

**“Decoding ESG: Measuring Clarity and Greenwashing in Corporate Disclosures”**

**Vishwas Khandelwal, Shravani Mahadeshwar, Padma Priya Botsa**

Under the Guidance of **Prof. Dr. Lucas Böttcher**

**Master in Applied Data Science**

**Batch 2025**

**Submitted to**

**Frankfurt School Of Finance and Management**

**Background**

Environmental, Social, and Governance (ESG) reporting has become a cornerstone of corporate communication and disclosure practices. Large corporations now routinely include ESG-related sections in their annual filings, highlighting sustainability, diversity, ethics, and accountability initiatives. However, while ESG *scores* and *performance data* are widely studied, the linguistic quality and credibility of these disclosures remain largely unexplored. Most ESG research evaluates *what* is reported - metrics, emissions, or governance structures - but not *how* this information is communicated within formal, regulated reports.

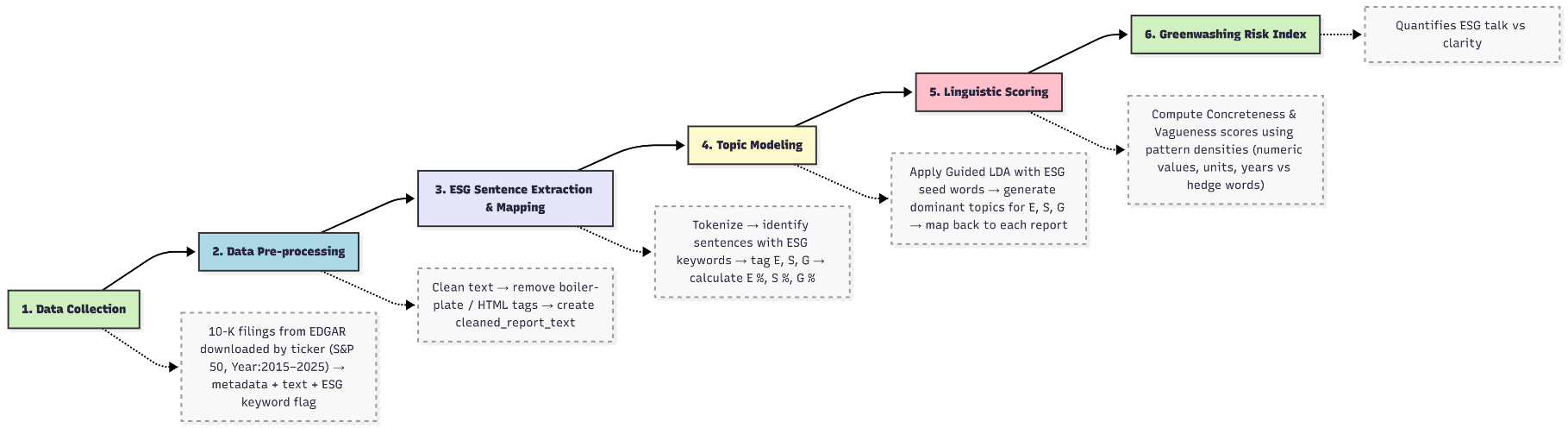
To bridge this gap, our study adopts a language-based analytical perspective using Natural Language Processing (NLP) and probabilistic topic modeling. We analyze SEC 10-K filings from S&P 50 companies between 2015 and 2025, providing a decade-long view of corporate ESG communication in official, audited filings. Since these reports are legally binding and standardized, they provide an ideal foundation for measuring linguistic consistency and transparency in ESG reporting.

**Objective and Framework**

This study aims to:

1. Identify and categorize dominant ESG themes in corporate filings using **Guided Latent Dirichlet Allocation (LDA)**.
2. Quantify the linguistic **clarity and factuality** of ESG communication using a **Concreteness Score**.
3. Measure the **balance between ESG emphasis and clarity** using a custom metric - the **Greenwashing Risk Index (GRI)**.

Our integrated pipeline combines **topic modeling** to uncover “what companies talk about” and **linguistic scoring** to assess “how clearly they communicate it.” This dual approach provides an empirical framework for assessing the **credibility and maturity** of ESG discourse in corporate reporting.



**Mathematical Framework**

1. Guided Latent Dirichlet Allocation (Guided LDA)

We use *Guided LDA*, a probabilistic model that uncovers hidden topics in text while incorporating prior knowledge through seed words.  
It assumes that each document is a mixture of topics and each topic is a distribution over words.

where is the topic mixture for document , and controls how concentrated the topic mixture is.

The topic–word distribution is also drawn from a Dirichlet prior:

where is the word distribution for topic and is a custom prior vector.  
For seed words belonging to predefined ESG topics, higher prior weight is given:

Each word in a document is assigned a topic :

and generated from that topic’s word distribution:

The posterior word probability for topic is given by:

where is the count of word assigned to topic .

The posterior estimate of topic proportion in document is:

2. Concreteness and Vagueness Scoring

The linguistic clarity of ESG sections is quantified by computing densities of numeric, unit-based, and temporal expressions versus hedge or vague terms.

Higher concreteness values imply more measurable, factual language.

3. Token Count and ESG Percentages

To quantify focus across ESG pillars:

These measures show the proportion of text allocated to Environmental (E), Social (S), and Governance (G) topics.

4. Greenwashing Risk Index (GRI)

To evaluate the balance between ESG talk and linguistic clarity:

A higher GRI indicates more ESG emphasis relative to clarity (possible rhetorical inflation), while a lower GRI suggests balanced and credible reporting.

5. Topic Coherence Evaluation

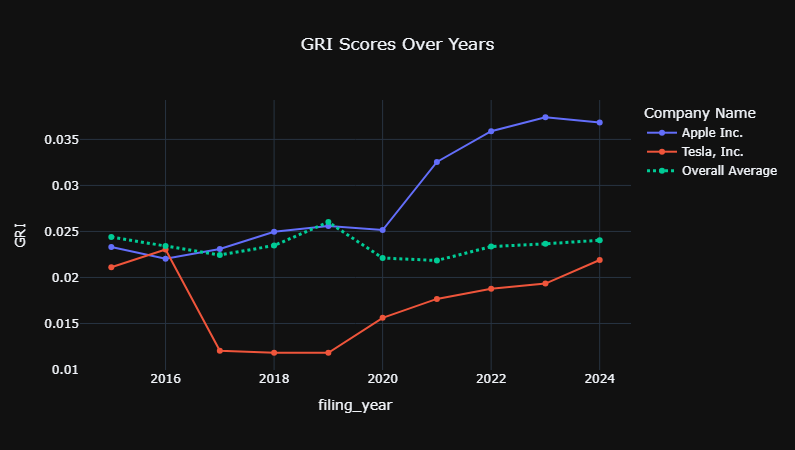
To validate topic quality, we computed Coherence Scores, which measure semantic similarity among top words within each topic.

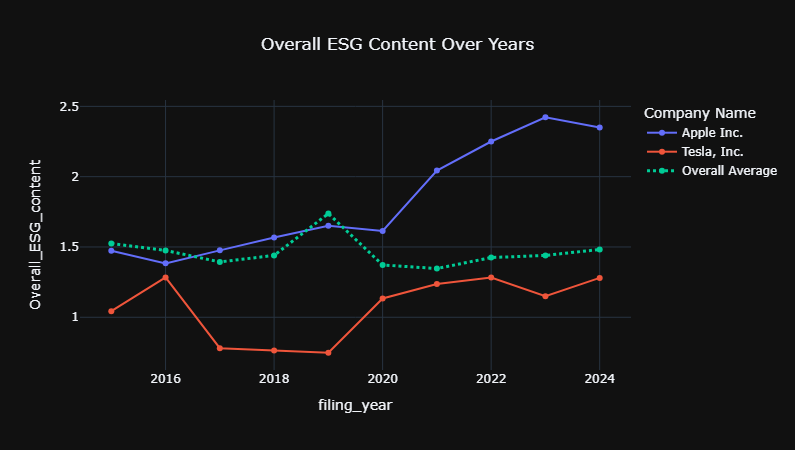
Higher values indicate more meaningful, interpretable topics.

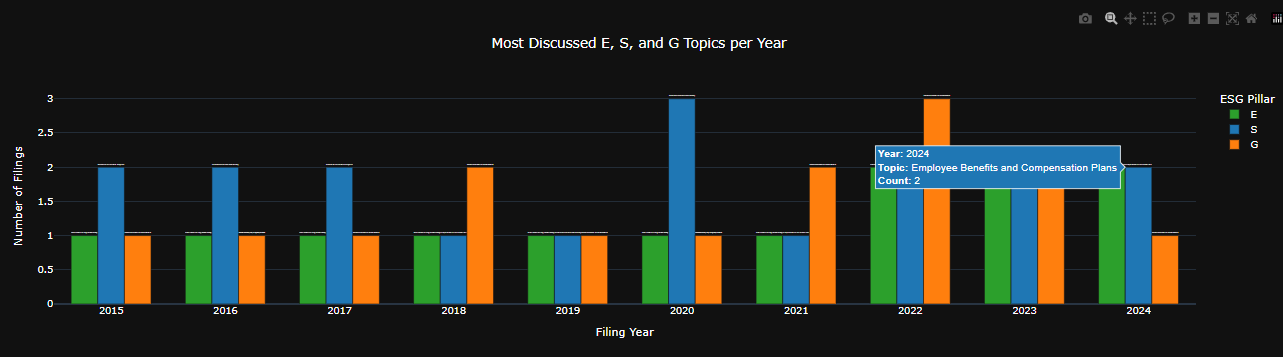
Table

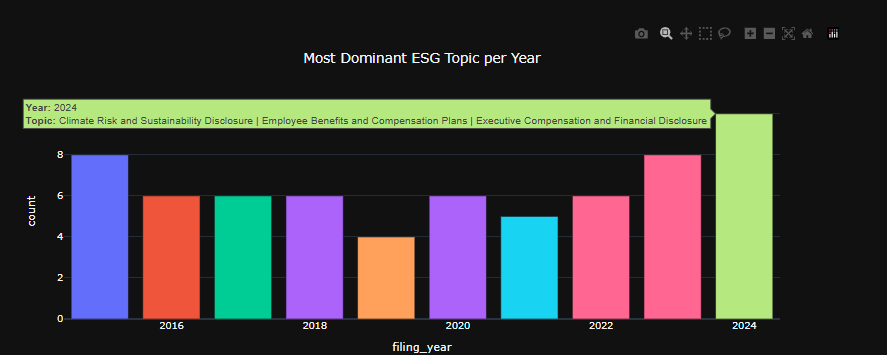
| **ESG Pillar** | **Coherence Score** | **Interpretation** |
| --- | --- | --- |
| **E (Environmental)** | **0.3834** | Moderate coherence due to limited environmental content in filings. |
| **S (Social)** | **0.4459** | Reasonably coherent, covering consistent themes like workforce and inclusion. |
| **G (Governance)** | **0.6397** | Strong coherence, reflecting structured governance disclosures. |

**Results:**

****

****

****

****

**References:**

1. Loughran, T. & McDonald, B. (2021). Measuring Textual Disclosure of Sustainability and CSR in 10-K Filings. Journal of Accounting Research.

2. Li, Y., Chen, H., & Wang, Q. (2023). ESG Disclosures in U.S. Public Filings: An Empirical Text Analysis. SSRN Preprint. <https://ssrn.com/abstract/4402165>

3.Yu, J. & Zhang, L. (2022). Quantifying ESG Reporting in Corporate Annual Filings. Sustainability Accounting Journal, 11 (3), 45-60.

**AI Tools and Assistance Disclosure:**

Some portions of this project and its related materials were refined with the assistance of artificial intelligence tools, including ChatGPT (GPT-5 by OpenAI) and Claude Sonnet (Claude 4.5 by Anthropic). These tools were used to support code refinement, language improvement, and grammatical corrections.

All AI-assisted outputs have been carefully reviewed, validated, and approved by the project team to ensure technical accuracy, quality, and consistency with the project’s objectives. The final responsibility for all content and results remains solely with the authors.

**Code:**

<https://github.com/Vikh110/Decoding-ESG-Measuring-Clarity-and-Greenwashing-in-Corporate-Disclosures>

**Data:**

Data for this study were extracted directly from the U.S. Securities and Exchange Commission (SEC) EDGAR database using the edgar library in Python. Filings were queried by company ticker for all S&P 50 firms, with the Form Type set to 10-K, ensuring a consistent and legally audited dataset.

For each filing, the following metadata were collected: Ticker, Filing Date, Report Date, Accession Number, and the full Report Text. The dataset covers the period 2015–2025, representing ten years of annual filings for each company.

A preliminary ESG keyword flag was created to identify reports containing sustainability-related terminology. This flag was used only for validation purposes and not for filtering, ensuring the dataset remained complete and unbiased.

The resulting corpus comprises the complete 10-K texts for all S&P 50 tickers, forming the foundation for subsequent Natural Language Processing (NLP), Guided Latent Dirichlet Allocation (LDA) topic modeling, and linguistic analysis.