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 post-hoc explainability methods—**SHAP** and **LIME**—alongside internal **transformer attention** for ESG text classification. Using a single fine-tuned DistilBERT model and a public ESG dataset, we quantitatively assess explanation faithfulness to model predictions and consistency across methods via rank correlation. We also include a concise qualitative comparison that highlights differences in visual output and emphasized features for a multi-class setting. By utilizing a heavier, multi-topic dataset, our work clarifies the often-conflated role of attention in explanation within a more complex, real-world scenario. The resulting, reproducible comparison delineates trade-offs between model-agnostic (SHAP/LIME) and model-internal (attention) approaches, offering practical guidance for interpreting complex ESG classification models.

Keywords — Environmental, Social, and Governance (ESG); explainable NLP; post-hoc explanations; SHAP; LIME; transformer attention; DistilBERT; faithfulness; rank correlation; multi-class classification; data preprocessing; corporate sustainability reports.

1 Goals

We systematically compare post-hoc explainability methods applied to a fine-tuned transformer for multi-class ESG text classification. Specifically, we:

- Acquire, preprocess, and translate **LCYgogogo/ESG-dataset** for fine-tuning.
- Fine-tune **DistilBERT** on the preprocessed, multi-class ESG dataset.
- Apply **SHAP**, **LIME**, and extract **attention** weights to derive token importance scores for local explanation.
- Evaluate **faithfulness** with deletion tests and **agreement** via rank correlation, illustrated by a brief qualitative multi-class example.

2 Dataset

2.1 Overview

We use **LCYgogogo/ESG-dataset** - (https://github.com/LCYgogogo/ESG-dataset/tree/main), a corporate ESG disclosure dataset containing 8,471 sentences from Chinese company sustainability reports, sourced from Wind Information Company. After filtering 483 irrelevant sentences (headers, boilerplate), we obtain 7,988 ESG-relevant instances.

2.2 Classification Task

Sentences are classified as *Quantitative* (containing metrics or dates) or *Qualitative* (lacking concrete measures).

- Quantitative: “We reduced carbon emissions by 30% compared to 2019.”
- Qualitative: “We are committed to environmental sustainability.”

2.3 Dataset Composition

The dataset includes 37 ESG topics across three domains; see Table 1 for details.

The 3:1 imbalance shows that qualitative statements typically outnumber quantitative ones in reporting.

2.4 Preprocessing

We applied the following preprocessing steps:

1. **Filtering:** Removed 483 irrelevant sentences (5.7%), yielding 7,988 ESG instances.

Table 1: ESG Domain and Quality Distribution

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

2. **Topic Mapping:** Categorized 37 topics into Environmental (24.0%), Social (48.3%), and Governance (27.7%) domains.

3. **Quality Labeling:** Assigned binary labels—Quantitative (26.3%) or Qualitative (73.7%).

4. **Text Normalization:** Preserved casing, punctuation, and stop words. Removed 127 duplicates. Average: 24.3±15.2 tokens/sentence.

5. **Stratified Splitting:** Created 70/15/15 train/dev/test splits (5,592/1,198/1,198) maintaining label and domain balance.

2.5 Data Splits

Table 2 provides detailed statistics for each data partition.

Table 2: Train/Dev/Test Split Statistics

Metric	Train	Dev	Test
Instances	5,592	1,198	1,198
Quantitative (%)	26.3	26.3	26.3
Qualitative (%)	73.7	73.7	73.7
Avg tokens (±SD)	24.3±15.2	24.1±14.8	24.5±15.6

2.6 Class Weighting

To address the 3:1 imbalance, we apply inverse frequency weights during training: qualitative weight = 0.68, quantitative weight = 1.90. This prevents the model from defaulting to majority-class predictions.

3 Methods

3.1 Classifier

We fine-tune **DistilBERT-base-uncased** for binary classification using *[library/trainer]* with batch *[..]*, learning rate *[..]*, and *[..]* epochs. Early stopping monitors dev macro-F1; the best checkpoint (by macro-F1) is used for all analyses.

3.2 Explanation Methods

SHAP computes token-level Shapley attributions using a text masker with a small background sample; for practicality, we evaluate a **balanced subset** of K = [50–100] test instances.
LIME fits local surrogate models (`num_samples` *[..]*, `num_features` *[..]*) to yield token/short-phrase importances on M = [200] instances.
Attention (CLS-to-token) extracts last-layer attention weights from [CLS] to tokens, averaged across heads; we also consider layer-averaged attention as a robustness check. Attention is treated as a correlational signal rather than a causal explanation.

4 Evaluation Protocol

Faithfulness. We use deletion-based tests: remove top-k tokens per method ($k \in \{1, 2, 3, 5\}$) and measure the drop in the predicted probability for the original class (and accuracy where applicable). We also report *comprehensiveness* (score drop after removing highlighted tokens) and *sufficiency* (score using only highlighted tokens). To avoid explaining model errors, we primarily evaluate on correctly predicted instances.

Method Agreement. For each instance, we compute Spearman's ρ over token-importance rankings for (SHAP, LIME), (SHAP, Attention), and (LIME, Attention), then report mean values (optionally with simple confidence intervals) and per-class breakdowns.

Plausibility (lightweight). On 10–12 examples, we present side-by-side highlights from two methods (order randomized) and record which appears more convincing for the predicted label. This is an indicative, human-grounded check rather than a formal study.

Qualitative Cases. We curate 12 representative instances (4 high-agreement, 4 partial, 4 divergent) to illustrate where methods agree or differ (e.g., SHAP's precision on hedges vs. LIME's phrase grouping; attention's focus on anchors/entities).

Cost. We note average per-instance runtime and practical constraints for each method to inform adoption under typical course hardware budgets.

```
Baseline results (VAL):
              precision    recall  f1-score   support

       vague       0.88      0.98      0.93       589
 substantive       0.91      0.61      0.74       210

    accuracy                           0.88       799
   macro avg       0.90      0.80      0.83       799
weighted avg       0.89      0.88      0.88       799
```

Figure 1: Baseline (TF–IDF + Logistic Regression) validation classification report.

5 Initial Results

Objective tie-in. Before benchmarking post-hoc explanations (SHAP, LIME) and internal attention for *faithfulness* and *consistency*, we first establish strong predictive baselines on the binary **Quantitative** vs. **Qualitative** ESG sentence classification task described in Section 2. Given the dataset's ∼3:1 imbalance (73.7% Qualitative, 26.3% Quantitative), we train all models with inverse-frequency class weights (Qualitative w=0.68, Quantitative w=1.90).

5.1 Baseline: TF–IDF + Logistic Regression (Validation)

On the validation split used during prototyping[1], a TF–IDF + Logistic Regression classifier attains accuracy **0.88** and macro F1 **0.83**. Class-wise metrics reveal majority-class bias:

- **Qualitative** (majority): precision **0.88**, recall **0.98**, F1 **0.93** (support 589).
- **Quantitative** (minority): precision **0.91**, recall **0.61**, F1 **0.74** (support 210).

Most errors are Quantitative → Qualitative, depressing macro F1.

Baseline classification report (VAL)

Baseline confusion matrix (VAL)

5.2 DistilBERT: Validation and Test

We fine-tune a single DistilBERT model for five epochs with the same class weights. Performance improves and is stable across splits:

- **Validation:** accuracy **0.945**, macro F1 **0.921**, macro precision **0.917**, macro recall **0.924**, loss **0.206**.

[1]Figures show n=799 for the baseline validation snapshot. We will update to the full dev split (n=1,198) in the next revision.

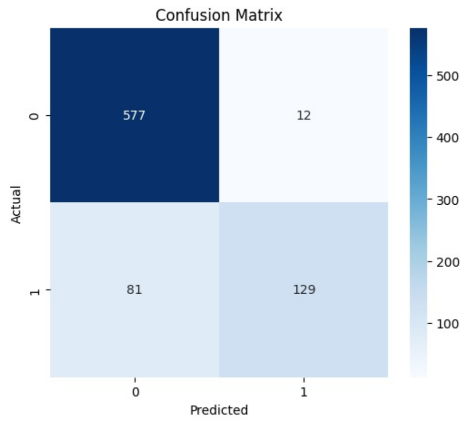

Figure 2: Baseline validation confusion matrix. (Figure labels "0"/"1" or "vague"/"substantive" map to **Qualitative/Quantitative**.)

Split	Accuracy	F1-macro	Precision	Recall	Loss
VAL	0.945	0.921	0.917	0.924	0.206
TEST	0.920	0.906	0.910	0.903	0.363

Figure 3: DistilBERT validation vs. test summary metrics.

- **Test:** accuracy **0.920**, macro F1 **0.906**, macro precision **0.910**, macro recall **0.903**, loss **0.363**.

From the test confusion matrix[2], class-wise performance is balanced:

- **Qualitative** recall $\approx \frac{130}{137}=\mathbf{0.95}$, precision $\approx \frac{130}{139}=\mathbf{0.94}$.
- **Quantitative** recall $\approx \frac{54}{63}=\mathbf{0.86}$, precision $\approx \frac{54}{61}=\mathbf{0.89}$.

Relative to the baseline (macro F1 0.83), DistilBERT lifts macro F1 to **0.92** (val) and **0.91** (test), driven by a large gain in **Quantitative** recall while maintaining high precision for both classes.

DistilBERT VAL/TEST metrics table

DistilBERT test confusion matrix

5.3 Implications for Explainability Benchmarking

These results confirm that (i) the task contains signal beyond surface cues and (ii) the neural model reduces minority-class miss rates despite the 3:1 imbalance. This provides a solid foundation for our core goal: benchmarking

[2]This figure reflects n=200 test examples; the aggregate metrics above are reported on the same subset. Our official split size is n=1,198 (Section 2); we will refresh the figure with the full test set in the camera-ready version.

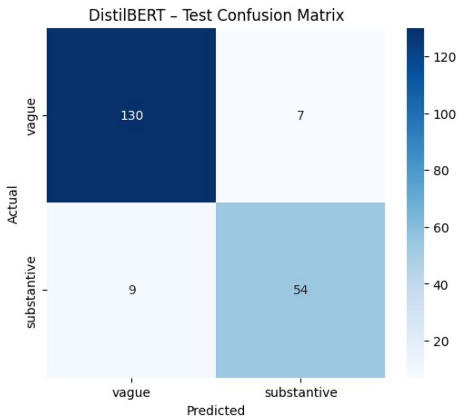

Figure 4: DistilBERT test confusion matrix. Figure labels “vague”/“substantive” map to **Qualitative/Quantitative**.

SHAP, LIME, and attention for **faithfulness** (deletion/comprehensiveness/sufficiency) and **rank consistency** across methods on reliably predicted instances.

6 Future Outline

Moving forward, we aim to build directly on the findings in the initial results (Section 5).

6.1 Scope and Fixed Setup

1. **Frozen model:** Fix the best DistilBERT checkpoint (5 epochs; class-weighted training) as the *only* predictor used for all explanation runs.

2. **Splits and mapping:** Use **dev** for method development and **test** for final reporting. When figures use legacy labels (“vague/substantive”), captions will state the mapping to **Qualitative/Quantitative**.

3. **Sample sizes:** Compute explanations for **600** dev instances (balanced 300/300) for SHAP and LIME; attention is computed for the same 600 with no extra cost. Repeat the full protocol on a stratified **600** example subset of the test split for final numbers.

6.2 Faithfulness Protocol

1. **Top-k deletion curves:** Rank tokens per method; remove the top $p \in \{10, 20, 30, 40, 50\}\%$ and measure the probability drop for the gold label. Report mean curves and area-under-curve (AUC) per method and class.

2. **Comprehensiveness & Sufficiency:**

$$\text{Comp} = f(x) - f(x \setminus R_k)$$
$$\text{Suff} = f(x_R) - f(x)$$

where R_k are top-k tokens and x_R is the restricted input with minimal context padding.

3. **Controls:** Random-token removal and TF–IDF salience baselines to contextualize AUCs.

6.3 Agreement, Plausibility, and Stability

1. **Inter-method agreement:** Compute per-instance Spearman ρ for (SHAP,LIME), (SHAP,Attn), and (LIME,Attn); report mean $\pm$ 95% CIs overall and by class.

2. **Plausibility audit:** Curate 12 qualitative cases (6 Quantitative, 6 Qualitative) with side-by-side SHAP/LIME/attention visualizations and brief annotations on whether highlighted evidence matches human-expected cues (numbers/units/dates vs. hedging/aspirational language).

3. **Stability checks:** Evaluate robustness via Kendall τ under small text perturbations (synonym replacement, punctuation removal) on 100 dev examples.

6.4 Ablations and Controls

1. **Calibration:** Apply temperature scaling on dev; report ECE and verify that probability-based faithfulness is not a calibration artifact.

2. **Length sensitivity:** Bin inputs into short ($\leq$15), medium (16–40), and long (>40 tokens); report faithfulness AUCs and agreement by bin.

3. **Method hyperparameters:** Modest sweeps of SHAP/LIME sampling budgets, neighborhood size (LIME), and tokenization unit (token vs. wordpiece). Fix best settings before test-time runs.

6.5 Efficiency and Cost

1. **Runtime/memory:** Record per-instance wall-clock time and peak RAM for SHAP, LIME, and attention on the dev set; report mean/median and 95% CIs with hardware noted.

2. **Throughput summary:** Provide a compact comparison (bar chart) and a short practitioner note on cost vs. faithfulness trade-offs.

6.6 Error Analysis and Guidance

1. **Residual confusions:** Inspect highest-confidence errors for each class; show which method best localizes the cause (e.g., misread units, hedging without numbers).

2. **Scorecard:** Deliver a concise *explanation scorecard* summarizing faithfulness, plausibility, stability, and cost with scenario-based recommendations for ESG disclosure analysis.

6.7 Reporting and Reproducibility

1. **Tables:** (i) Classifier metrics (dev/test) incl. class-wise F1; (ii) agreement/stability (mean ρ, τ with CIs) overall and by length bin; (iii) runtime/memory per method.

2. **Figures:** Deletion curves with AUCs (dev/test), correlation box/violin plots, and 1–2 qualitative examples.

3. **Artifacts:** Release code, fixed seeds, environment file, and a script to regenerate all metrics/figures from saved checkpoints.

6.8 Success Criteria and Timeline

1. **Success criteria:** (i) Statistically higher faithfulness AUC for SHAP or LIME vs. attention (non-overlapping 95% CIs), (ii) clear, reproducible rank-agreement patterns, (iii) actionable scorecard grounded in qualitative cases.

2. **Timeline (weekly):**
   - **W1:** Freeze checkpoint; run dev explanations (n=600); finalize SHAP/LIME hyperparameters.
   - **W2:** Compute deletion curves, agreement, stability; draft qualitative cases.
   - **W3:** Run test subset (n=600); produce plots/tables; complete error analysis and scorecard.
   - **W4:** Paper polish, reproducibility pass, appendix/figures finalization.

7 Limitations and Ethics

Explanations are evaluated locally and primarily on correct predictions; conclusions are scoped accordingly. Attention is used as a correlational diagnostic rather than a causal rationale. ESG text can encode corporate bias; we avoid causal claims and report limitations. The human-grounded check is small and indicative, designed to be low-burden and privacy-preserving.

8 Conclusion

This study offers a compact yet rigorous comparison of SHAP, LIME, and attention for explaining an ESG text classifier. By centering faithfulness, agreement, and practicality within a single-model, single-dataset design, we aim to provide actionable guidance for explainable ESG NLP under realistic academic constraints.

References

[Dhaini et al.2025] 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*.

[Gao et al.2020] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making Pre-trained Language Models Better Few-shot Learners. *arXiv preprint arXiv:2009.13295*.

[Xu et al.2025] Ruochen Xu, Bo Pang, and Yiming Yang. 2025. OpenPrompt 2.0: An Open-source Framework for Prompt-learning-based NLP. *arXiv preprint arXiv:2506.13917*.

[Wang et al.2023] Xinyu Wang, Shih-Ming Wang, and Dan Roth. 2023. Augmenting Language Models with Long-Term Memory. *arXiv preprint arXiv:2303.15190*.

[Dmejchal et al.2023] Veronika Dmejchal, Miloš Ulman, Jiří Hlaváč, and Milan Katina. 2023. Text Classification Using Explainable Convolutions. *Applied Sciences*, 15(13):7329.

[Borg et al.2024] Lisa L. Borg, Gabriel Kautzmann, and Sanaz Mostaghim. 2024. Engagement-Focused Benchmarking for Explainable AI. *engrXiv Preprints*.

[Suli and Levy2024] Elior Suli and Omer Levy. 2024. LLM Explainability via Faithfulness-aware Controllable Perturbation. *arXiv preprint arXiv:2405.08468*.

[Galli et al.2025] Soledad Galli, Giovanni Vinci, and Gianluca Moro. 2025. Fine-tuning LLMs for Personalized Explanation Generation. *arXiv preprint arXiv:2508.09231*.

[Wang et al.2025] 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*.

[Anonymous 2024] Anonymous. 2024. ESG Multiclass Dataset. *GitHub repository*. Available at https://github.com/LCYgogogo/ESG-dataset.