Greenwashing Detection with Causal Explanation: A Novel Multi-layered Approach

The pervasive issue of greenwashing in corporate sustainability reports necessitates robust and transparent detection mechanisms. Existing methods despite their advancements, often lack resilience to adversarial manipulations and frequently fail to provide interpretable insights into why a statement is deemed deceptive, limiting their real-world reliability. This work addresses these limitations by introducing a novel, multi-layered framework for greenwashing detection that integrates advanced language modeling with a causal inference approach.

Our methodology is grounded in a robust dataset and a principled approach to overcoming data scarcity. We curated a dataset of 975 corporate sustainability reports spanning six sectors (Retail, Technology, Automotives, Financial Services, Pharmaceuticals, Consumer Goods), from 100 NASDAQ-listed companies, spanning over five years (2020 - 2024), augmented with media articles from Bloomberg and Reuters, and environmental risk scores from Sustainalytics. To address the challenge of limited labeled examples, we employed a synthetic data generation technique to create positive and negative instances of greenwashing.

The core of our detection framework involves three distinct classification models: (i) a fine-tuned RoBERTa-Large model, (ii) a simple term frequency -inverse document frequency (TF-IDF) + support vector machine (SVM) based classifier, and (iii) a specialized ClimateBERT model [1]. These models are benchmarked against our curated dataset to establish their performance in identifying greenwashing claims.

Our key contribution is the proposal of a novel Green Authenticity Index (GAI), which provides a quantitative and interpretable measure of greenwashing. The GAI is a composite score,

$$\texttt{GAI} = \alpha * \texttt{Certainty} + (1 - \alpha) * \texttt{Agreement}$$

designed to move beyond simple binary classification. The Certainty component assesses linguistic features such as clarity, factuality, and specificity by evaluating a statement's ambiguity, verifiability, and concrete detail (e.g., key milestone, location, practical timeline with commitments). The Agreement component measures a claim's alignment with independent, external evidence from media articles and environmental risk scores. By integrating these metrics, the GAI provides a measurable indicator for potential greenwashing, allowing for a more nuanced and transparent assessment.

Lastly, we introduce a causal inference framework to provide causal explanations for our model's predictions. By connecting the classification outputs to the GAI's sub-scores, our framework not only classifies a statement as greenwashing but also explains why it is deceptive. The GAI demonstrates strong correlation (r=0.76) with expert human assessments. while the causal inference framework successfully identifies key linguistic patterns indicative of greenwashing, providing actionable insights for regulators and investors navigating corporate sustainability claims [1]

References

[1] N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold, "Climatebert: A pretrained language model for climate-related text," *arXiv preprint arXiv:2110.12010*, 2021.

[1] Work in progress: dataset and code supporting this study will be made publicly available upon completion of the research.