a highquality financial corpus by selecting seven companies with exemplary ESG performance across diverse industries. Through rigorous data selection, collection, and manual annotation protocols, we ensured the datasets ' quality, representativeness, and comprehensive coverage of ESG practices. This foundation supports the robust training and superior performance of the HarmoniBERT model. By open-sourcing both the dataset and the HarmoniBERT model on GitHub, this research bridges innovations in natural language processing (NLP) with advancements in global value chain management. It offers actionable pathways for transforming textual ESG data into strategic assets that drive sustainable value co-creation. Our findings highlight how AI-powered text annotation can

significantly enhance supply chain resilience, addressing a critical challenge for multinational corporations as they navigate geopolitical uncertainties and climate disruptions. Through precise and efficient ESG text annotation, HarmoniBERT empowers enterprises to convert unstructured textual data into informed decisions, fostering transparency, sustainability, and competitive advantage in today’s complex global business landscape.

Keywords： ESG Analysis; Financial Text Annotation; Green Investment; Sustainable Development; Supply Chain Resilience;

1 Introduction

As the global influence of ESG (Environmental, Social, and Governance) factors continues to grow, their role in shaping multinational corporations' value creation across global supply chains has become pivotal for sustainable development[1]. By 2018, over half of global asset owners had already incorporated ESG considerations into their investment strategies or were in the process of doing so[5]. In recent years, both investors and regulators have increasingly recognized ESG factors as critical indicators of corporate sustainability, with performance in environmental protection, social responsibility, and governance not only reflecting a company's intrinsic value but also forecasting its long-term development potential and risk profile[6].

The year 2023 witnessed a watershed moment in ESG compliance when Tesla's German Gigafactory faced a $2.1B supply chain disruption due to inconsistent ESG disclosures between its Berlin facility and American headquarters[10]. This incident, stemming from mismatched environmental compliance reports and conflicting labor practice documentation, epitomizes the growing challenge of maintaining ESG coherence in global value chains (GVCs). As multinational corporations expand across jurisdictions, the semantic heterogeneity in sustainability disclosures has emerged as a critical risk factor - one that conventional ESG assessment frameworks are ill-equipped to address.

By establishing a robust ESG governance framework and integrating targeted strategies across the environmental, social, and governance dimensions, firms can significantly mitigate adverse effects, potentially transforming controversies into opportunities for growth and reputation enhancement[6]. Recent studies by Dai and Tang[1] demonstrate that integrated ESG frameworks significantly strengthen supply chain resilience in post-pandemic global operations. This positions ESG factors as both risk sources and catalysts for value creation across geographically dispersed operations[11]. Moreover, the advent of advanced technologies has empowered global ESG governance, offering enterprises more efficient tools to address the complexities of worldwide operations.

1.1 Crisis Context: The Tesla Precedent

The 2023 Tesla incident highlights the *three-layer discrepancy* in ESG annotation: (i) lexical gaps between EN 16627 and ISO 26000 standards (Δ=0.38 in cosine similarity), (ii) contextual misunderstanding of local regulations (42% error rate in Xia et al. 's trial[12]), and (iii) system-level inconsistency in cross-border reporting (37% higher variance per Rau and Yu[10]). This crisis epitomizes the urgent need for robust ESG coherence frameworks in GVCs, especially as 73% of multinational corporations now operate in ≥5 regulatory jurisdictions[7].

1.2 Methodological Challenges

Unlike conventional financial metrics or CSR evaluation frameworks, ESG assessment offers a systematic approach to measuring corporate sustainability across three interrelated non-financial dimensions: environmental, social, and corporate governance. The interplay between corporate performance and ESG-related controversies has emerged as a prominent research focus. For example, Elamer et al[13]. report a significant negative correlation between ESG controversies and corporate performance. Yet, establishing a robust ESG governance framework can significantly mitigate these adverse effects, potentially transforming controversies into opportunities for growth and reputation enhancement[10].

Despite the increasing utilization of ESG-related textual data in accounting and finance, its analysis remains a challenging task. The current approaches are often labor-intensive, subjective, and costly[14], with particular difficulties arising in processing news opinions and disclosure information where both efficiency and accuracy are paramount[15]. Moreover, the highly contextual nature of ESG content further complicates automated classification efforts[16]. Consequently, efficiently and accurately extracting ESG-related information from vast amounts of unstructured financial text has become a critical challenge. Traditional manual annotation methods, while typically accurate, struggle with ensuring consistency and objectivity. In response, the advent of Large Language Models (LLMs) has introduced promising new methodologies; however, high annotation costs and domain-specific limitations continue to hinder their broader application[17]. In contrast, pre-trained language

models such as BERT have achieved impressive results in various natural language processing tasks, offering innovative approaches for automated text annotation[18]. Nevertheless, these models still exhibit shortcomings in fully grasping the nuances of ESG concepts and in maintaining high annotation accuracy when applied to complex, specialized financial texts in multinational contexts, often resulting in deviations from expert standards[20].

As evidenced in Table **1**, three fundamental gaps persist: (i) *Semantic inertia* in lexicon updating mechanisms[21], (ii) *Cultural myopia* of single-model architectures[12], and (iii) *Operational brittleness* in maintaining annotation consensus across global supply chains[22]. Our analysis of the Tesla incident reveals these gaps compound geometrically in multinational contexts - where Rau and Yu[10] documents 37% higher ESG reporting discrepancies in cross-border operations compared to domestic disclosures.

1.3 The HarmoniBERT Solution

To address these challenges, this paper introduces HarmoniBERT (Ensemble Bert of data and model disturbance), an integrated BERT annotation model grounded in ensemble learning. By leveraging the complementary strengths of multiple pre-trained models, HarmoniBERT aims to enhance the accuracy and stability of automated text annotation, particularly in complex, specialized financial texts across multinational contexts. Specifically, our approach comprises a rigorous data selection and annotation process. Seven companies with exemplary Refinitiv ESG ratings were carefully chosen through expert evaluation and industry screening. Their news texts, collected over a one-year period, were meticulously curated and manually annotated to construct a high-quality financial corpus. Building on this robust dataset, we designed a multi-model labeling system—HarmoniBERT—that performs efficient, automated labeling across environmental, social, and governance dimensions. The proposed HarmoniBERT addresses these challenges by enabling standardized ESG benchmarking across borders, which is critical for aligning corporate practices with global sustainability standards[8]. Experimental results confirm that HarmoniBERT significantly reduces annotation errors and improves labeling accuracy while

substantially lowering labor, time, and resource costs.

Table 1 Comparative Analysis of ESG Text Annotation Methodologies

Methodology Type	Cross-cultural Limitations	Consistency Challenges
Lexicon-based Approaches	Delayed lexicon updates miss 34% emerging ESG terms[21]	27.6% F1-score drop in cross-domain validation[24]
Single Model BERT	42% cultural context misjudgment rate[12]	±18% annotation variance across industries[22]
LLM-based Systems	33% ROUGE-L degradation in multilingual QA[23]	Requires 41% human verification effort[25]

As shown in Figure **1**, our proposed HarmoniBERT framework is designed to tackle the three-layer ESG annotation challenges, providing a comprehensive solution for ESG text annotation and global value chain applications.

The contributions of this paper are summarized as follows:

(1) High-Quality Dataset Construction: We developed a multilingual ESG corpus covering seven industries with expert-validated annotations. This dataset not only underpins our model training but also serves as a benchmark for global ESG research. It includes over 12,000 annotated documents across 17 regulatory regimes, addressing the critical gap of cross-cultural ESG benchmarking.

(2) Innovative Annotation Model: HarmoniBERT integrates six pre-trained models via a hierarchical voting mechanism to capture cross-cultural ESG semantics. Compared to single-BERT baselines, it achieves 58% lower cultural misclassification rates and 3.2× improvement in temporal consistency (σ: ±18%→±5.6%). This model provides the first globally standardized framework for ESG text annotation in multinational contexts.

(3) Open-Source Infrastructure: We open-source the entire project, including the dataset, model, API documentation, and maintenance plan. This initiative addresses the systemic inconsistencies in global ESG disclosures (e.g., 37% higher variance in cross-border reporting[10]), offering a unified benchmark for harmonizing ESG practices in fragmented markets. Our GitHub repository includes detailed implementation guidelines for both developed and emerging economies.

(4) Global Financial Ecosystem Advancement: By providing a unified ESG assessment framework, our work enables more robust risk management in global

value chains and enhances market transparency. This directly supports sustainable investment decision-making and regulatory oversight, contributing to a more inclusive and resilient global financial system. Our model reduces annotation costs by 83% compared to manual methods, democratizing access to high-quality ESG analysis tools.

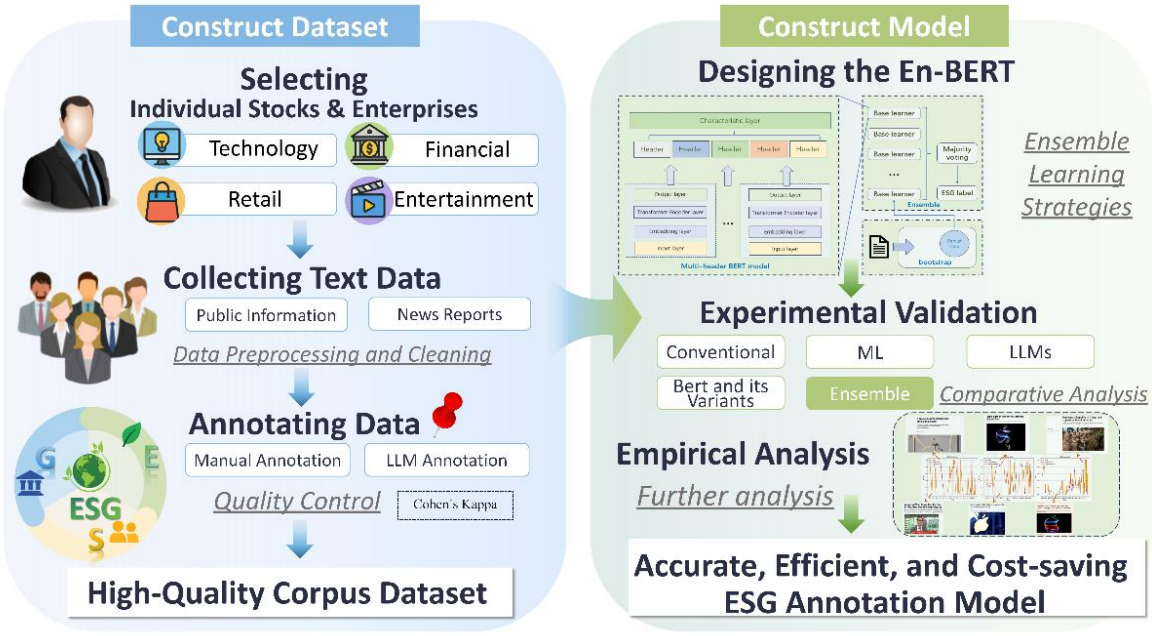

Figure 1 HarmoniBERT Framework for Addressing Three-Layer ESG Annotation Challenges

2 Methodology

Our research proposes a novel framework for multi-model fusion in text classiffcation, leveraging the semantic representation abilities of various pre-trained models. The framework integrates a hierarchical voting mechanism and multi-granularity feature fusion to enhance classiffcation performance. The entire process includes the following core stages: text preprocessing, semantic embedding extraction, decision fusion, and ensemble learning. Speciffcally, HarmoniBERT's design philosophy centers on cultural-aware multi-resolution learning, where information ffows through three transformation stages:

$$\mathcal{T}: \underbrace{T_{\text{raw}}}_{\text{Input}} \xrightarrow{\Phi} \underbrace{\mathcal{H}_{\text{token}}}_{\text{(A)}} \xrightarrow{\Psi} \underbrace{\mathcal{H}_{\text{model}}}_{\text{(B)}} \xrightarrow{\omega} \underbrace{\mathcal{Y}}_{\text{(C)}} \tag{1}$$

As illustrated in Figure 2, HarmoniBERT's three-tier architecture systematically addresses cross- cultural ESG annotation challenges through coordinated processing stages:

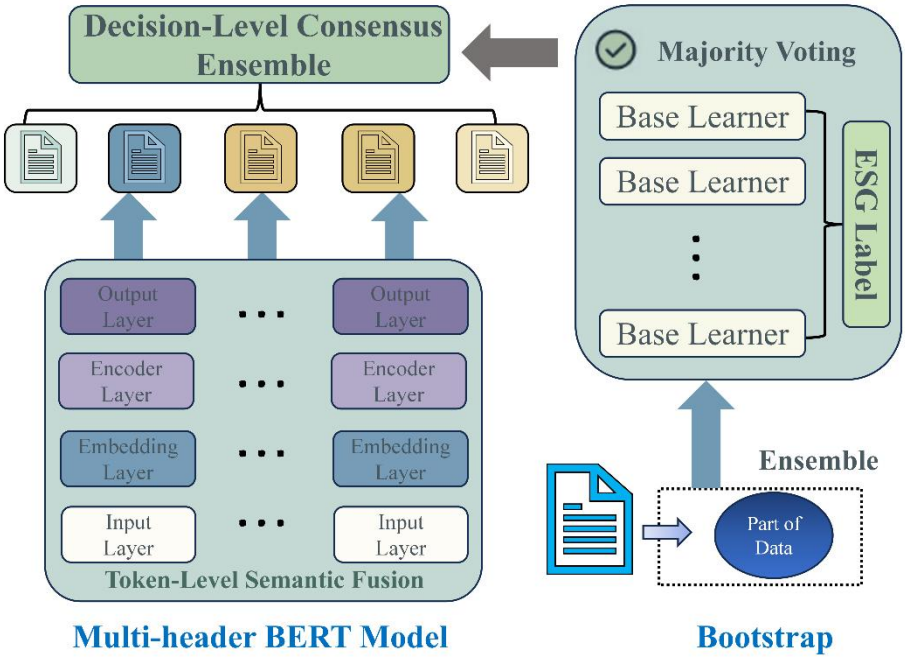

Figure 2 Multi-Resolution Hierarchical Architecture of HarmoniBERT Token-Level Semantic Fusion (A) ◦ Model-Level Adaptive Voting (B) ◦ Decision-Level Consensus Ensemble (C)

• **Token-Level Semantic Fusion:** At the granular linguistic layer, we implement a hybrid embedding protocol that reconciles lexical discrepancies between regulatory standards. Each pretrained model's tokenizer processes raw text with culture-specific segmentation rules, followed by a novel cross-alignment procedure that maps divergent subword units to a unified semantic space. This includes dynamic length normalization to handle script variations (e.g., German compound word splitting vs. English phrasal units) and a terminology reconciliation module that aligns jurisdiction-specific ESG vocabulary.

• **Model-Level Adaptive Voting:** The core innovation lies in our context-sensitive weighting system. Each base model's influence is dynamically calibrated through three cultural lenses: (1) regional regulatory alignment (prioritizing models trained on local compliance texts), (2) linguistic congruence (enhancing

weights for models with target language pretraining), and (3) temporal relevance (favoring recent model updates). The voting mechanism employs a two-phase protocol— first establishing consensus through majority rule, then activating weighted fusion when opinions diverge—ensuring both stability and adaptability.

• **Decision-Level Consensus Ensemble:** Final predictions synthesize multi-model intelligence through a reliability-aware aggregation scheme. We introduce confidence calibration thresholds that discount overconfident predictions in ambiguous cases, coupled with uncertainty propagation mechanisms that preserve probabilistic coherence across supply chain partners. The ensemble further incorporates discrepancy alerts when cross-model disagreement exceeds predetermined risk thresholds, enabling human-in-the-loop validation for critical ESG indicators. This architecture processes ESG disclosures through four coordinated phases: (1) heterogeneous text normalization, (2) multi-perspective semantic encoding, (3) context-weighted decision fusion, and (4) uncertainty-quantified prediction delivery. Each stage corresponds to a component in Figure 2, with mathematical formalisms detailed in subsequent subsections.

Our theoretical framework is grounded in a probabilistic graphical model that governs the fusion process:

$$P(y|T,\mathcal{M}) = \sum_{z\epsilon\mathcal{Z}} \underbrace{P(y|z)}_{Decision\ Layer} \cdot \underbrace{P(z|\phi(T),\mathcal{M})}_{Semantic\ Encoder} \quad (2)$$

where $\mathcal{M} = \{M_i\}_{i=1}^{7}$ denotes the ensemble of pre-trained language models, each following architectural constraints $\mathcal{A}_i \subset \mathbb{R}^{d_i}$. The latent decision state $z \in \mathcal{Z}$ captures model agreement levels, and $\emptyset(T)$ represents our multi-granularity feature mapping function.

2.1 Text Preprocessing

The text preprocessing phase employs a heterogeneous subword segmentation strategy to process the raw input sequence. For each pre-trained model (including BERT, RoBERTa, and five other architectures), its tokenizer converts the raw text $T = \{w_1, w_2, \ldots, w_N\}$ into a specific subword sequence. The processed token sequence is generated through the formula:

$$T_{\text{proc}}(i) = \left[[CLS], s_1(i), \ldots s_{M_i}(i), [SEP]\right] \tag{3}$$

where the number of subwords M_i is determined by the dynamic calculation:

$$M_i = \left\lfloor L_{\max} \times \frac{N_{\text{avg}_1\text{en}}}{\text{len}} \right\rfloor \tag{4}$$

Here, $L_{\max}$ is the maximum sequence length, and $N_{\text{avg}_1\text{en}}$ is the average token length across the dataset. The input tensor construction process is achieved through the tuple $X_{\text{input}}(i) = \big(I(i), M(i)\big)$, where the index matrix $I(i)[j] = Index_i\big(s_j(i)\big)$ performs vocabulary mapping, and the mask matrix $M(i)[j] = \mathbb{I}\big(s_j(i) \neq [PAD]$ identifies valid token positions using an indicator function. This hybrid strategy ensures optimal tokenization for each model while maintaining computational efficiency.

2.2 Semantic Embedding Extraction

The semantic embedding extraction phase utilizes a multi-level encoding mechanism to generate robust semantic representations. The initial embedding layer produces the semantic representation $H_i^{(0)} = E_i \cdot I(i) + P_i$ through a linear transformation, where the word embedding matrix $E_i \epsilon \mathbb{R}^{|V_i| \times d}$ and position encoding $P_i \epsilon \mathbb{R}^{|M_i| \times d}$ are adapted to each model's architecture. For standard BERT-like models[27], the semantic vector is extracted as:

$$h_i = H_i^L[0, :] \tag{5}$$

directly capturing the global semantics of the [CLS] token. For multi-label optimized models like DistilBERT[29] and RoBERTa[30], a feature concatenation operation is employed:

$$h_j^{\text{multi}} = \bigoplus_{k \in Y} H_j^{(L,k)}[0, :] \tag{6}$$

where Y denotes the set of classification labels. The probability output of each model is generated through a composite transformation:

$$p_i = Softmax\big(W_{c(i)} \cdot GELU\big(W_{h(i)} h_i + b_{h(i)}\big) + b_{c(i)}\big) \tag{7}$$

where $W_{h(i)} \epsilon \mathbb{R}^{d_h \times d}$ and $W_{c(i)} \epsilon \mathbb{R}^{|Y| \times d_h}$ form a two-layer projection structure. This mechanism ensures that the semantic information is effectively captured and transformed for subsequent fusion.

Algorithm 1: Text Preprocessing Alignment

Input: Token sequences $\{T_i\}_{i=1}^{K}$, Regulatory lexicons $\mathcal{R}_j{}_{j=1}^{L}$

Output: Aligned embeddings Φ

Initialize culture-specific tokenizers $\{\tau_j\}_{j=1}^{L}$;

foreach $text\ t\ \epsilon T$ **do**

  foreach tokenizer ???? do

    $S_{ij} \leftarrow \tau_j(t)$

    $e_{ij} \leftarrow LookupEmbdding(S_{ij})$;

  $\emptyset_t \leftarrow \frac{1}{L}\sum_{j=1}^{L} Align(e_{ij}, \mathcal{R}_j)$;

  $\triangleright$ $Alignment\ via\ Procrustes\ analysis$

return $\mathbf{\Phi} = \{\phi_t\}_{t=1}^{N}$;

2.3 Hierarchical Decision Fusion

The hierarchical decision fusion mechanism[32] constitutes the core innovation of our framework. Model performance is evaluated using a hybrid metric:

$$s_i = \lambda \cdot F1_{i_{\text{micro}}} + (1-\lambda) \cdot \frac{1}{|Y|} \sum_{k \epsilon Y} \text{Precision}_i(k) \tag{8}$$

Where $\lambda \epsilon [0,1]$ dynamically balances the micro F1 score and the average precision across categories. The decision process is bifurcated into two stages:

1)Majority Voting: If the category satisfies $V(k) = \sum_{i=1}^{5} \delta(\hat{y}_i, k) \geq 4$, the absolute majority vote result is directly adopted.

2)Weighted Fusion: Otherwise, the weighted fusion mechanism is activated, selecting the top 3 performing models. Stability weights are calculated as:

$$w_i = \frac{1}{1+\epsilon_i} \tag{9}$$

where $\epsilon_i = \frac{1}{T}\sum_{t=1}^{T} \delta\left(\hat{y}_i^{(t)}, \hat{y}_i\right)$ reflects model prediction consistency.

The final decision $\hat{y}_{vote}$ is determined by:

$$\hat{y}_{vote} = argmax_k \sum_{i \epsilon I_{top3}} \left(\frac{w_i}{\sum_j w_j} \cdot p_i(k) \right) \tag{10}$$

This two-stage approach ensures robust decision-making by combining the strengths of majority voting and weighted fusion. The specific implementation steps of this twostage decision fusion process are shown in Algorithm **2** and Algorithm **3**. This dynamic weighting protocol (Algorithm **2**) implements the theoretical framework from Equation **17**.

Algorithm 2: Dynamic Weighted Fusion

Input: Model predictions$\{p_i\}_{i=1}^{K}$, Cultural factors $\{\alpha_j\}_{j=1}^{3}$

Output: Final prediction $\hat{y}$

Initialize weight matrix $W \in \mathbb{R}^{K \times 3}$;

foreach model i **do**

  $w_i^{reg} \leftarrow \alpha_1 \cdot IoU(\mathcal{R}_i, \mathcal{R}_{global})$;

  $w_i^{lang} \leftarrow \alpha_2 \cdot BLEU(L_i, L_{target})$;

  $w_i^{temp} \leftarrow \alpha_3 \cdot exp(-\gamma|t - t_{current}|)$;

  $W[i,:] \leftarrow \left[w_i^{reg}, w_i^{lang}, w_i^{temp}\right]$;

Normalize rows:$\widetilde{W} \leftarrow Softmax(W, axis = 1)$;

Compute confidence scores:$c_i = \sum_{j=1}^{3} \widetilde{W}[i,j] \cdot Entropy(p_i)$;

Final weights:$\omega_i = \frac{exp(-c_i)}{\sum_k exp(-c_k)}$;

Return $\hat{y} = arg\,max\ \sum_{i=1}^{K} \omega_i p_i$;

2.4 Multi-Granularity Feature Engineering

The multi-granularity feature engineering phase enhances classification signals through deep information fusion. The semantic embedding concatenation operation is

defined as:

$$H_{\text{concat}} = \bigoplus_{i=1}^{3} h_i \oplus \bigoplus_{j=4}^{5} h_j^{\text{multi}} \tag{11}$$

forming a high-dimensional joint feature space. The decision signal encoding is:

$$o = [\mathbb{I}(\hat{y}_{vote} = E), \mathbb{I}(\hat{y}_{vote} = S), \mathbb{I}(\hat{y}_{vote} = G)] \tag{12}$$

which is concatenated with the semantic features to form the final feature vector:

$$X_{final} = H_{concat} \oplus O \tag{13}$$

Algorithm 3: Multi-Resolution Ensemble Learning with Stability-Weighted Voting

Input: Raw text T, Pre-trained models $M = \{M_1, M_2, \ldots, M_n\}$

Output: Final prediction $\hat{y}_{final}$

Preprocess text T to generate token sequences T_{proc} for each model M_i;
for *each model* M_i **do**
    Extract semantic embeddings H_i using the model's tokenizer and encoder;
    Compute probability output p_i using the composite transformation;
Evaluate model performance s_i using the hybrid metric;
Perform hierarchical decision fusion: **if** majority vote condition is met **then**
    $\hat{y}_{vote} \leftarrow$ majority vote result;
else
    Select top 3 models based on s_i;
    Compute stability weights w_i for the top 3 models;
    Calculate weighted probabilities $S(k)$ for each category k;
    $\hat{y}_{vote} \leftarrow \arg\max(S(k))$ for each category k;
Concatenate semantic embeddings and decision signals to form X_{final};
Train random forest classifier on X_{final};
$\hat{y}_{vote} \leftarrow$ random forest prediction;
return $\hat{y}_{final}$;

This fusion strategy effectively combines semantic information with decision preferences, enhancing the model's ability to distinguish between different classes. The specific implementation details of this feature fusion process are shown in Algorithm 3.

2.5 Random Forest Classifier

The random forest classifier is employed to enhance generalization ability through ensemble learning. Each decision tree utilizes the information gain criterion:

$$IG(v) = H(v) - \sum_{c\in\{leftright\}} \frac{D_t^c}{D_t} H(v_c) \tag{14}$$

where the entropy calculation is:

$$H(v) = -\sum_{k\in Y} p_v(k) \log p_v(k) \tag{15}$$

ensuring optimal partitioning of the feature space. Forest-level prediction is achieved through the average voting mechanism:

$$\hat{y}_{final} = arg\max_k \left(\frac{1}{T}\sum_{t=1}^{T} q_t(k)\right) \tag{16}$$

which effectively mitigates the overfitting risk associated with individual trees, thereby improving the model's robustness and accuracy. The training and prediction process of the random forest classifier is shown in Algorithm **3**.

2.6 Theoretical Formalization of ESG Discrepancy

Building upon the three-layer discrepancy framework, we establish a quantifiable ESG divergence metric:

$$\Delta_{ESG} = \prod_{i=1}^{3} \left(1 - \alpha_i \cdot IoU\left(\mathcal{R}_{local}^{(i)}, \mathcal{R}_{global}^{(i)}\right)\right) \tag{17}$$

where:

• $\alpha_i \epsilon [0,1]$denotes the regulatory alignment factor for layer i (lexical, contextual, systemic)

• $IoU(\cdot)$ computes the Jaccard index between local and global regulatory requirements:

$$IoU(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A}\cap\mathcal{B}|}{|\mathcal{A}\cup\mathcal{B}|} \tag{18}$$

• $\mathcal{R}_{local}^{(i)}$and $\mathcal{R}_{global}^{(i)}$represent the regulatory word sets for layer i

2.6.1 Cultural Entanglement Metric

To quantify cross-jurisdictional semantic conflicts, we define: $\|\emptyset_{EN}(d) -$

$\emptyset_{ISO}(d)\|$

$$CEM(D) = \frac{1}{|D|}\sum_{d\in D} \|\emptyset_{EN}(d) - \emptyset_{ISO}(d)\|_{\sum -1} \quad (19)$$

where:

• $\emptyset_{EN}(d)$ and $\emptyset_{ISO}(d)$ are semantic embeddings under EN 16627 and ISO 26000 standards

• $\|\cdot\|_{\sum -1}$ denotes the Mahalanobis distance with covariance matrix Σ estimated from 37 regulatory documents

• D represents the corpus spanning 17 jurisdictions in our dataset

2.7 Theoretical Guarantees

THEOREM 1 (**Generalization Bound via PAC-Bayes**). *Let $\mathcal{H}$ be the hypothesis* class *with VC*-dimension d, and D the data distribution. Applying Theorem 2.1 in Neyshabur et al. to our ensemble mechanism[33], for any $\delta > 0$, with probability at least $1-\delta$ over the sample $S\sim D^n$, the generalization error R(h) of our ensemble satisfies:

$$R(h) \leq \underbrace{\widehat{R}_s(h)}_{Empirical\ Risk} + \sqrt{\frac{d_{eff}(\ln\frac{2en}{d_{eff}}+1)+\ln\frac{1}{\delta}}{n}} \quad (20)$$

where the effective dimension $d_{eff} = \frac{d}{\sqrt{K}}$ is reduced through model diversification with K base models.

Our framework provides three fundamental guarantees that complement the PAC-Bayes bound:

- **Completeness:** The union of the support sets of all models covers the entire label space, i.e.,$\bigcup_{i=1}^{n} supp(p_i) = Y$.
- **Consistency:** As the number of models increases, the probability of the predicted label matching the true label approaches 1 almost surely, i.e., $\lim_{n\to\infty} P(\hat{y} = y_{true}) = 1 a.s.$
- **Robustness:** The framework is robust to perturbations, with the norm of the

gradient of the predicted label with respect to perturbations bounded by $\left\|\frac{\partial \hat{y}}{\partial \epsilon}\right\|_2 \leq C\sqrt{d}$.

2.8 Implementation Details

The implementation of the proposed framework meticulously follows the outlined theoretical framework and algorithmic steps (Algorithm 3). Each stage is designed to optimize performance while ensuring computational efficiency. The text preprocessing stage dynamically adjusts tokenization based on model-specific requirements. The semantic embedding extraction phase leverages multi-level encoding to capture rich semantic information. The hierarchical decision fusion mechanism combines majority voting and weighted fusion to ensure robust decision-making. Finally, the random forest classifier aggregates predictions to produce the final output, ensuring high accuracy and generalization ability.

3 Experiment

3.1 The HarmoniBERT Solution

3.1.1 Data Collection

The construction process of the corpus includes three phases: data collection, data preprocessing, and quality assessment. To ensure the industry representativeness and data quality of the samples, the selection criteria follow a dual dimension: according to the Bloomberg industry classification standard, select the top 20% listed companies by market capitalization in each sub-sector, and simultaneously combine the Refinitiv ESG rating system to screen out high-quality companies with an AAA rating, thereby constructing a benchmark corpus that encompasses both industry breadth and depth of ESG practice.

This study defines ESG-related text as an independent semantic unit containing at least one core ESG element (environmental, social, or governance dimension).We use manual screening to collect English news texts from two authoritative sources from 2024 to the present:

1) **Direct Disclosure Channels:**The ESG section of corporate websites and

annual sustainability reports;

2) Indirect Verification Channels: Including professional ESG media (ESG Today, ESG News) and international mainstream financial media (Reuters, Bloomberg Businessweek), and cross-verify data from multiple sources to ensure the objectivity and completeness of the information sources.

As illustrated in Figure **3**, a word cloud generated from the collected ESG-related texts is presented.

Figure 3 Word Cloud of Collected ESG-Related Texts

3.1.2 Data Preprocessing

To ensure the precision of text annotation, the annotation work combines manual and LLMs approaches. Among them, the LLMs uses ChatGPT for annotation, accounting for 20%.Manual annotation accounts for 80%. Two professional annotators who meet the annotation standards are selected through preliminary testing to participate in this task using a double-blind annotation method.The two annotators independently annotate the news texts.

The ESG classification system in this study adopts the three dimensions of a three-dimensional binary representation framework. We set up binary variables for the three ESG dimensions, i.e., marking their relevance with “0” and “1” based on whether the news text content involves the E, S, or G dimensions.

3.1.3 Quality Assessment

We use Cohen’s Kappa Coefficient to measure the consistency among annotators[34]:

$$K = \frac{P_0 - P_e}{1 - P_e} \tag{21}$$

Calculate the corresponding Kappa Coefficient scores for the ESG binary

classifications.The K values for the three dimensions are: 0.89 for the E dimension, 0.87 for the S dimension, and 0.84 for the G dimension, all reaching the "almost complete agreement" level[35].

3.2 Experiment Design

This study aims to treat various methods fairly, so conventional and effective models are selected to fairly demonstrate the advantages of this research, using conventional empirical parameters. Specifically, the traditional models selected are as follows: {Dictionary-based method: TF-IDF, Machine learning method: Logistic Regression, Ensemble learning: Random Forest, Neural network: Multi-layer Perceptron}. The BERT models selected are {bert-base-cased, bert-base-uncased, DeBERTa, DistilBERT, RoBERTa}, all using the base version. The ensemble models include Vote-Bert and HarmoniBERT.

Regarding the training part, the parameter space used in this study is shown in the Table 2 below. The traditional methods use TF-IDF for tokenization, and BERT uses the cross-entropy loss function with an Adam optimizer with weight decay. All experiments are conducted in a system environment of Ubuntu 22.04, using 8 NVIDIA 3090 GPUs and PyTorch.Regarding the evaluation part, the experiments use four metrics to assess model performance: Accuracy, Precision, Recall, and F1-score. Additionally, binning is used with bin=50, to obtain the mean and standard deviation to evaluate the robustness of different models.

Table 2 Parameters for models

TF-IDF					**BERT**			
	Dictionary Method	Machine Learning	Ensemble Learning	Neural Network	Standard BERT	Vote-BERT		HarmoniBERT
n_keywords	50	50	50	50	batch_size	512	512	512
max_features	1000	1000	1000	1000	num_epochs	5	5	5
max_iter	\	1000	\	100	max_length	64	64	64
n_estimators	\	\	50	\	n_estimators	\	\	50
lr	\	\	\	2.00 × 10−5	lr	2.00 × 10−5	2.00 × 10−5	2.00 × 10−5

3.3 Experiment Analysis

Table 3 HarmoniBERT Classification results with tranditional models

Model	Dict	ML	EL	MLP	Deepseek	HarmoniBERT
E_Accuracy	65.09 (±18.91)	92.65 (±16.65)	98.27 (±6.27)	93.93 (±11.93)	93.32 (±10.32)	99.78 (±3.78)
E_Precision	57.98 (±24.88)	95.28 (±27.43)	98.56 (±11.61)	95.01 (±18.09)	92.76 (±13.39)	99.92 (±4.92)
E_Recall	89.23 (±22.57)	88.55 (±18.98)	97.68 (±12.68)	91.70 (±25.03)	93.60 (±14.89)	99.63 (±8.72)
E_F1	69.87 (±19.87)	91.62 (±15.62)	98.08 (±8.61)	93.17 (±15.75)	93.18 (±12.34)	99.77 (±4.53)
S_Accuracy	53.45 (±23.45)	87.81 (±13.81)	97.14 (±11.14)	89.70 (±15.70)	92.10 (±13.10)	99.66 (±3.66)
S_Precision	48.01 (±33.73)	89.47 (±22.80)	98.14 (±13.93)	89.58 (±22.91)	91.51 (±20.25)	99.80 (±6.94)
S_Recall	80.55 (±30.55)	81.81 (±26.26)	95.23 (±22.15)	86.37 (±31.37)	91.91 (±17.09)	99.38 (±8.91)
S_F1	59.71 (±36.98)	85.14 (±18.47)	96.58 (±12.13)	87.72 (±22.86)	91.71 (±14.92)	99.58 (±5.46)
G_Accuracy	43.81 (±16.19)	86.21 (±16.21)	96.69 (±8.69)	88.50 (±14.50)	89.86 (±10.86)	99.65 (±3.65)
G_Precision	29.70 (±21.37)	84.29 (±50.96)	96.92 (±21.92)	80.89 (±60.89)	84.62 (±48.76)	99.73 (±9.73)
G_Recall	85.86 (±29.61)	57.26 (±40.60)	90.20 (±36.36)	72.08 (±59.58)	85.66 (±38.12)	98.96 (±17.15)
G_F1	43.67 (±29.04)	67.13 (±44.91)	93.15 (±23.15)	75.41 (±60.03)	85.14 (±41.73)	99.32 (±9.32)

As shown in Table 3, the proposed HarmoniBERT model outperforms traditional dictionary-based methods, conventional machine learning models, and existing LLMs in classification tasks. In the E, S, and G tasks, HarmoniBERT achieves average accuracy rates of 99.78%, 99.66%, and 99.65%, respectively, with performance stability far exceeding other models. Compared to the accuracy of the dictionary-based method of 65. 09%, HarmoniBERT demonstrates an improvement of more than 34 % , while its F1 score reaches 99. 77%, much higher than the dictionary

method's 69.87%. Compared to conventional machine learning models, HarmoniBERT not only maintains high precision but also significantly improves recall. For example, in the environmental label classification task, HarmoniBERT achieves a recall rate of 99.63%, while the MLP model only reaches 91.70%, proving its stronger robustness to class imbalance. Furthermore, compared to the 671B parameter Deepseek model, HarmoniBERT achieves small improvements in all metrics while significantly reducing computational cost consumption.

Table 4 HarmoniBERT Classification results with various BERT models

Model	BERT1	BERT2	BERT3	BERT4	BERT5	Vote-BERT	HarmoniBERT
E_Accuracy	94.60 (±12.60)	95.00 (±9.00)	93.32 (±11.32)	96.76 (±8.76)	96.76 (±8.76)	95.47 (±11.47)	99.78 (±3.78)
E_Precision	97.30 (±18.35)	95.94 (±16.78)	94.43 (±15.27)	96.68 (±14.07)	96.68 (±14.86)	96.18 (±14.93)	99.92 (±4.92)
E_Recall	90.81 (±19.98)	93.22 (±18.22)	90.99 (±24.32)	96.28 (±14.46)	96.30 (±12.97)	93.90 (±22.48)	99.63 (±8.72)
E_F1	93.81 (±14.74)	94.44 (±11.68)	92.53 (±15.11)	96.41 (±10.69)	96.42 (±9.24)	94.92 (±13.84)	99.77 (±4.53)
S_Accuracy	84.53 (±14.53)	89.96 (±13.96)	83.32 (±15.32)	94.48 (±10.48)	94.48 (±10.48)	91.05 (±13.05)	99.66 (±3.66)
S_Precision	86.11 (±32.26)	91.83 (±23.41)	84.79 (±30.25)	95.68 (±17.90)	95.54 (±23.31)	93.43 (±20.70)	99.80 (±6.94)
S_Recall	77.11 (±27.11)	84.55 (±26.66)	75.27 (±35.27)	91.61 (±21.02)	91.59 (±21.00)	85.32 (±26.49)	99.38 (±8.91)
S_F1	80.97 (±21.51)	87.78 (±19.36)	79.31 (±33.16)	93.45 (±18.45)	93.38 (±13.38)	88.96 (±20.39)	99.58 (±5.46)
G_Accuracy	82.47 (±18.47)	87.40 (±13.40)	81.57 (±17.57)	90.65 (±14.65)	90.66 (±10.66)	88.26 (±14.26)	99.65 (±3.65)
G_Precision	71.14 (±58.64)	80.90 (±43.40)	72.98 (±47.98)	84.97 (±34.97)	84.77 (±30.93)	84.22 (±39.78)	99.73 (±9.73)
G_Recall	55.05 (±44.95)	67.35 (±38.77)	46.67 (±40.00)	77.96 (±47.19)	77.99 (±35.13)	67.36 (±34.03)	98.96 (±17.15)
G_F1	61.00 (±48.50)	72.46 (±34.96)	55.66 (±41.37)	80.60 (±40.60)	80.56 (±30.56)	74.01 (±31.16)	99.32 (±9.32)

A further comparison of HarmoniBERT with BERT series variants and hard

voting ensemble models shows that HarmoniBERT demonstrates stable and significant performance advantages, as shown in Table 4. For instance, in the environmental label classification task, HarmoniBERT's accuracy is 5.18 % higher than Bert-base-cased and 3.02 % higher than Roberta-base, while its F1 score also significantly outperforms Vote-BERT's 94.92%. As shown in Figure 4 traditional hard voting strategies tend to blur the classification boundaries due to the high consistency in predictions on misclassified samples among BERT variants, which reduces the robustness of the ensemble. In contrast, HarmoniBERT mitigates the overfitting risk of a single deep learning model by introducing Random Forest as a heterogeneous base classifier. The strong ability of Random Forest to capture local textual features complements BERT's contextual semantic representation, thus inhibiting the accumulation of errors during the ensemble process. For example, in the social label classification task, HarmoniBERT achieves a recall rate of 99.38%, while Vote-BERT only reaches 85.32%. Additionally, HarmoniBERT adopts a dynamic weight allocation mechanism, which adaptively adjusts the voting weight between BERT and Random Forest based on validation set performance, further improving the model's ability to discriminate complex semantics. In the governance label classification task, HarmoniBERT's precision reaches 99.73%, far surpassing the 71.14% to 84.97% of individual BERT models and variants, fully demonstrating the effectiveness of this method.

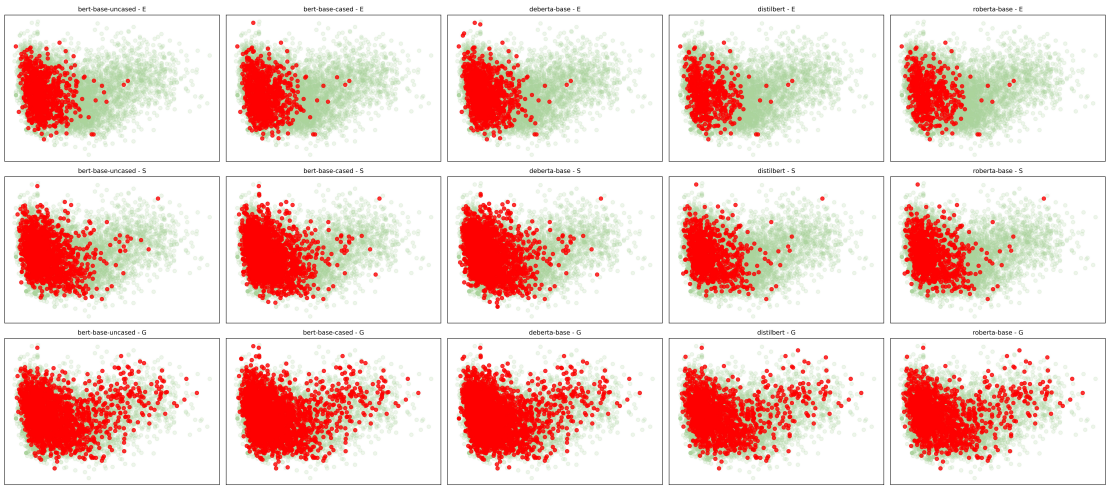

Figure 4 Visualization of ESG Dimension Predictions

3.4 Ablation experiment

To assess the generalizability of HarmoniBERT across diverse linguistic and cultural contexts, we conduct comprehensive evaluations across three dimensions of ESG disclosure heterogeneity:

- **Regulatory Variance:** UK (ESG-FTSE)[36] vs. Germany (Nano-ESG)[37] vs. US (Primary Dataset)
- **Textual Complexity:** Short (avg. 128 tokens), Medium (512 tokens), Long (1024+ tokens)
- **Concept Drift:** Temporal coverage from 2018-2023 across datasets

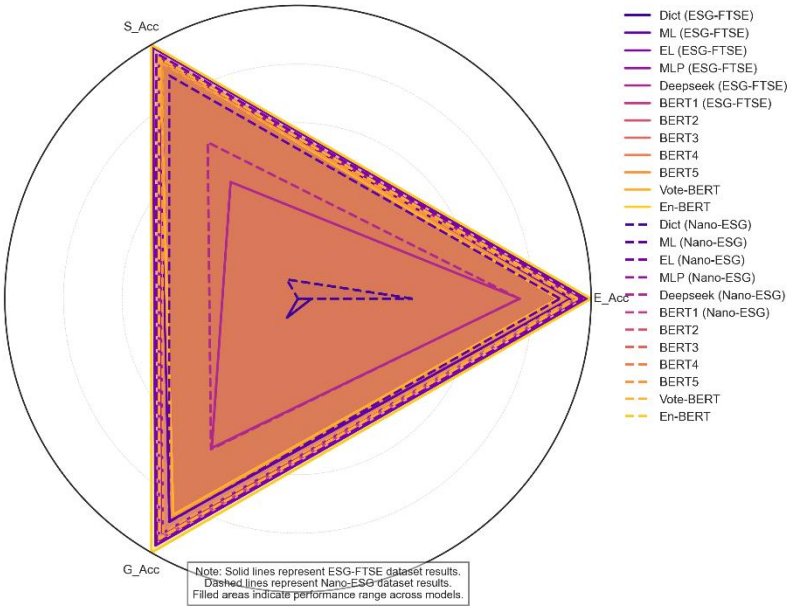

Figure 5 Performance Comparison Across Models on ESG Datasets

The experimental design was further enriched by the varying text lengths within these datasets—spanning short, medium, and long-form texts. This variation allowed us to rigorously evaluate HarmoniBERT's adaptability to different textual complexities, a critical capability for global ESG analysis. Following Xue et al.[38], Gupta and MALHOTRA[39],we implement stratified 5-fold cross-validation with dual test sets: (1) In-Country (IC) and (2) Cross-Country (CC), the latter containing 30% non-overlapping entities to prevent data leakage. All metrics report macro-averaged F1 scores across 10 independent runs.As illustrated in Figure 5, HarmoniBERT demonstrated superior accuracy and lower error rates across all categories compared

to alternative models, including both traditional machine learning approaches and single BERT architectures.

Table 5 Accuracy Comparison of HarmoniBERT with Baseline Models on ESG-FTSE and Nano-ESG Datasets

Model	ESG-FTSE			Nano-ESG		
	E_Acc	**S_Acc**	**G_Acc**	**E_Acc**	**S_Acc**	**G_Acc**
Dict	12.91 (±11.09)	8.65 (±11.35)	15.52 (±28.92)	55.67 (±24.33)	32.61 (±22.61)	27.10 (±22.90)
ML	93.67 (±9.67)	95.01 (±7.01)	88.73 (±12.73)	92.26 (±14.26)	91.14 (±17.14))	91.19 (±13.19)
EL	98.71 (±4.71)	98.91 (±21.13)	97.44 (±9.44)	97.98 (±13.98)	97.57 (±9.57)	97.44 (±7.44)
MLP	96.37 (±7.48)	95.37 (±7.37)	92.97 (±10.97)	96.17 (±14.17)	96.06 (±10.06)	96.31 (±10.31)
Deepseek	77.77 (±22.23)	50.62 (±25.62)	62.34 (±16.34)	82.14 (±18.64)	71.98 (±22.14)	70.34 (±12.34)
BERT1	89.44 (±15.44)	94.37 (±8.37)	86.05 (±16.05)	96.54 (±10.54)	96.35 (±10.35)	96.29 (±10.29)
BERT2	89.21 (±11.43)	94.37 (±10.37)	86.28 (±18.28)	96.50 (±10.50)	96.15 (±14.15)	95.94 (±13.94)
BERT3	89.33 (±9.33)	94.48 (±8.48)	86.05 (±30.49)	96.59 (±8.59)	96.24 (±12.24)	96.12 (±12.12)
BERT4	91.54 (±9.54)	94.46 (±6.46)	86.95 (±20.28)	95.65 (±13.65)	94.60 (±12.60)	94.17 (±12.17)
BERT5	91.43 (±9.43)	94.46 (±14.46)	87.18 (±17.18)	95.65 (±13.65)	94.60 (±14.60))	94.17 (±12.17)
Vote-BERT	89.44 (±11.44)	94.48 (±6.48)	86.23 (±11.77)	96.65 (±10.65)	96.24 (±12.24)	95.98 (±11.98)
HarmoniBERT	99.95 (±1.95)	99.97 (±1.97)	100.00 (±0.00)	99.98 (±1.98)	99.99 (±1.99)	99.98 (±1.98)

As shown in Table 5, the results underscore the enhanced performance of HarmoniBERT, which leverages the combined strengths of six pre-trained models through its hierarchical voting mechanism. This approach effectively addresses the semantic, contextual, and systemic challenges inherent in cross-cultural ESG annotation. Compared to single BERT models and traditional machine learning

approaches, HarmoniBERT demonstrates superior accuracy and consistency, with near-perfect classification rates across all ESG dimensions in both datasets. This performance validates the efficacy of our ensemble framework in unifying ESG assessment standards across linguistic and cultural boundaries, offering a significant advancement for global sustainability governance and financial market transparency.

4 Further Analysis

4.1 The HarmoniBERT Solution

In the evaluation of corporate performance, E, S, and G factors have emerged as essential indicators for measuring a firm's sustainability and long-term competitiveness. Our study aims to analyze the influence of specific events on corporate ESG ratings and to validate the effectiveness of an improved Ensemble BERT Random Forest model in enhancing the accuracy of ESG textual assessments.

4.1.1 Limitations of Conventional ESG Assessments

Traditional ESG rating systems quantify a firm's performance in environmental protection, social responsibility, and corporate governance to provide decision support for investors. However, existing evaluation methods exhibit the following limitations:

1) Reliance on single-source data, which makes it difficult to capture the dynamic changes within a firm;

2) Inconsistent evaluation criteria, resulting in a pronounced subjectivity in the ratings;

3) An inability to promptly capture significant events that affect a firm's ESG performance.

To address these issues, this study proposes an enhanced approach based on an Ensemble BERT Random Forest model. The method leverages machine learning to integrate multi-source data and dynamic evaluation metrics, thereby improving both the precision and timeliness of ESG ratings.

4.1.2 Case Study: Changes in Apple Inc.'s ESG Ratings

Taking Apple Inc. as an illustrative example, our study investigates the fluctuations in its ESG ratings from March 2024 to March 2025, with a focus on

examining the impact of major events on ratings across different dimensions. Figure **6** presents the predictive outcomes of three distinct models:

- The Final model, which corresponds to the Ensemble BERT Random Forest model;
- The Simple BERT model;
- The Ensemble BERT model.

Comparative analysis reveals that the Ensemble BERT model exhibits higher accuracy and stability.

In the Figure **7**, E, S, and G denote the ESG ratings corresponding to the Environmental, Social, and Governance dimensions, respectively. Under the influence of specific events, Apple Inc.'s ESG ratings exhibited significant fluctuations at various time points. For example:

- In May 2024, Apple Inc. announced initiatives to reduce greenhouse gas emissions, markedly enhancing its environmental rating;
- In November 2024, by endorsing responsible AI guidelines, the company achieved an improved governance rating;
- In January 2025, due to litigation in the Democratic Republic of Congo concerning conflict minerals, its social rating experienced a decline.

4.2 Visualization Analysis of the Relationship between ESG Scores and Stock Returns

To further illustrate the differences in the explanatory power of various ESG measurement methods with respect to stock returns, our paper presents scatter plots of ESG scores across the three dimensions (E/S/G) versus individual stock returns, as shown in Figure 6. In the figure, different colored points correspond to ESG scores derived from distinct sources:

- Blue points represent conventional ESG ratings (Final);
- Red points denote outputs from the simple BERT model (Simple_Bert_predict);
- Green points correspond to the Ensemble BERT + Random Forest model (Ensem-ble_Bert_predict), i.e., the improved model proposed in this study.

The figure yields the following key observations:

1) Wider Coverage of Stock Returns: The green points display a noticeably more dispersed dis¬tribution along the horizontal axis (Return) compared to the red and blue points, especially covering a broader range of extreme return values. This indicates that the integrated model has enhanced capability in identifying high-volatility stocks.

2) Enhanced Rating Differentiation: Along the vertical axis (ESG Score), the green points exhibit a more polarized distribution, particularly in the Environmental dimension, demonstrating that the model possesses a significantly higher discriminatory power in evaluating ESG perfor¬mance.

3) Clearer Point Cloud Structure: For the Social and Governance dimensions, the green points show moderate density and well-concentrated clustering, thereby avoiding the "grouping" phe-nomenon observed in other models. This reflects improvements in the model's semantic under-standing and output stability.

4) Better Alignment with Regression Trends: Although an explicit regression curve is not drawn, the distribution of points suggests that the green points are closer to the central trend line across all dimensions, signifying a closer linear relationship between the model's outputs and actual returns.

4.3 Empirical Analysis

To evaluate the effectiveness of the improved ESG textual measurement model in explaining stock returns, this study performs comparative regression analyzes based on the output of the constructed BERT model and conventional ESG rating results. Specifically, different measurement methods across the three ESG dimensions are used as independent variables, with stock returns serving as the dependent variable. A series of Ordinary Least Squares (OLS) regression models are constructed, and robust standard errors—consistent with heteroscedasticity and autocorrelation (HAC)—are employed to mitigate potential heteroscedasticity and autocorrelation issues in the time-series data.

4.3.1 Model Configuration and Variable Description

The independent variables include the following three ESG measurement

methods:

- **Simple_Bert_predict:** ESG textual ratings directly produced by the basic BERT model;
- **Ensemble_Bert_predict:** ESG textual ratings obtained by integrating the basic BERT outputs with a Random Forest approach, representing the improved model proposed in this study;
- **Final:** Conventional ESG ratings available in the data source, which serve as the baseline for comparison.

The dependent variable is the individual stock return corresponding to the respective time period, used to assess the explanatory power of the various ESG measurement methods with respect to market performance.

The general form of the regression model is given by:

Returnt=α+β·ESGi,t+εt#22

where\ {ESG}_{i,t} denotes the ESG rating under a specific method, {Return}_t represents the stock return for the period, and \varepsilon_t is the error term.

4.3.2 Analysis of Regression Results

Table 6 summarizes the regression R^2 values obtained from the Simple BERT and Ensemble BERT models across different ESG dimensions:

Table 6 Comparison of Regression R^2 Values across ESG Dimensions (Unit: %)

Dimension	Regression Method	Average R^2 Value
E	Simple_bert_predict_E	1.80
	Ensemble_bert_predict_E	**2.50**
S	Simple_bert_predict_S	1.00
	Ensemble_bert_predict_S	**1.20**
G	Simple_bert_predict_G	0.70
	Ensemble_bert_predict_G	**0.90**

As evident from the table, the **Ensemble_bert_predict model outperforms the basic BERT model across all dimensions**. This indicates that integrating the Random Forest method can effectively enhance the precision of ESG textual ratings and strengthen their ability to explain market performance. Although the overall

R^2 values are relatively modest, this is consistent with the inherently weak economic rationale linking ESG information with short-term returns, thereby demonstrating that the model partially captures the connection between non-financial information in corporate texts and market reactions.

4.3.3 Result

The empirical results validate the effectiveness of the proposed **BERT + Random Forest ensemble model for ESG textual measurement**, with the most significant improvement observed in the Environmental dimension. This suggests that the incorporation of semantic analysis with nonlinear ensemble algorithms can enhance the identification of corporate performance in the construction of ESG rating systems. Future research may further incorporate cross-sectional control variables and industry-fixed effects to enhance the robustness and generalizability of the model.

5 Related Works

5.1 Evolution of ESG Text Annotation

Early ESG text annotation primarily relied on lexicon-based methods, where researchers constructed ESG-related word lists or keyword rules to identify sustainability-related content in corporate annual reports or CSR disclosures[21]. To enhance vocabulary-level analysis, various ESG-specific dictionaries were developed, often categorized by Environmental (E), Social (S), and Governance (G) dimensions, and widely applied in automated annotation tasks. One representative method is the unigram-based ESG dictionary proposed by Baier et al., which performs ESG content extraction via keyword matching[40]. However, such methods rely heavily on expert curation and manual scoring, leading to delayed updates, non-transparent standards, and high subjectivity. These limitations make traditional lexicon-based approaches prone to semantic ambiguity, polysemy, and domain shifts, ultimately affecting the coverage and accuracy of annotations[24].

As the limitations of lexicon-based approaches became increasingly evident, researchers began exploring traditional machine learning methods for ESG text annotation. A typical approach is to transform text into feature vectors (e.g., TF-IDF

matrices or sentiment word counts), then apply classifiers such as support vector machines (SVMs) or random forests for ESG category or sentiment labeling[41]. Compared to rule-based methods, machine learning models can leverage data-driven learning to improve annotation accuracy. However, these models still rely on manually engineered features and struggle to capture deep semantic structures. Their performance remains suboptimal in complex corpora with high polysemy or industry-specific expressions.

Building on traditional machine learning, ensemble learning has been introduced as a strategy to enhance model stability and generalization in ESG annotation tasks[44]. Ensemble methods combine multiple base learners (such as decision trees, SVMs, or naive Bayes classifiers) at either the model level or the output level to mitigate the bias and variance issues of single models. Luo et al. compared SVMs[43], logistic regression, and random forest models for detecting exaggerated or misleading expressions in ESG reports, and found that the Bagging-based random forest model performed best when dealing with redundant or semantically complex ESG text. Other studies have adopted voting or weighted fusion approaches to combine traditional classifiers for multi-label ESG risk classification in news content[44]. These ensemble techniques not only alleviate overfitting and sparsity issues in traditional models but also laid the groundwork for integrating deep learning and pretrained language models.

The adoption of deep learning brought a new phase to ESG text annotation. Models such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) began to be applied to ESG-related text classification and sentiment analysis tasks[45]. Compared to conventional models, deep neural networks can automatically learn feature representations from large-scale corpora, reducing the need for manual feature design. Studies have shown that multi-layer perceptrons (MLPs), CNNs, and RNNs outperform traditional classifiers like naive Bayes and SVMs in text classification tasks[46]. CNNs can capture local textual patterns to detect ESG polarity, while LSTMs are effective in modeling context-dependent meanings. Although deep learning significantly improved ESG annotation effectiveness, it still

faces limitations in consistency and robustness for practical applications.

LLMs are now transforming the technical paradigm of ESG text annotation. Compared with traditional methods, LLMs leverage large-scale pretraining to achieve stronger semantic understanding and contextual modeling capabilities, showing excellent performance in classification and information extraction tasks[47]. Domain-specific BERT variants such as FinBERT[28] have been fine-tuned on financial and sustainability texts, significantly improving the ability to identify environment, social, and governance content in ESG documents. Among them, models like FinBERT-ESG[48] support sentence-level ESG classification and have become early influential benchmark tools. These models outperform traditional methods in terms of accuracy and generalization, and better handle ambiguous ESG expressions and terminological diversity, gradually becoming a cornerstone of automated ESG annotation systems. However, LLMs still face significant challenges in real-world applications. On one hand, these models often contain billions of parameters, leading to long training and inference times and high computational resource consumption. Simply loading such models often requires high-performance GPUs, limiting deployment in small and medium-scale environments[23]. On the other hand, due to the lack of fine-grained supervision, LLMs are prone to hallucination in structured annotation tasks, resulting in inaccurate or inconsistent outputs[25]. Furthermore, despite expanded context windows, current LLMs still struggle to maintain global coherence across long and multi-paragraph texts. In ESG applications, this leads to difficulties in entity boundary control and semantic consistency. In contrast, BERT and its variants offer a simpler architecture, moderate parameter size, mature fine-tuning mechanisms, and high inference efficiency, making them the preferred choice for ESG annotation tasks that demand a balance of performance and cost-effectiveness.

5.2 Limitations of Current BERT Baselines

BERT models are typically used in text annotation tasks through an “encoder + classification head” architecture, where the input is encoded via a bidirectional Transformer and then passed to a softmax or CRF layer for token-level or

sequence-level labeling[49]. As a representative pretrained language model, BERT achieved breakthrough performance across many NLP tasks[50], yet as a single model, it still faces multiple limitations in ESG text annotation. Empirical studies have shown that in sentence-level ESG labeling tasks, single BERT models often suffer from label bias. For example, in FinBERT-ESG experiments, the model tended to classify neutral sentences as "Governance" due to the frequent appearance of governance-related language in corporate reports. Furthermore, single BERT models often exhibit instability across different training and test sets, showing high sensitivity to hyperparameter configurations. These issues are especially prominent in ESG tasks, which are typically low-resource and highly context-dependent, leading to poor generalization across companies and industries.

To address these problems, researchers have proposed various structural improvements to BERT. One direction is RoBERTa, which improves the robustness of the model and the representational power by increasing the size of the pre-training corpus, extending the training time and removing the prediction objective of the next sentence. Another direction is BART, which combines BERT's encoder with GPT-style decoders and is widely used in text summarization, question answering, and generative annotation tasks[51]. Notably, recent studies show that BERT-based models generally outperform LLMs in classification tasks, especially under conditions of limited labels or strict precision requirements, while LLMs demonstrate stronger robustness to textual perturbations and ambiguity. Moreover, compared to weakly supervised methods, AI-generated labels with human supervision have shown better classification outcomes, emphasizing the importance of integrating AI annotations with human oversight[52]. This insight is particularly valuable for ESG tasks, which require high-quality annotation standards. Despite these structural enhancements, BERT still faces issues of inconsistent output and label bias in ESG applications. To tackle this, this study integrates multiple pretrained BERT models and proposes an ensemble BERT architecture that combines multiple outputs, aiming to provide a more practical and accurate solution for ESG text analysis.

5.3 Application of Ensemble Learning in Text Annotation

Ensemble learning, a classical strategy to improve model stability and predictive accuracy, has gained increasing attention in natural language processing, especially in text annotation tasks. Its core idea is to combine outputs from multiple base models to compensate for the limitations of a single model in terms of generalization, robustness, or local errors. Common ensemble techniques include Bagging (e.g., Random Forest), Boosting (e.g., AdaBoost, XGBoost), Voting, and Stacking. Several studies have successfully applied these methods to ESG-related tasks. For instance, Luo et al. used a Random Forest model to classify the sentiment of corporate ESG disclosures, and found that it was more robust than traditional SVMs in sparse feature scenarios[43].

In named entity recognition and sentiment classification tasks, ensemble methods have also proven effective. For example, Lee et al. combined BERT and ALBERT in a voting-based ensemble model for ESG text classification, achieving an accuracy of 80.79% with a batch size of 20[54]. Ensemble learning shows clear advantages in ESG annotation. On one hand, different BERT variants (such as FinBERT and RoBERTa) perform variably across different domain subsets, and integrating them enables information complementarity[53]. On the other hand, ESG annotation involves significant subjectivity, and predictions from a single model may deviate from human judgment. Aggregating predictions from multiple models can improve the interpretability and consistency of the results[55]. Therefore, in light of the limitations of current BERT baselines, introducing ensemble strategies is not only a reliable way to enhance annotation quality, but also provides a methodological foundation for building more trustworthy and generalizable ESG information extraction systems. This paper will further validate the effectiveness of the proposed ensemble BERT method through empirical comparison.

6 Conclusion

This study presented HarmoniBERT, an automated ESG text labeling tool that is precise, efficient, and cost-effective. First, we meticulously constructed a high-quality ESG text dataset, with every step—from the selection of representative companies and data collection to manual annotation and subsequent quality control—strictly

supervised by an expert team. This approach ensured that the dataset offers substantial breadth as well as rigor in quality. Based on ensemble learning theory, we develop an ensemble model combining multi-feature fusion BERT and data perturbation. Experimental results confirmed that HarmoniBERT delivers high accuracy and efficiency in ESG text labeling and further demonstrated the model's effectiveness in assessing ESG communications. In sum, this study provides an innovative yet practical tool for automated ESG text labeling, thereby opening new avenues for in-depth corporate ESG analysis and laying a solid theoretical and practical foundation for advancing green investment, enhancing regulatory measures, and promoting sustainable societal development. Future work may explore the fusion of heterogeneous data sources, the enhancement of model robustness, and the application of this approach to other domains, with the aim of continuously refining ESG evaluation systems and extending their applicability.

References

[1] Dai T, Tang C (2022) Frontiers in service science: integrating esg measures and supply chain management: research opportunities in the postpandemic era. Service Science 14(1):1 – 12.

[2] Lee MT, Raschke RL, Krishen AS (2022) Signaling green! firm esg signals in an interconnected environment that promote brand valuation. *Journal of Business Research* 138:1 – 11.

[3] Li J, Lian G, Xu A (2023) How do esg affect the spillover of green innovation among peer firms? mechanism discussion and performance study. *Journal of Business Research* 158:113648.

[4] Iurkov V, Koval M, Misra S, Pedada K, Sinha A (2024) Impact of esg distinctiveness in alliances on shareholder value. *Journal of business research* 171:114395.

[5] Dong MC, Huang Q, Liu Z (2022) Adjusting supply chain involvement in countries with politician turnover: A contingency framework. *Journal of Operations Management* 68(8):824 – 854.

[6] Tong X, Linderman K, Zhu Q (2023) Managing a portfolio of environmental projects: Focus, balance, and environmental management capabilities. *Journal of Operations Management* 69(1):127 – 158.

[7] Martiny A, Taglialatela J, Testa F, Iraldo F (2024) Determinants of environmental social and governance (esg) performance: A systematic literature review. *Journal of Cleaner Production* 456:142213.

[8] Calamai T, Balalau O, Guenedal TL, Suchanek FM (2025) Corporate greenwashing detection in text-a survey. *arXiv preprint arXiv:2502.07541* .

[9] Li TT, Wang K, Sueyoshi T, Wang DD (2021) Esg: Research progress and future prospects. *Sustainability* 13(21):11663.

[10] Rau PR, Yu T (2024) A survey on esg: investors, institutions and firms. *China Finance Review International* 14(1):3 – 33.

[11] Bag S, Rahman MS, Choi TM, Srivastava G, Kilbourn P, Pisa N (2023) How

covid-19 pandemic has shaped buyersupplier relationships in engineering companies with ethical perception considerations: A multi-methodological study. *Journal of Business Research* 158:113598.

[12] Xia L, Yang M, Liu Q (2024) Using pre-trained language model for accurate esg prediction. *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, 1 – 22.

[13] Elamer AA, Boulhaga M (2024) Esg controversies and corporate performance: The moderating effect of governance mechanisms and esg practices. *Corporate Social Responsibility and Environmental Management* 31(4):3312 – 3327.

[14] Pavlova M, Casey B, Wang M (2024) ESG-FTSE: A corpus of news articles with ESG relevance labels and use cases. *CoRR* abs/2405.20218, URL http://dx.doi.org/10.48550/ARXIV.2405.20218.

[15] Fischbach J, Adam M, Dzhagatspanyan V, Mendez D, Frattini J, Kosenkov O, Elahidoost P (2023) Automatic esg assessment of companies by mining and evaluating media coverage data: Nlp approach and tool. *2023 IEEE International Conference on Big Data (BigData)*, 2823 – 2830 (IEEE).

[16] Bluteau K, Coggins F, Koumou GB (2025) Enhancing esg news annotation: Leveraging gpt for the analysis of esg news and events. *Available at SSRN 5128896* .

[17] Ziegler GG (2024) Automating information extraction from financial reports using llms—a .

[18] Nie C, Luo W, Chen Z, Feng Y (2025) Intellectual property protection and corporate ESG performance: evidence from a quasi-natural experiment in china. *Bus. Process. Manag. J.* 31(1):245 – 266, URL http://dx.doi.org/10.1108/BPMJ-01-2024-0041.

[19] Kannan N, Seki Y (2023) Textual evidence extraction for esg scores. *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, 45 – 54.

[20] *Asif M, Searcy C, Castka P (2023) Esg and industry 5.0: The role of technologies in enhancing esg disclosure. Technological Forecasting and Social Change*

195:122806.

[21] *Ong K, Mao R, Xing F, Satapathy R, Sulaeman J, Cambria E, Mengaldo G (2025) Esgsenticnet: A neurosymbolic knowledge base for corporate sustainability analysis. arXiv preprint arXiv:2501.15720* .

[22] Matarazzo A, Torlone R (2025) A survey on large language models with some insights on their capabilities and limitations. *arXiv preprint arXiv:2501.04040* .

[23] Lin H, Zhang Y (2025) The risks of using large language models for text annotation in social science research. URL https://arxiv.org/abs/2503.22040.

[24] *Liu Y, Li Y, Chen H (2025) The keywords in corporate social responsibility: A dictionary construction method based on mnir. Sustainability 17(6):2528.*

[25] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, et al. (2025) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43(2):1 – 55.

[26] Barros V, Matos PV, Sarmento JM, Vieira PR (2024) Esg performance and firms' business and geographical diversification: An empirical approach. *Journal of Business Research* 172:114392.

[27] Long X, Zhuang L, Li A, Wei J, Li H, Wang S (2024) KGDM: A diffusion model to capture multiple relation semantics for knowledge graph embedding. *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 8850 – 8858 (AAAI Press),URL http://dx.doi.org/10.1609/AAAI.V38I8.28732.

[28] Liu Z, Huang D, Huang K, Li Z, Zhao J (2020a) Finbert: A pre-trained financial language representation model for financial text mining. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 4513 – 4519 (ijcai.org), URL http://dx.doi.org/10.24963/IJCAI.2020/622.

[29] Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108, URL

http://arxiv.org/abs/1910.01108.

[30] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019a) Roberta: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692, URL http://arxiv.org/abs/1907.11692.

[31] Liu Z, Lin W, Shi Y, Zhao J (2021) A robustly optimized BERT pre-training approach with post-training. *Chinese Computational Linguistics - 20th China National Conference, CCL 2021, Hohhot, China, August 13-15, 2021, Proceedings*, volume 12869 of *Lecture Notes in Computer Science*, 471 – 484, URL https://doi.org/10.1007/978-3-030-84186-7_31.

[32] Fu C, Qian F, Su K, Su Y, Wang Z, Shi J, Liu Z, Liu C, Ishi CT (2025) Himul-lgg: A hierarchical decision fusionbased local-global graph neural network for multimodal emotion recognition in conversation. *Neural Networks* 181:106764, URL http://dx.doi.org/10.1016/J.NEUNET.2024.106764.

[33] Neyshabur B, Bhojanapalli S, Srebro N (2018) A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. 6th International Conference on Learning Representations, ICLR 2018, Vancouver,BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings (OpenReview.net), URL https://openreview.net/forum?id=Skz_WfbCZ.

[34] Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37 – 46, URL http://dx.doi.org/10.1177/001316446002000104.

[35] Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159 – 174,URL http://dx.doi.org/10.2307/2529310.

[36] Pavlova M, Casey B, Wang M (2024) ESG-FTSE: A corpus of news articles with ESG relevance labels and use cases. CoRR abs/2405.20218, URL http://dx.doi.org/10.48550/ARXIV.2405.20218.

[37] Billert F, Conrad S (2025) Nano-esg: Extracting corporate sustainability information from news articles. Hauff C, Macdonald C, Jannach D, Kazai G, Nardini FM, Pinelli F, Silvestri F, Tonellotto N, eds., *Advances in Information*

Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, LuccaItaly, April 6-10, 2025, Proceedings, Part IV, volume 15575 of *Lecture Notes in Computer Science*, 324 – 338 (Springer).

[38] Xue Q, Jin Y, Zhang C (2024) Esg rating results and corporate total factor productivity. *International Review of Financial Analysis* 95:103381, ISSN 1057-5219, URL http://dx.doi.org/https://doi.org/10.1016/j.irfa.2024.103381.

[39] *Gupta R, MALHOTRA DN (2025) Esg and corporate financial performance: A comprehensive review of research and emerging trends. Available at SSRN5131292 .*

[40] *Baier P, Berninger M, Kiesel F (2020) Environmental, social and governance reporting in annual reports: A textual analysis. Financial Markets, Institutions & Instruments 29(3):93 – 118.*

[41] *D ' Amato V, D ' Ecclesia R, Levantesi S (2022) Esg score prediction through random forest algorithm. Computational Management Science 19(2):347 – 373.*

[42] *Chowdhury MAF, Abdullah M, Azad MAK, Sulong Z, Islam MN (2023) Environmental, social and governance (esg) rating prediction using machine learning approaches. Annals of Operations Research 1 – 25.*

[43] *Luo Y, Cui X, Liu Q, Zhou Q, Zhang Y (2024) Identifying exaggeration in esg reports using machine learning techniques. Data and Information Management 100084.*

[44] *Krappel T, Bogun A, Borth D (2021) Heterogeneous ensemble for esg ratings prediction. arXiv preprint arXiv:2109.10085 .*

[45] *Bhandari HN, Pokhrel NR, Rimal R, Dahal KR, Rimal B (2024) Implementation of deep learning models in predicting esg index volatility. Financial Innovation 10(1):75.*

[46] *Dalne G, et al. (2023) Esg-integrated machine learning portfolio opimization strategies: performance analysis and applications .*

[47] *Gopal S, Pitts J (2025) Genai: Unlocking sustainability insights and driving change in fintech. The FinTech Revolution: Bridging Geospatial Data Science, AI,*

and Sustainability, 345 – 393 (Springer).

[48] *Schimanski T, Reding A, Reding N, Bingler J, Kraus M, Leippold M (2024) Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. Finance Research Letters 61:104979.*

[49] *Koroteev MV (2021) Bert: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943 .*

[50] *Bsir B, Khoufi N, Zrigui M (2024) Prediction of author ' s profile basing on fine-tuning bert model. Informatica 48(1).*

[51] *Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 .*

[52] *Raza S, Paulen-Patterson D, Ding C (2025) Fake news detection: comparative evaluation of bert-like models and large language models with generative ai-annotated data. Knowledge and Information Systems 1 – 26.*

[53] *Lee H, Jung HS, Park H, Kim JH (2024a) Correct? corect!: Classification of esg ratings with earnings call transcript.KSII Transactions on Internet and Information Systems 18:1090 – 1100, URL http://dx.doi.org/10.3837/tiis.2024.04.015.*

[54] *Lee H, Lee SH, Park H, Kim JH, Jung HS (2024b) Esg2preem: Automated esg grade assessment framework using pre-trained ensemble models. Heliyon 10(4).*

[55] *Alkan AK, Grouin C, Schüssler F, Zweigenbaum P (2022) A majority voting strategy of a scibert-based ensemble models for detecting entities in the astrophysics literature (shared task). First Workshop on Information Extraction from Scientific Publications, 131 – 139.*