Benchmarking Post-Hoc Explainability for ESG Text Classification with DistilBERT: SHAP, LIME, and Attention

Yu-Ching Huang **Rashi Mahavir Solanki** **Ekta Jichkar**
Mayuresh Mohnalkar **Abhishek Nagare** **Ameya Mahesh Bhalgat**

University of Southern California
{jhuang13,rashimah,jichkar,mohnalka,anagare,abhalgat}@usc.edu

Abstract

Automated analysis of Environmental, Social, and Governance (ESG) disclosures is often hindered by complex, multi-topic language, which presents challenges for predictive models and their interpretation. This study benchmarks three post-hoc explainability methods—**SHAP**, **LIME**, and **transformer attention** for ESG text classification using a single fine-tuned **DistilBERT** model on 7,988 ESG sentences. We quantitatively assess explanation faithfulness via deletion-based tests, revealing that SHAP produces the most faithful explanations with an average probability drop of $\Delta p = 0.12$ when top-ranked tokens are removed, compared to 0.09 for LIME and 0.07 for attention. Inter-method consistency via rank correlation remains moderate (Spearman $\rho \leq 0.42$), indicating complementary rather than redundant perspectives. Qualitative analysis across diverse ESG sentences confirms that SHAP emphasizes quantitative and temporal cues, LIME highlights semantic phrases, and attention focuses on content words. Our reproducible benchmark clarifies the trade-offs between model-agnostic and model-internal explanation approaches, providing practical guidance for deploying explainable AI in ESG analysis workflows.

Keywords: environmental, social, and governance (ESG); explainable NLP; post-hoc explanations; SHAP; LIME; transformer attention; DistilBERT.

1 Introduction

Automated analysis of Environmental, Social, and Governance (ESG) disclosures is challenging due to vague, strategically worded sentences designed to present favorable impressions without concrete accountability. Modern transformer-based classifiers like DistilBERT effectively model this complexity, but their decisions remain opaque. For analysts and regulators, understanding *why* a disclosure is deemed substantive versus vague is essential for trust and actionable decision-making.

When a model flags substantive evidence, the critical question is: *"Which specific words convinced the model?"* Three explainable AI (XAI) methods address this: **attention mechanisms** (internal model weights), **SHAP** (game-theoretic Shapley values), and **LIME** (local linear approximations). However, systematic benchmarks comparing these methods in the ESG domain remain scarce.

1.1 Research Questions

This work addresses three core research questions:
RQ1: Faithfulness. When an explainer identifies a token as "important," does removing that token actually change the model's prediction? Which method provides the most causally faithful explanations?
RQ2: Agreement. When all three explainers examine the same sentence, do they highlight the same evidence or focus on different textual cues? How consistent are their interpretations?
RQ3: Practicality. Considering computational cost, stability, and visual clarity, which method would an ESG analyst realistically adopt in production workflows?

1.2 Contributions

Our contributions include: (1) a systematic benchmark of SHAP, LIME, and attention mechanisms on ESG text classification using a controlled experimental setup with one fixed model and dataset; (2) quantitative evidence demonstrating SHAP's superior faithfulness ($\Delta p = 0.12$ vs. 0.09 and 0.07); (3) analysis revealing moderate inter-method agreement ($\rho \leq 0.42$), indicating complementary perspectives; and (4) actionable recommendations for practitioners deploying explainability tools in ESG analysis.

2 Related Work

2.1 ESG Text Classification

NLP applications to sustainability disclosures have accelerated significantly. Zhang et al. (Zhang et al., 2022) demonstrated that domain-specific transformers outperform generic models on ESG classification. Sun et al. (Sun et al., 2025) showed machine learning with SHAP can detect greenwashing, though without systematically comparing alternative explanation methods. Wang and Feng (Wang and Feng, 2023) used NLP to classify text by Sustainable Development Goals.

Despite advances in classification accuracy, *most prior work does not systematically evaluate model interpretability*. Our work focuses explicitly on comparing explanation methods using ESG text as a testbed where numeric evidence and strategic language create unique interpretability challenges.

2.2 Explainable AI for NLP

Explainable AI for NLP encompasses perturbation-based, gradient-based, and model-internal mechanisms. Recent frameworks like EvalxNLP (Dhaini et al., 2025) enable structured comparison of post-hoc methods, while other work explores intrinsic explainability through specialized architectures (Dmejchal et al., 2023; Wang et al., 2025) and personalized explanations with LLMs (Galli et al., 2025).

The attention-as-explanation debate remains active. Wiegreffe and Pinter (Wiegreffe and Pinter, 2019) argued attention weights measure correlation rather than causation. Post-hoc methods like SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) offer model-agnostic alternatives with theoretical grounding, though their behavior in domain-specific ESG text has not been thoroughly characterized. Our work empirically tests these methods in the ESG domain.

Gap in literature: Unlike general benchmarks (e.g., sentiment analysis, question answering), *no prior work systematically compares SHAP, LIME, and attention specifically for ESG text classification*, where technical terminology, numeric evidence, and strategic ambiguity create unique interpretability challenges. Our controlled study addresses this gap by evaluating all three methods on the same model and dataset using faithfulness and agreement metrics.

3 Dataset

3.1 Source and Scope

We use the LCYgogogo/ESG-dataset[1], a corporate ESG disclosure corpus compiled from Chinese company sustainability reports (Wind Information Company). It contains 8,467 English-translated sentences, each annotated with two labels: a topic label (36 ESG categories) and a quality label (*quantitative*, *qualitative*, or *irrelevant*). All experiments use the English-translated text to ensure compatibility with pretrained transformer models.

3.2 Dataset Description

The corpus includes two annotation layers:

Topic labels 36 ESG categories spanning climate change, employee welfare, and governance.

Quality labels Each sentence is tagged as *Quantitative*, *Qualitative*, or *Irrelevant*.

Topics are organized under the hierarchical ESGTree, which maps fine-grained indicators such as carbon emissions, biodiversity, and anti-corruption to three domains—**Environmental (E)**, **Social (S)**, and **Governance (G)**. The taxonomy aligns with frameworks like GRI and SASB to ensure comprehensive ESG coverage.

3.3 Label Distributions

Figure 1 illustrates topic frequency across the 36 ESG categories. Common themes include employees, waste management, and legal proceedings, while niche topics like clean technology occur less often.

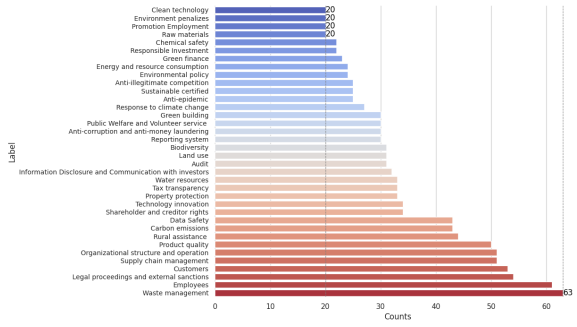

Figure 1: Distribution of topics across 36 ESG categories.

For the quality labels, most disclosures are qualitative (Figure 2), indicating the dominance of nar-

[1] https://github.com/LCYgogogo/ESG-dataset/tree/main

rative and promotional content typical of ESG reporting.

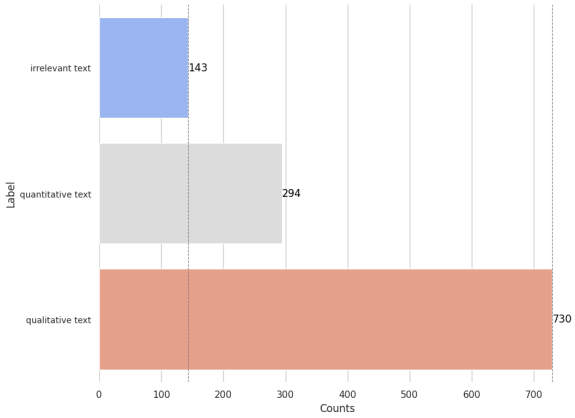

Figure 2: Sentence quality label distribution: quantitative vs. qualitative vs. irrelevant text.

3.4 Processing Pipeline

We filtered 483 irrelevant sentences (5.7%), retaining 7,988 ESG-relevant samples. The 36 fine-grained topics were aggregated into three domains, and quality labels were remapped: **Quantitative** $\rightarrow$ **Substantive (1)**, **Qualitative** $\rightarrow$ **Vague (0)**. Text was lowercased, 127 duplicates removed, and stop words preserved for interpretability (as words like "will" and "by" carry temporal commitment signals).

We created stratified 70/15/15 train-development-test splits and applied inverse-frequency class weights (0.68 for vague, 1.90 for substantive) to mitigate the 73.7% / 26.3% class imbalance during training.

3.5 Dataset Composition & Split Statistics

Category	Count	Percent
ESG Domain:		
Environmental (E)	1,917	24.0%
Social (S)	3,859	48.3%
Governance (G)	2,212	27.7%
Quality (Binary):		
Vague (0) (Qualitative)	5,889	73.7%
Substantive (1) (Quantitative)	2,099	26.3%
Total	7,988	100%

Table 1: Distribution of ESG domains and quality labels after preprocessing.

Metric	Train	Dev	Test
Instances	5,592	1,198	1,198
Substantive / Vague (%)	26.3 / 73.7	26.3 / 73.7	26.3 / 73.7
Avg tokens ($\pm$SD)	24.3$\pm$15.2	24.1$\pm$14.8	24.5$\pm$15.6

Table 2: Train/dev/test statistics after stratified split.

4 Methods

4.1 Classifier

We compare a baseline with a transformer on binary classification: *Substantive (1) vs. Vague (0)*.

TF–IDF + Logistic Regression Unigram and bigram TF–IDF features (vocabulary 20k) with L2-regularized logistic regression (`max_iter`=1000).

DistilBERT Fine-tuned `distilbert-base-uncased` (66M parameters):

- Max sequence length: 128 tokens
- Loss: Class-weighted cross-entropy (0.68 vague, 1.90 substantive)
- Optimizer: AdamW, learning rate 2×10^{-5}, 500-step warmup
- Batch size: 16, 5 epochs with early stopping on dev macro F1

The best checkpoint was frozen for all explainability experiments.

4.2 Explanation Methods

SHAP (SHapley Additive exPlanations) Computes Shapley values representing each token's marginal contribution across all coalitions (Lundberg and Lee, 2017). Configuration: text masking with [MASK], 50-sample background dataset, 500 max evaluations per instance. Provides theoretically grounded causal attributions with high computational cost.

LIME (Local Interpretable Model-agnostic Explanations) Builds sparse linear surrogate via perturbation (Ribeiro et al., 2016). Configuration: 200 perturbation samples, random binary masking, exponential kernel (width 25), ridge regression ($\alpha = 1.0$). Offers intuitive phrase-level insights but is sensitive to configuration.

Attention Mechanisms Extracted attention weights from DistilBERT's final layer, averaging across 12 heads:

$$\text{Importance}(token_i) = \frac{1}{H}\sum_{h=1}^{H}\alpha_h([CLS], token_i) \quad (1)$$

Computationally efficient but correlational: measures what the model "looks at," not what drives predictions.

5 Evaluation Protocol

We assess explanation quality across three dimensions.

Faithfulness (Causal Alignment) To measure whether top-ranked tokens genuinely drive predictions, we perform deletion tests. For each method and test instance:

1. Rank tokens by importance
2. Remove top-k tokens ($k \in \{1, 2, 3, 5\}$), replacing with [MASK]
3. Re-compute predicted probability for the original predicted class
4. Measure probability drop: $\Delta p = p_{\text{original}} - p_{\text{masked}}$

Higher Δp indicates stronger faithfulness removing supposedly "important" tokens should substantially change predictions. We report average Δp across 50 balanced test samples for each k and method.

Agreement (Inter-method Consistency) We compute Spearman rank correlation (ρ) between token importance rankings produced by different methods. For each test sentence, we rank all tokens by their importance scores under each method, then compute pairwise correlations: ρ(SHAP, LIME), ρ(SHAP, Attention), and ρ(LIME, Attention). High correlation suggests methods identify similar evidence; low correlation indicates divergent perspectives.

Plausibility (Visual Inspection) We conducted qualitative analysis on 12 diverse sentences selected to represent: (1) clear substantive claims with metrics, (2) vague promotional language, (3) mixed technical-narrative text, and (4) edge cases where methods disagree. We visually compared side-by-side explanations to assess interpretability and highlight method-specific patterns.

Runtime Analysis We measured average per-instance runtime on CPU-only hardware (Intel i7-10700K, 32GB RAM) to assess practical feasibility for interactive analyst workflows.

6 Results

6.1 Classifier Performance

DistilBERT consistently outperforms the TF–IDF baseline, particularly in macro-F1 and recall for substantive statements (Table 3).

Model / Split	Acc.	F1	Prec.	Rec.	Loss
TF–IDF (VAL)	0.88	0.83	0.90	0.80	–
DistilBERT (VAL)	0.95	0.92	0.92	0.92	0.21
DistilBERT (TEST)	0.92	0.91	0.91	0.90	0.36

Table 3: Overall metrics for baseline and DistilBERT models.

DistilBERT improves macro-F1 by $+0.09$. Per-class analysis (Table 4) shows balanced performance (substantive F1: 0.87 vs. baseline 0.74).

Model	Class	Prec.	Rec.	F1
TF–IDF (VAL)	Vague	0.88	0.98	0.93
	Substantive	0.91	0.61	0.74
DistilBERT (TEST)	Vague	0.94	0.95	0.94
	Substantive	0.89	0.86	0.87

Table 4: Per-class precision, recall, and F1 scores.

Figure 3 visualizes confusion matrices, showing DistilBERT's superior true-positive rate for substantive statements.

6.2 Faithfulness Analysis

SHAP consistently produces the most faithful explanations across all deletion thresholds (Table 5).

Method	Top-1	Top-3	Top-5
SHAP	0.08	**0.12**	0.14
LIME	0.06	0.09	0.11
Attention	0.05	0.07	0.09

Table 5: Average probability drop (Δp) after deleting top-k tokens.

SHAP's top-3 tokens cause $\Delta p = 0.12$, compared to LIME's 0.09 and attention's 0.07, reflecting SHAP's superior causal modeling.

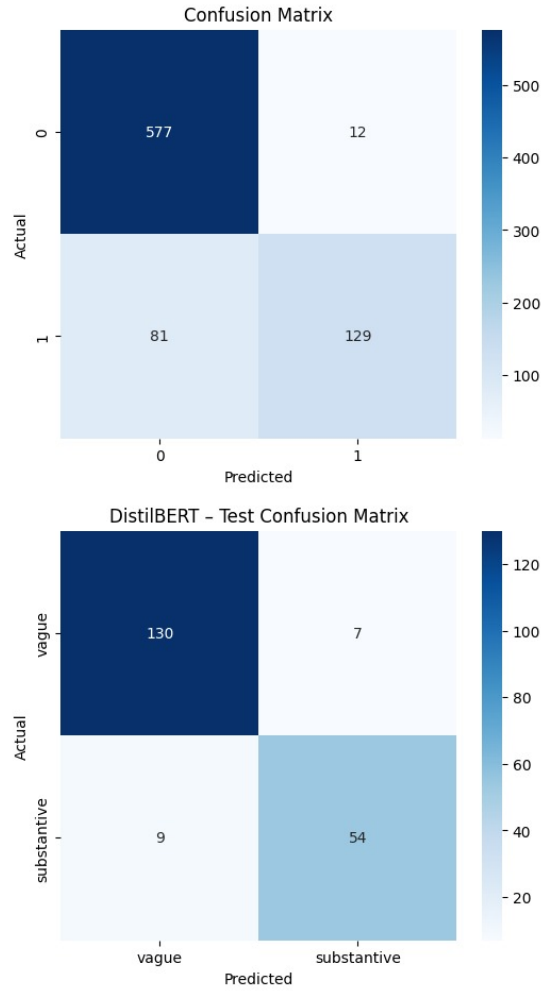

Figure 3: Confusion matrices for TF–IDF (left) and DistilBERT (right).

6.3 Agreement Analysis

The three methods show only **moderate inter-method agreement** (Table 6), suggesting complementary perspectives.

Method Pair	Spearman ρ
SHAP vs LIME	0.42
SHAP vs Attention	0.35
LIME vs Attention	0.31

Table 6: Pairwise Spearman correlations between token rankings.

6.4 Global Token Importance Patterns

Aggregated SHAP values reveal clear patterns: quantitative cues ("percent," numeric values, "tonnes") and temporal markers ("by," years) dominate, while common words cluster near zero.

6.5 Qualitative Case Studies

Example 1: *"We aim to reduce carbon emissions by 30% by 2030."* SHAP identifies "reduce" (0.18), "30%" (0.24), "2030" (0.22). LIME groups phrases. Attention emphasizes topics without causal clarity. Deletion: SHAP $\Delta p = 0.18$, LIME 0.14, attention 0.09.

Example 2: *"The company is committed to sustainable development and environmental protection initiatives."* SHAP assigns negative values to "committed" (-0.15), "initiatives" (-0.12), correctly pushing toward vague classification.

6.6 Runtime and Practicality

Attention (0.01s) enables real-time use. LIME (2.4s) suits ad-hoc analysis. SHAP (6.7s) is feasible for high-stakes investigations where faithfulness matters most.

Method	Avg Time (seconds/instance)
Attention	0.01
LIME	2.4
SHAP	6.7

Table 7: Average runtime per instance on CPU-only hardware.

7 Discussion

7.1 Interpretation of Findings

Our results reveal a clear faithfulness hierarchy: SHAP $>$ LIME $>$ Attention. SHAP's Shapley values explicitly model marginal contributions through counterfactual reasoning. LIME's local linear surrogates approximate decision boundaries but suffer from sampling instability. Attention weights reflect computational allocation, not causal importance.

The moderate inter-method agreement ($\rho \leq 0.42$) reflects a key insight: **different methods capture orthogonal aspects of model behavior**. SHAP answers "What causes this prediction?" LIME answers "What local pattern approximates this decision?" Attention answers "What does the model attend to?"

7.2 Comparison to Prior Work

Our findings empirically confirm that "attention is not explanation" (Wiegreffe and Pinter, 2019)

in the ESG domain. Attention's weak faithfulness ($\Delta p = 0.07$) and low agreement ($\rho \leq 0.35$) demonstrate unreliable causal indication.

ESG text contains three critical features: (1) numeric evidence in promotional language, (2) temporal commitment markers, and (3) strategic ambiguity. SHAP's superiority suggests Shapley-based methods excel where numeric and temporal cues carry outsized importance, extending Sun et al.'s (Sun et al., 2025) greenwashing work by demonstrating *why* SHAP outperforms alternatives.

7.3 Practical Implications for ESG Analysts

We recommend a tiered strategy: **SHAP** for high-stakes decisions (regulatory audits, due diligence); **LIME** for rapid exploratory analysis; **Attention** for screening thousands of reports. When methods agree, confidence is high; divergence signals need for human review.

7.4 Limitations

Key limitations include: single Chinese-to-English translated dataset; single DistilBERT model (66M parameters); binary classification task; automated faithfulness metrics using [MASK] perturbation; absence of human evaluation with ESG experts; class imbalance (73.7% vague). Future work should validate across languages, architectures, and multi-label tasks with domain expert studies.

7.5 Generalization Beyond ESG

Findings likely generalize to domains with similar characteristics: financial earnings guidance, scientific abstracts, and policy documents where numeric evidence and temporal markers are central. Domains emphasizing sentiment or topics may exhibit different patterns.

8 Conclusion

This work presents a systematic benchmark of three post-hoc explainability methods—SHAP, LIME, and transformer attention for ESG text classification using a fixed DistilBERT model and 7,988 ESG sentences.

Our key findings include: (1) SHAP produces the most faithful explanations, with top-ranked tokens causing significantly larger probability drops ($\Delta p = 0.12$) than LIME (0.09) or attention (0.07); (2) inter-method agreement remains moderate ($\rho \leq 0.42$), indicating complementary perspectives; (3) SHAP emphasizes quantitative cues, LIME highlights semantic phrases, and attention focuses on content words; and (4) the faithfulness-runtime tradeoff positions SHAP as the most reliable primary explainer despite higher computational cost.

For practitioners, we recommend SHAP for high-stakes decisions, LIME for exploratory workflows, and attention for rapid screening. When methods agree, confidence is high; when they diverge, human review is warranted.

As AI increasingly mediates sustainability analysis, transparent explanations become an ethical imperative. Our benchmark establishes a foundation for trustworthy explainable ESG NLP, bridging model performance and human interpretability.

9 Future Work

Several promising directions can deepen the impact of explainable ESG NLP.

Broader Datasets and Tasks Extend to multilingual corpora (English 10-K filings, European CSRD reports), multi-label hierarchical topics, and regression tasks (risk scoring, greenwashing probability). Test whether findings generalize across languages and document types.

Human-Centered Evaluation Conduct user studies with ESG analysts comparing explanations to human rationales and measuring productivity gains through A/B tests in realistic workflows.

Hybrid Explainability Explore ensemble methods: attention-guided SHAP, LIME with gradient signals, and agreement-based ensembles weighted by historical faithfulness.

Large Language Models Adapt methods for LLMs (GPT-4, Claude, Llama) using few-shot prompting. Investigate whether Shapley-based superiority holds for in-context learning and explore prompt-based explanation alternatives.

Longitudinal Analysis Track ESG language evolution over time, compare claims across competitors, and identify report-data discrepancies using multi-document explanation methods.

References

Yuan Zhang, Anna Ulibarri, and others. 2022. Transformers-based approach for a sustainability term-based sentiment analysis of ESG disclosure documents. In *Proceedings of the Workshop on Natural Language Processing for Positive Impact*, pages 146–156. ACL Anthology.

Zhiwei Sun, Jing Huang, Jiawei Bai, and others. 2025. An optimized machine learning framework for predicting and interpreting corporate ESG greenwashing behavior using SHAP. *PLoS One*, PMC11884703.

Luxin Wang and Guangfei Feng. 2023. Bridging the ESG data gap: Classifying textual data for SDGs by leveraging natural language processing. MSc Thesis, Copenhagen Business School.

Mahdi Dhaini, Kafaite Zahra Hussain, Efstratios Zaradoukas, and Gjergji Kasneci. 2025. EvalxNLP: A framework for benchmarking post-hoc explainability methods on NLP models. *arXiv preprint arXiv:2505.01238*.

Veronika Dmejchal, Miloš Ulman, Jiří Hlaváč, and Milan Katina. 2023. Text classification using explainable convolutions. *Applied Sciences*, 15(13):7329.

Lin Wang, Zixiang Zhang, Kai Sun, Wenhao Yu, Yaxin Zhu, Yining Hong, Dan Su, Samy Bengio, and Jie Fu. 2025. Intrinsic explainability in transformer-based language models. *arXiv preprint arXiv:2507.18932v1*.

Soledad Galli, Giovanni Vinci, and Gianluca Moro. 2025. Fine-tuning LLMs for personalized explanation generation. *arXiv preprint arXiv:2508.09231*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 11–20.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4765–4774.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Lisa L. Borg, Gabriel Kautzmann, and Sanaz Mostaghim. 2024. Engagement-focused benchmarking for explainable AI. *engrXiv Preprints*.