GreenVerifier: Evaluation of ESG Disclosures with an LLM Approach

University of Southern California

Alexis Moumtzis
moumtzis@usc.edu

Ray Zhang
zhangray@usc.edu

Abstract

ESG reports are an increasingly important source for disclosing a company's environmental, social, and governance (ESG) performance, with many stock exchanges now mandating them for listed companies. However, the unstructured nature of these documents makes it difficult to quantify disclosure levels and assess performance, limiting their ability to substantiate a company's claims.

1 Introduction

GREENVERIFIER automatically verifies corporate sustainability claims against evidence in company reports. Given (i) a short claim, e.g., "Scope 1+2 emissions fell 42% vs 2019; net-zero by 2030," and (ii) the corresponding sustainability or TCFD report in PDF, it classifies the claim as Supported, Contradicted, or Not Enough Evidence, while also returning the best supporting or contradicting evidence snippet and a structured KPI extract. This transforms unstructured disclosures into verifiable information, enabling consistent ESG evaluation.

2 Engines and methods

2.1 Data Preparation

The first stage of our pipeline focuses on preparing ESG reports for automated verification. Company sustainability or TCFD reports, typically published in PDF format, are preprocessed to extract machine-readable text. This extracted content is then transformed into vector representations using text embeddings and stored in a retrieval-augmented generation (RAG) database.

2.2 ESG Metadata Processing

To complement text extraction, GREENVERIFIER adds a metadata processing step standardizing ESG information to frameworks like GRI, SASB, and ESRS. It comprises three components: Entities, defining KPI attributes; Extensions, adding domain-specific knowledge and search terms for reporting variations; and Expressions, structuring prompt templates and output formats for consistent verification. This processing boosts both classification accuracy (Supported/Contradicted/Not Enough Evidence) and KPI extraction quality across diverse ESG reports.

2.3 Language Model

GREENVERIFIER relies on a Transformer-based language model to encode claims and report text into embeddings for retrieval and to perform the final classification task, trained in a supervised learning setup with cross-entropy loss, gradient-based optimization, and evaluation through inference on claim–evidence pairs.

3 Hypotheses and Evaluation Framework

We hypothesize that GREENVERIFIER can assess ESG claims with at least 80% macro-F1 when classifying statements as *Supported*, *Contradicted*, or *Not Enough Evidence*, using bootstrapping to compute confidence intervals. For *Supported* or *Contradicted* claims, we expect the top retrieved evidence snippet to match human judgment at $\geq$90% precision and score 80% on a rubric for number, scope, baseline, and unit clarity. For quantitative disclosures, extracted KPI mini-tables should match ground truth with $\geq$95% accuracy within value and unit tolerances.

We verify these hypotheses using a 70/15/15 company-level split, applying classification metrics (accuracy, macro-F1, per-class F1), evaluating evidence quality, and measuring numeric consistency (exact match $\pm$1% and unit error rate). Robustness is tested on messy tables, restatements, and boundary shifts, requiring performance not to drop by more than 10 F1 points than in clean cases.

References

- Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication: https://www.sciencedirect.com/science/article/pii/S1544612324000096
- Optimizing Large Language Models for ESG Activity Detection in Financial Texts: https://arxiv.org/html/2502.21112v1
- ESGReveal: An LLM-based approach for extracting structured data from ESG reports: https://www.sciencedirect.com/science/article/pii/S0959652624040216
- ESG-KIBERT: A new paradigm in ESG evaluation using NLP and industry-specific customization: https://www.sciencedirect.com/science/article/pii/S0167923625000417