# From Text Analysis to Metrics: AI Approaches to ESG Assessment

Qianxing Chen *

Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, China

* Corresponding Author Email: 23110523d@connect.polyu.hk

**Abstract.** In ESG assessments, disclosure-centric scorecards are being replaced by AI-based sentence-level systems, which offer organized aggregate based on precise criteria. The shift from topic models and human supervision to domain-adaptive Transformer classifiers and evidence-based language models with retrieval capabilities is traced in this work. In order to create trustworthy textual metrics from news, earnings calls, and reports, we concentrate on multi-label phrase annotation by combining non-ESG controls, industry materiality, and financial sentiment. The evaluation takes precision/recall, external validity, and structural variations among rating agencies into account. Decision-making applications in supply chain visibility, investment research, and disclosure assurance are covered, as well as governance elements like documentation, energy use, and human-computer interaction. We conclude by outlining potential avenues for future research and development, such as cross-lingual robustness, retrieval stability, benchmark design, and low-carbon AI. Our suggested "text → metrics → econometrics" approach complements current ratings while improving the scalability, timeliness, comparability, and auditability of ESG measurement.

**Keywords:** ESG measurement, natural language processing, Transformer models, sentence-level multi-label classification, FinBERT sentiment.

## 1. Introduction

In today's society, environmental, social, and governance (ESG) evaluations have gained a lot of attention. Technological developments have replaced human disclosure frameworks and analyst scorecards with automated, AI-driven procedures. The substance that businesses should reveal (such as industry-specific vs common standards) and how rating agencies translated these disclosures into comparable scores shaped early practices. Although this approach has aided in standardizing sustainability reporting, it has also brought to light enduring problems, such as uneven indicators and scope among providers, notable disparities in ratings, narrative-heavy reporting using complicated language, and sluggish update cycles. Because of this, there is a lot of data available for monitoring the ecosystem, but its timeliness, comparability, and auditability differ.

In consideration of this, ESG has progressively incorporated automated methods. Rule-based and vocabulary-based methods, which map checklist items to scores or count keywords, were the first to develop. Subsequently, themes and disclosure intensity were compiled using statistical and topic models (e.g., TF-IDF and LDA/STM). They were wide in scope, but their semantic accuracy was also restricted. Sentence-level, multi-label understanding of E/S/G information is now possible thanks to a recent trend toward deep learning, especially Transformer-based models optimized for ESG. This shift integrates financial sentiment to a moderate degree and modifies sector importance weights according to industrial relevance. In the meantime, structured information extraction and evidence tracking are being facilitated by large language models (LLMs) with retrieval-augmented generation (RAG) capabilities, which promise increased transparency into the ways in which textual statements support scores.

This analysis offers a roadmap for the future development of ESG assessments, connecting manual disclosure requirements and grading techniques with waves of automation: rules → topic models → supervised machine learning → Transformer pipelines → LLM, including multimodal extensions. With an emphasis on managing "non-ESG" content, sentence-level multi-label annotation, and external validity checks for established ratings, the review also improves datasets, annotation schemes,

and evaluation procedures. Third, it methodically determines use cases and consistency—how companies can utilize text-based evaluations for supply chain visualization, investment research, monitoring, and disclosure reviews, as well as how they can be compared to external standards or market variables. Last but not least, it highlights pertinent concerns regarding rating consistency, cross-lingual robustness, explainability, anti-greenwashing proof chains, and low-carbon AI. It also presents a workable agenda for governance, reporting, and benchmarking.

## 2. Historical evolution

### 2.1. Manual Disclosure

The disclosure requirements, which outline what businesses need to report, are the initial part of the ESG. With effect from January 1, 2023, the GRI General Standard (2021) creates a single standard for narrative-based sustainability reporting and redefines material subjects [1]. For periods starting on or after January 1, 2024, the ISSB's IFRS S1 (2023) creates a global standard for data on sustainability-related risks and opportunities that is helpful to primary readers of general-purpose financial reports [2]. The SASB's "Materiality Map/Finder" includes 26 sustainability concerns across 77 industries, operationalizes industry-specific materiality and is still commonly used to match disclosures and metrics with industry relevance [3].

### 2.2. Semi-Automated Evaluations

These frameworks serve as the foundation for rating agencies' systematized scorecards, which combine structured data and rule-based evaluations. By standardizing thousands of variables into problem scores and overall ratings on a 0–10 scale, MSCI evaluates management's exposure to 33 "key issues" [4]. Sustainalytics has developed a versioned methodology and describes "unmanaged ESG risk" for businesses as a fundamental concept [5]. After eliminating dispute, LSEG/Refinitiv presents percentile-ranked category scores and total ESG risk [6]. In order to rectify missing data and guarantee that ratings may be linked to input data supplied by the company, Bloomberg places a strong emphasis on disclosure factors [7]. Hand-coded "strength/concern" indicators were previously given by KLD STATS (1991-2014), which served as the basis for a great deal of empirical research [8]. According to Berg, Kölbel, and Rigobon (2022), who used six lead raters, discrepancies among providers still exist. They broke down these disagreements into three categories: scope, measurement, and weight variations. The biggest contributor was measurement [9].

### 2.3. Early Automation

Initially, researchers used topic models and vocabulary lists to overcome the drawbacks of fixed lists and manual reading. In order to show how themes change over time and within firms, a representative study applied LDA to all corporate social responsibility/environmental, social, and governance (CSR/ESG) reports from major European indices [10]. A follow-up study linked thematic content to information asymmetry by performing sentence-level topic modeling on 1,667 ESG reports (2007-2020). Richer linguistic representations are sought after because these statistical techniques increase coverage and comparability but have trouble with negation, context-dependent governance language, and cross-lingual transfer.

### 2.4. Transformer

After that, the field moved toward Transformers that have already been trained specifically for ESG text. A well-known pipeline pre-trained ESG models on more than 13.8 million news texts and reports, creating three 2,000 expert-labeled datasets for classification at the sentence level and combining them into company-year metrics that were verified in panel regressions against Bloomberg/Refinitiv/RobecoSAM [11]. This pipeline proved that fine-grained, auditable building

blocks for communication intensity and cross-pillar share measures are provided by sentence-level multi-label classifiers.

### 2.5. Customization of Domains

Domain priors have been incorporated into the Transformer through further development. By incorporating FinBERT's sentiment model and SASB's importance weights, ESG-KIBERT improves BERT with a keyword-guided (hard) attention mechanism derived from the standard setter's vocabulary. It also reports higher accuracy, especially lowering "no/non-ESG" false positives that can taint aggregated results [12]. In order to facilitate repeatable evaluation, community-shared tasks (FinNLP FinSim4-ESG, IJCAI-ECAI 2022) further formalize rich and durable sentence prediction benchmarks for ESG taxonomies [13].

### 2.6. LLM-Aided Extraction

Evidence-based extraction has started to replace categorization in ESG text processing with the emergence of LLM. LLM and retrieval-augmented generation (RAG) were used by Bronzini et al. (2024) to extract structured ESG efforts from sustainability reports. They demonstrated the range of subjects outside current taxonomies by graphical analysis [14]. Evidence-based fields can help with downstream scoring and auditing, as demonstrated by ESGReveal's proposed LLM+RAG system, which achieved extraction accuracy of roughly 77% and disclosure analysis accuracy of roughly 84% for Hong Kong Stock Exchange reports [15].

ESG assessment has progressed beyond the ISSB and GRI baselines, starting with manual standards and analyst scorecards and continuing with statistical text mining and, more recently, domain-adapted Transformers and LLM-based extraction. This trend illustrates a broader shift in natural language processing from bag-of-words to retrieval-grounded generation to contextual encoders. For fine-grained evaluation, current best practices in technique combine sentiment modulation and materiality-aware weighting with sentence-level, multi-label classification to capture relevance and intensity. External validity checks against several third-party ratings and careful interpretation of cross-provider divergence are two strategies to improve credibility [9]. While this evolution resolves longstanding difficulties with size, speed, and comparability, new governance requirements pertaining to transparency, audit trails, and greenwashing detection are beginning to be addressed by evidence-linked, RAG-enabled LLM procedures.

## 3. Annotation, Evaluation, and Datasets

The three text sources that are typically incorporated into ESG pipelines are news, dividend conference transcripts, and annual and sustainability reports; third-party ratings serve as external controls. The phrase level is often used to separate large-scale corpora in order to offer balanced industry coverage for downstream analysis to represent industry materiality. The use of manually coded indicators such as KLD (ESG) STATS, 1991–2014, as heritage "truth" or counterfactual controls for automatic labeling has also been common in the past [8].

Sentence-level multi-label annotation, which covers E/S/G and expressly includes a "None/Non-ESG" category to prevent the over-inclusion of irrelevant items when aggregated to the company-year level, is now the most frequently recognized approach [11, 12]. In reality, expert arbitration is frequently conducted after this has been pre-screened using a terminology from an ESG glossary or a standard-setting organization. Categories (E/S/G), action/goal cues, and optional sentiment or intensity tags are examples of annotation dimensions [13]. The dataset employs stratified sampling by report chapter to reduce bias and caps the number of sentences per firm due to the repetitive and templating nature of disclosure material [3].

At both the label and macro levels, studies usually report Cohen's $\kappa$ or Krippendorff's $\alpha$. Disagreements usually revolve around challenging topics like negation, forward-looking claims, and

language pertaining to governance [10]. To increase reproducibility, they also provide annotation rules and positive/negative examples, and they monitor class imbalance, especially G vs. None [13].

Sentence-level results are integrated with SASB industry weightings to create company-year signals like cross-pillar share, communication intensity, and weighted sentiment intensity. Credibility is evaluated using two-way fixed-effects panel regressions and correlations with top institutions (Bloomberg, MSCI, LSEG/Refinitiv, Sustainalytics), which are interpreted in light of recognized sources of rating disagreement (weight, range, and metric). In panel testing, robust standard errors are frequently clustered at the company or company × industry level; a temporal training/validation/test split stops information from leaking from text within the same year.

## 4. Methodology Taxonomy for ESG Evaluation

In order to transform phrases into auditable firm-year indicators, this part systematizes techniques from scorecards to Transformer pipelines and LLM+retrieval, building on datasets and labeling standards.

The question of what to measure is central to ESG assessment. Rating agencies operationalize the framework established by disclosure standards. The ISSB's IFRS S1 creates a global baseline for sustainability data used in decision-making, whereas the GRI Common Standard specifies modular themes and a general reporting structure. Many pipelines then employ the industry-specific correlations that are provided by the SASB's materiality map/finder to modify features and weightings. These inputs are converted into easily manageable scores using major ratings agency methodologies: Bloomberg stresses disclosure factors, so scores remain traceable to the areas reported by companies; LSEG/Refinitiv reports percentile ranking pillars and an overall ESG score, ignoring controversies; Sustainalytics treats analysis as unmanaged ESG risks; and MSCI aggregates exposure and management of issue-level indicators into a 0–10 sub-score and issuer rating. This hierarchy breaks down vendor divergence into variations in scope, measurement, and weighting while remaining open about the inputs and incorporating subjective assessments of scope and weighting [9].

Prior to the development of deep context encoders, lexicons, TF-IDF, and topic models were employed by academics for extensive analysis. They identified key themes and their changes over time and across firms based on a multi-year CSR/ESG corpus [10]. In large report sets, topic content and information friction were connected using sentence-level topic models. Although these techniques increased coverage, they had trouble with boilerplate, negation, and context-dependent governance language.

Now, pre-trained Transformers can classify sentences spanning E, S, and G using multiple labels. According to Schimanski et al. (2024), the representative pipeline is pre-trained on a corpus of millions of sentences from news and reports, refined using expert labels, aggregated to company-year metrics, and validated for external validity against multiple raters using a panel regression model. The prediction unit has clear aggregation criteria and can be audited.

Relevance is improved by financial semantics and domain priors. Standard-setting vocabulary in the encoder is given priority using a keyword-guided attention mechanism [12]. FinBERT uses domain sentiment to control strength [16]. Aggregation is matched with industry significance using a weighting system based on the SASB [3]. The evaluation technique and label space are standardized by shared tasks like FinSim4-ESG [13].

Labels give way to structured fields in retrieval-enhanced LLMs. Goals, strategies, and timetables can be extracted while maintaining verbatim evidence anchors, as case studies show [14]. Exchange-specific corpora exhibit correctness at the field level, making them appropriate for citation assurance and subsequent scoring [15].

FinBERT sentiment models, SASB weighted aggregation, keyword-guided attention, sentence-level multi-label classification with explicit non-ESG categories, domain-adaptive pre-training, and multi-provider validity checks are all examples of practical design patterns. In situations where auditability is necessary, evidence association extraction is added.

By matching expected units with actual presentation in disclosures, sentence-level modeling lowers the possibility that a few ESG-rich portions may overshadow the remainder of the report [11]. Additionally, it makes error identification and rectification manageable because auditors may readily identify misclassifications by linking them to certain sentences. Thresholding and aggregation that are transparent. Unambiguous counting methods, industry important weights, and unambiguous inclusion thresholds should be used to create features like communication intensity, sentiment-weighted intensity, and cross-pillar share after classification in order to guarantee reproducibility across several runs. In order to deduplicate several phrases within a single section, resolve conflicting signals, and eliminate non-ESG content, teams should disclose concise aggregation techniques.

According to Schimanski et al. (2024), threshold calibration and targeted data augmentation are still required since high-precision pipelines still have trouble with negation, forward-looking promises, and governance wording that is dependent on local law. In order to reduce over-inclusion without stifling important disclosures, curators can modify criteria, do recurring confusion audits on the Non-ESG class, and keep a limited library of hard negatives and borderline positives. Although scope and measurement decisions will continue to create cross-provider divergence, comparability will improve until extraction and aggregation are recognized as first-class artifacts [9]. Users can analyze drift and ask others to reproduce results using their own corpora by publishing versioned label guides, model cards, and change logs.

## 5. Risk, Governance, and Countering Greenwashing

NLP-specific failure modes and inter-rater heterogeneity are inherent to automated ESG. The validation story has to recognize that these differences are structural rather than coincidental. Large models raise the possibility of retrieval drift and hallucinations, whereas disclaimers, linguistic drift, and governance phrasing can deceive generalized models [11].

Models with standards-based features match disclosures that are helpful in making decisions. Topics and disclosures are arranged for comparability using the GRI methodology. The information that main users need to make financial decisions is made clear by the ISSB baseline.

Data lineage, sampling, labeling criteria, model versions, parameters, and energy/computational footprints should all be documented by governance. An upgrade path and a brief checklist for these disclosures are provided by the AI-ESG protocol.

Outcomes connected to evidence improve auditability. A human-checkable trail is maintained using retrieval-based extraction, which also highlights discrepancies between statements and deeds [14]. To ensure reproducible outcomes, the production system keeps anchor points and hashed source versions [15].

## 6. Challenges and Future Directions

After establishing the boundaries, we list unresolved issues and viable research avenues to make ESG NLP decision-grade across disclosure styles, industries, and languages. Since disclosures are a combination of languages and styles and governance language is context-sensitive, cross-lingual robustness is still inconsistent as of right now [11]. Instead of merely using ad hoc weighting, the generalization of materiality perception necessitates the use of common standards to connect sentence evidence to industry-specific themes. When non-ESG content predominates, it is still difficult to balance recall and precision, even with keyword-guided attention and sentiment modules [12]. Although FinBERT adds extra hyperparameters that need documentation and stress testing, it aids in providing strength cues [16]. Reliable retrieval, consistent citations, and repeatable logging are essential for large-scale evidence context extraction [14]. Field-level extraction using evidence anchors and sentence-level multi-label classification with non-ESG categories should be covered by community benchmarks [13].

## 7. Conclusions

This review shows that accurate ESG text measuring is a multi-layered procedure rather than a single model. Modern natural language processing (NLP) offers the semantic granularity to evaluate disclosures at the sentence level, industry materiality maps offer domain relevance, and standards like GRI and ISSB benchmarks specify what should be reported. In reality, a robust pipeline refines a domain-adapted encoder, aggregates using auditable, materiality-aware algorithms, and divides data into sentences while maintaining clear non-ESG categories. Measurement findings can be understood across the same system when evidence-related extraction capabilities (paragraph or page anchors stored alongside normalized variables) are introduced. This is essential for regulatory investigations, board reporting, and attestation. Instead of copying a single provider, validation should be seen as a convergence of approaches since structural variations can result from choices about measurement and scope. By recording energy/compute footprints, versioning data and models, releasing dynamic labeling guidelines, and conducting sensitivity tests on alternative aggregation rules, the team is improving operational robustness. Building a multilingual corpus and benchmarks with verified evidence anchors, improving the retrieval stability of RAG-assisted extraction, and standardizing the generalization of materiality perceptions to facilitate more insightful cross-industry comparisons are the explicit strategic goals. These protections allow automated text-based measures to supplement current ratings and lower the cost of guaranteeing the timeliness, comparability, and auditability of ESG data.

## References


[1] GRI. GRI universal standards 2021 [S]. 2021.

[2] IFRS Foundation/International Sustainability Standards Board. IFRS S1: General requirements for disclosure of sustainability-related financial information [S]. 2023.

[3] SASB/IFRS Foundation. Materiality finder / materiality map [EB/OL]. (n.d.).

[4] MSCI. MSCI ESG ratings—Methodology [EB/OL]. (n.d.).

[5] Sustainalytics (Morningstar). ESG risk ratings—Methodology v3.1 (abstract) [EB/OL]. 2024.

[6] LSEG/Refinitiv. LSEG ESG scores—Methodology [EB/OL]. 2025.

[7] Bloomberg. Environmental & social (ES) scores—Methodology overview [EB/OL]. (n.d.).

[8] MSCI/KLD. KLD (ESG) STATS—Data overview (1991–2014) [EB/OL]. 2015.

[9] Berg F, Kölbel J F, Rigobon R. Aggregate confusion: The divergence of ESG ratings [J]. Review of Finance, 2022, 26(6): 1315–1344.

[10] Goloshchapova I, Poon S-H, Pritchard M, Reed P. Corporate social responsibility reports: Topic analysis and big data approach [J]. The European Journal of Finance, 2019, 25(17): 1637–1654.

[11] Schimanski T, Reding A, Reding N, Bingler J, Kraus M, Leippold M. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication [J]. Finance Research Letters, 2024, 61: 104979.

[12] Lee H, Kim J H, Jung H S. ESG-KIBERT: A new paradigm in ESG evaluation using NLP and industry-specific customization [J]. Decision Support Systems, 2025, 193: 114440.

[13] Kang J, El Maarouf I. FinSim4-ESG shared task: Learning semantic similarities for the financial domain [C]// ESG insights. 2022: 211–217.

[14] Bronzini M, Nicolini C, Lepri B, Passerini A, Staiano J. Glitter or gold? Deriving structured insights from sustainability reports via large language models [J]. EPJ Data Science, 2024, 13(1): 41.

[15] Zou Y, Shi M, Chen Z, Deng Z, Lei Z, Zeng Z, Yang S, Tong H, Xiao L, Zhou W. ESGReveal: An LLM-based approach for extracting structured data from ESG reports [EB/OL]. arXiv:2312.17264, 2023.

[16] Araci D. FinBERT: Financial sentiment analysis with pre-trained language models [EB/OL]. arXiv:1908.10063, 2019.