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

AI Researcher

Outline

- Background
- Motivation
- Dataset description
- Methodology
- Result and analysis
- Conclusion

Background

- Corporate sustainability reporting suffers from widespread greenwashing
- Regulators and stakeholders need transparent reasoning to make decisions or investments
- Existing methods study greenwashing (text classification, environmental claim detection, regression modeling, semantic analysis, etc.) but do not focus towards the cause of why a claim is deceptive.

Figure: Corporate greenwashing

Source:https://reputationtoday.in/sustainability-not-just-a-front-anymore

Why we need interpretability?

Limitations of existing approach:

- Models predict greenwashing/authentic claims based on data pattern
- Variables influencing greenwashing are unknown
- Variables influencing both greenwashing and authentic claims are unknown
- Unknown variations are not modeled explicitly but comprehensively processed

Figure: Wordcloud generated from corporate text data, larger font size mean more usage

Problem statement: Develop a multi-layered framework for greenwashing detection that integrates advanced language modeling with a causal inference approach

Dataset description

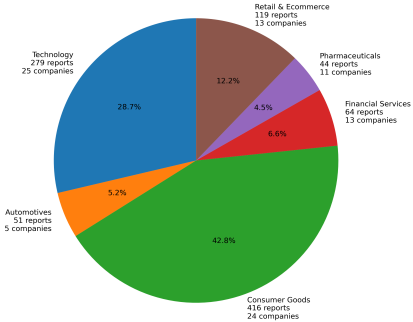

Figure: Pie chart illustrating the corporate reports distribution across North America

- Curated third party media articles from Bloomberg, Reuters, and other sources
- Obtained environmental risk scores published by Sustainanalytics

Data preprocessing

- Climate reports are unstructured data: tabular data, texts, images.
- First processing: Domain specific data extraction tool (Reportparse)
- Second processing: Customized stop word, url, etc removal

Methodology

- Trained 3 greenwashing classifiers
  - Pretrained RoBERTa on unlabelled 1.3 million samples of corporate text data. Finetuned on 10,000 samples of annotated and augmented greenwashing labels
  - Finetuned domain specific ClimateBERT model on 10,000 samples of annotated and augmented greenwashing labels
  - Used classical Term Frequency Inverse Document frequency (TF-IDF) for feature extraction and support vector machine (SVM) for classification using 10,000 samples of annotated and augmented greenwashing labels
- Compute Green authenticity index(GAI): Certainty and Agreement
- Studied causal mediation analysis to interprete the greenwashing predictions from classifiers

Modeling Greenwashing

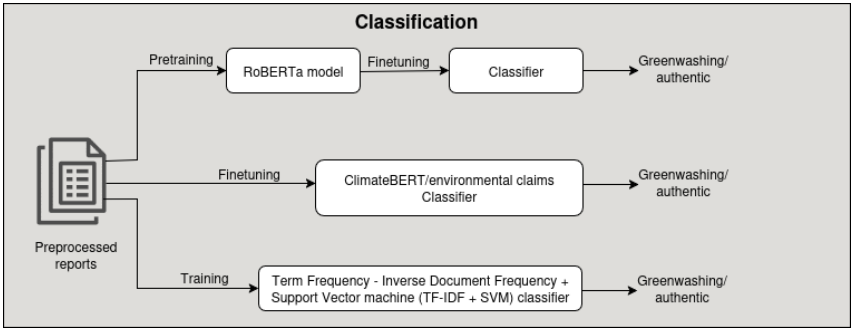

Figure: Classification modules for greenwashing detection

Green Authenticity Index (GAI)

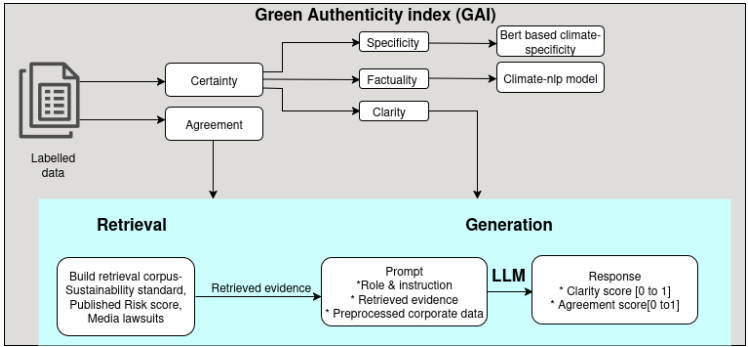

Figure: GAI: quantitative greenwashing measure

$GAI = \alpha \times Certainty + (1 - \alpha) \times Agreement$

High GAI $\rightarrow$ transparent & evidence-backed

Low GAI $\rightarrow$ high-risk greenwashing

GAI subcomponents: Certainty & Agreement

- Certainty assess whether a statement is vague or factual
- Certainty is a composite score: Clarity, Specificity and Factuality
  - Clarity metric measures ambiguity and complexity of a language
  - Specificity metric assess whether a statement provides concrete details about scope, location and time
  - Factual metric assess verifiable and concrete information
- Agreement metric measures whether a claim is true and align with independent and external evidence like media articles, published risk score, etc

GAI subcomponents: Clarity and Agreement

Retrieval augmented generation(RAG)

- Retrieval
  - Build retrieval corpus: Sustainalytics risk scores, Media articles, Verified disclosures (SBTi, CDP, TCFD), Legal filings/adverse reports
  - Preprocessed data, compute embeddings and stores metadata using FAISS index
- Generation
  - For each query Top-K relevant evidence are retrieved
  - Prompt design for a sustainability expert where the instructions are: "Is this sentence clearly written, unambiguous, and free from non-transparent words? Does the claim align with the evidence? - Provide output in a specific format"
  - LLM
    - Opensource Qwen2 Instruct model
    - Provide Clarity and Agreement score in the range of 0 to 1 with justification

Causal Mediation analysis

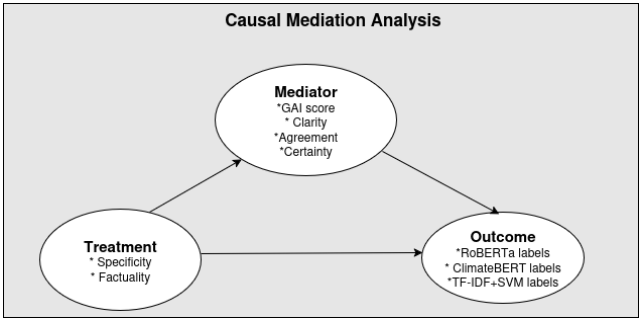

Figure: Causal mediation analysis

Objective: Under what mechanism does a treatment Variable affect an outcome ?

- Compute : Total Effect, Direct Effect, Indirect Effect
- Treatment, mediator and outcomes are analysed for different combinations

Results & analysis

Table: Classification model performance

Models	Accuracy	F1-score
RoBERTa	0.94	0.91
ClimateBERT	0.97	0.94
TF-IDF+SVM	0.84	0.83

Table: GAI subcomponents and their contribution towards the GAI metric

GAI subcomponents	correlation
Agreement	0.887
Clarity	0.800
Certainty	0.772
Specificity	0.398
Factuality	0.204

Results & analysis

Test sample: *We have deployed on-site solar at some of our U.S. distribution centers, including in Arizona, Connecticut, Massachusetts, Nevada, and Texas, as well as at our processing center in Australia. Because we lease, rather than own, nearly all our store locations, we have less flexibility in installing solar on store rooftops. That said, we are pleased to have installed solar at select stores in both the U.S. and the U.K. We continue to engage in conversations with certain landlords to explore the feasibility of installing rooftop solar panels at additional locations.*

- **Classification output:** Greenwashing
- **Specificity:**0.236
- **Factuality:** 0.156
- **Clarity:** 0.7
- **Agreement:**0.5
- **Certainty:** 0.364
- **GAI score**: 0.432

Causal mediation analysis

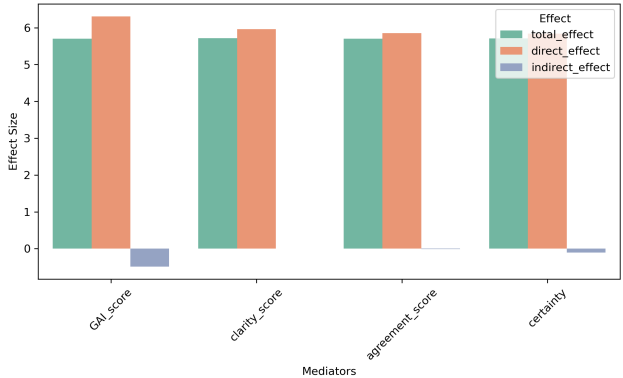

Figure: Bar chart illustrating mediator impacts

- Significant direct effects compared to indirect effects for all classifiers

Key findings

- GAI score aligns with classifier predicted labels
- ClimateBERT model outperformed RoBERTa and TF-IDF+SVM baselines
- Agreement and Clarity has dominant contributions in the GAI metric
- In causal mediation analysis, direct affect influences the outcome of all the three classifiers

Thank you