

Title Page

Scaling Sustainability Analysis: An Expert-Seeded AI Framework for Evaluating Disclosure Quality in CSRD Reports

Authors

Madhavan Vishnu Nampoothiri^a, Mohit Kumar^b

Affiliations

^a Xavier Institute of Management and Entrepreneurship (XIME), Kinfras Hi-Tech Park, Off HMT Rd, HMT PO, Kalamassery, Kochi, Kerala 683503, India.

Email: madhavanvee@gmail.com

(Corresponding author)

^b Department of Economics and Finance, Birla Institute of Technology & Science Pilani, K.K. Birla Goa Campus, Near NH 17B, Zuarinagar, Sancoale, Goa 403726, India.

Email: mohitk@goa.bits-pilani.ac.in

Declarations of interest: None

Scaling Sustainability Analysis: An Expert-Seeded AI Framework for Evaluating Disclosure Quality in CSRD Reports

Abstract

The EU's Corporate Sustainability Reporting Directive (CSRD) creates an unprecedented analytical challenge: processing thousands of complex narrative reports annually to assess disclosure quality. This paper introduces an AI-driven framework that makes this task tractable. We develop an expert-seeded, AI-validated approach that combines regulatory expertise with corpus-specific Word2Vec embeddings to construct context-specific disclosure measures. Our framework makes two methodological contributions. First, it introduces the distinction between strategic and compliance orientation, a feature absent in prior sustainability textual analysis but central to mandatory disclosure regimes. Second, it achieves marked efficiency while maintaining complete interpretability, with every score traceable to specific words.

Trained on 592 inaugural CSRD reports, the framework processes in hours what manual analysis would require months, supported by systematic validation including human-in-the-loop refinement. An empirical demonstration on 446 firms illustrates its pattern detection capabilities, revealing subtle governance dynamics not readily apparent through manual review. The primary contribution is providing scalable, transparent infrastructure for analyzing mandatory sustainability disclosures. While developed for CSRD, the framework's modular design allows adaptation to other regulatory contexts by modifying only the seed dictionaries.

Keywords

Sustainability Reporting; CSRD; Artificial Intelligence; Textual Analysis; Disclosure Quality; Corporate Governance;

1. Introduction

The European Union's Corporate Sustainability Reporting Directive (CSRD) presents a formidable computational challenge for sustainable finance. This new regulation requires thousands of European companies to produce detailed sustainability reports annually, creating an unprecedented volume of complex narrative data to analyze. While recent regulatory adjustments through the Omnibus Regulation are expected to reduce the scope from an initial 50,000 to less than 10,000 firms, the analytical burden remains substantial and compounds annually. Thousands of complex reports will be added every year to the analytical corpus. Longitudinal analysis is essential for tracking sustainability progress, identifying greenwashing patterns, and assessing regulatory effectiveness. Yet conducting such analysis requires processing these tens of thousands of documents. Companies produce comprehensive reports, typically 50 to 200 pages of sustainability content embedded within larger integrated reports, that blend strategic narratives with compliance documentation across multiple environmental, social, and governance dimensions.

These reports represent a fundamental departure from traditional financial disclosures. They must satisfy the double materiality principle, assessing both how sustainability matters affect the company (financial materiality) and how the company affects society and environment (impact materiality). They require extensive forward-looking information including transition plans, targets, and strategic commitments. They span diverse topics from climate change and biodiversity to human rights and business conduct, each with specific technical vocabulary and measurement requirements. The resulting documents mix quantitative metrics with qualitative assessments, regulatory compliance language with strategic positioning, and historical performance with future commitments, thereby creating analytical complexity that can overwhelm traditional assessment methods.

Traditional manual analysis requires expert readers to spend tens of hours per report, limiting scalability. Assessing the 592 early-adopter reports from 2024 alone would demand tens of thousands of hours, and covering the full CSRD universe of nearly 10,000 firms annually would require hundreds of analysts working full-time. When considering the need for longitudinal analysis to track disclosure evolution, commitment fulfilment, and regulatory impact over multiple years, the analytical challenge can become insurmountable through conventional means.

This paper introduces and validates a novel AI-driven framework that transforms this challenging task into a tractable one. Our methodology, which we term 'expert-seeded, AI-validated' textual analysis, achieves an order-of-magnitude gain in efficiency compared to manual expert review. After initial setup, our framework processes 592 reports in a matter of hours, a task that would require months of expert labor if performed manually. More importantly, our approach moves beyond simple keyword counting to capture granular attributes of disclosure quality that careful human readers might miss, such as the subtle balance between strategic and compliance language or the density of concrete forward-looking commitments versus vague aspirations.

Our framework makes two key methodological contributions. First, we develop an 'expert-seeded, AI-validated' approach that combines deep regulatory expertise with machine learning to create context-specific measurement tools for CSRD's unique linguistic environment. Rather than applying generic sentiment dictionaries or off-the-shelf text analysis tools, we begin with systematic analysis of all ESRS standards to identify regulatory terminology, then use corpus-specific Word2Vec models to discover how companies translate these requirements into corporate narrative. Our validated measures capture the distinction between strategic and compliance orientation, an important differentiation for assessing whether CSRD achieves strategic integration or mere compliance.

Second, we achieve high level of scalability while maintaining interpretability. Processing 592 reports in hours rather than months, our method can analyze the entire CSRD universe with every score traceable to specific words, not black-box predictions. This combination of scale and transparency enables detection of subtle patterns normally invisible to manual analysis, opening previously intractable research questions about disclosure dynamics in mandatory reporting regimes.

To develop and demonstrate our framework, we assembled a comprehensive corpus of 592 inaugural CSRD reports representing approximately 60,000 pages of sustainability disclosure. This corpus, drawn from early adopters across 27 EU member states and all major industries, captures the full linguistic diversity of this new regulatory regime. We train our language models on this complete corpus, ensuring our measures reflect actual corporate usage rather than theoretical regulatory language. We then apply our validated measures to a subset of 446 firms with complete financial and governance data from Bloomberg. The results offer a different view of disclosure quality determinants, showing that traditional resource proxies like firm size and age matter less than governance dynamics in this highly regulated context.

The implications extend far beyond our specific empirical findings. Our framework provides regulators with infrastructure to monitor compliance patterns systematically, identify emerging interpretation issues, and assess whether the CSRD achieves its strategic integration goals. It offers investors tools to evaluate the substance behind sustainability narratives, distinguishing genuine strategic commitment from sophisticated compliance exercises. For researchers, it opens previously intractable questions about disclosure evolution, the relationship between narrative and performance, and the effectiveness of mandatory sustainability reporting. As the CSRD corpus grows annually, our framework makes longitudinal analysis feasible. It enables tracking how disclosure language evolves, whether commitments materialize, and how regulatory learning occurs, with the same efficiency as analyzing a single year's reports.

The paper proceeds as follows. Section 2 motivates the need for AI-driven analysis in the CSRD context and reviews relevant literature. Section 3 details our methodological innovation, from corpus preparation through validation. Section 4 demonstrates the framework's capabilities through empirical application. Section 5 discusses implications for stakeholders and future development. Finally, Section 6 concludes with a summary of our findings and implications.

2. Background and Literature Review

2.1 Evolution of Textual Analysis Methods in Accounting and Finance

The application of textual analysis to corporate disclosures has evolved dramatically over the past two decades, progressing from simple word counts to sophisticated machine learning approaches. Early work by Li (2008) demonstrated that linguistic complexity in 10-K reports predicted earnings persistence, establishing that textual attributes contain information beyond numerical data. This foundational insight motivated widespread adoption of computational linguistics in accounting research.

The field's transformation accelerated with Loughran and McDonald (2011), who showed that applying general-purpose sentiment dictionaries misclassifies many words in financial texts, and developed domain-specific word lists better suited for accounting and finance. Their finance-specific dictionary became the standard for sentiment analysis in accounting, demonstrating that domain-specific language models significantly outperform generic alternatives. However, as Loughran and McDonald (2016) emphasize in their survey, static dictionaries face inherent limitations: they cannot adapt to evolving terminology, struggle with context-dependent meanings, and require continual manual updates.

These limitations prompted researchers to explore machine learning alternatives, as detailed in the review of Machine Learning applications in accounting and finance by Guo, Shi, and Tu

(2016). Hanley and Hoberg (2010) implemented a word content analysis of IPO prospectuses, decomposing text into ‘standard’ versus ‘informative’ content, and demonstrated that informative disclosure predicted more accurate offer pricing and reduced underpricing. Dyer, Lang, and Stice-Lawrence (2017) employed Latent Dirichlet Allocation (LDA) to identify topical evolution in 10-K disclosures over time. While these unsupervised methods captured patterns invisible to dictionary approaches, they often produced topics difficult to interpret and validate against regulatory requirements. This is a particular challenge for compliance-oriented analysis

2.2 Word Embeddings: From Linguistic Theory to Practical Application

The theoretical foundation for modern word embedding techniques traces to Harris (1954), whose distributional hypothesis established that words appearing in similar contexts tend to have similar meanings. This insight, formalized through neural network architectures by Mikolov et al. (2013) in Word2Vec, revolutionized textual analysis by learning semantic relationships directly from text corpora. Unlike dictionaries that treat words as discrete units, embeddings capture meaning through context, with words appearing in similar contexts receiving similar vector representations in high-dimensional space.

Li et al. (2021) demonstrated word embeddings' potential in measuring corporate culture from earnings calls, showing that firms with similar cultural profiles exhibited correlated performance patterns. Their approach combined expert-defined cultural dimensions with Word2Vec expansion, validating that machine-learned associations aligned with human understanding while discovering non-obvious relationships. This hybrid methodology, combining expert seeding with algorithmic expansion, represents an advance over purely automated or purely manual approaches.

Most relevant to our work, Lin et al. (2024) applied word embeddings to analyze global evolution of environmental and social (E&S) disclosure across 210,000 annual reports. Their methodology addresses three challenges that parallel those in CSRD analysis: (1) the multifaceted nature of E&S topics spanning diverse technical domains, (2) rapidly evolving terminology as sustainability concepts mature, and (3) the need to classify concepts with ambiguous boundaries. They validate their approach through extensive human coding, achieving κ statistics of 0.90 for overall E&S identification but only 0.52-0.71 for subtopic classification thereby establishing important benchmarks for machine-human agreement in this domain.

2.3 The Sustainability Disclosure Challenge: From Voluntary to Mandatory

The transition from voluntary to mandatory sustainability disclosure creates unique methodological challenges for textual analysis. Under voluntary regimes, firms' discretion enabled selective reporting and "greenwashing" through vague, boilerplate language (Hummel & Schlick, 2016). The EU's Non-Financial Reporting Directive (NFRD), effective from 2017, began addressing these issues through mandatory disclosure requirements. However, research revealed mixed results. While disclosure quantity increased substantially, quality improvements remained elusive. Jackson, Bartosch, Avetisyan, Kinderman, and Knudsen (2020) found that firms complied with the letter but not spirit of requirements, suggesting that simply mandating disclosure doesn't guarantee meaningful communication (Christensen, Hail, & Leuz, 2021 ; Vishnu Nampoothiri, Entrop and Annamalai, 2024).

The Corporate Sustainability Reporting Directive (CSRD) represents a fundamental structural break requiring new analytical approaches. Unlike NFRD's principles-based framework allowing substantial discretion, CSRD mandates compliance with detailed European Sustainability Reporting Standards (ESRS) containing hundreds of technical terms with precise

legal definitions. The directive's emphasis on double materiality (examining both corporate impacts on society and society's impacts on corporate value), forward-looking information, and strategic integration creates analytical complexity that overwhelms traditional assessment methods. With nearly 10,000 firms producing 100–150-page reports annually, manual analysis becomes prohibitively labour and time intense.

2.4 Measuring Disclosure Quality in Mandatory Regimes

Recent methodological advances enable quality assessment beyond simple volume metrics, particularly crucial in mandatory regimes where quantity becomes standardized. Lang and Stice-Lawrence (2015) developed measures for detecting boilerplate language, which refers to generic, recycled text providing minimal information value. They did so by identifying phrases repeated across firms and time periods Hope, Hu, and Lu (2016) complemented this by measuring disclosure specificity through references to particular people, places, organizations, times, and numbers, finding that specific disclosures correlate with improved analyst forecasts and reduced information asymmetry.

These quality dimensions become especially relevant for CSRD analysis, where firms must translate regulatory requirements into business language while choosing between strategic framing and compliance orientation. The CSRD context introduces unique measurement challenges absent in voluntary reporting. While Lin et al. (2024) successfully measure general E&S disclosure presence, CSRD has specific disclosure requirements that necessitates distinguishing strategic integration from mere compliance. This distinction is important for assessing whether the directive achieves its transformative goals. Moreover, the tension between board monitoring responsibilities and risk management concerns creates competing pressures on disclosure choices (Fama & Jensen, 1983). Independent directors might simultaneously push for strategic, high-quality disclosure while constraining forward-looking

commitments that create legal exposure, which is a dynamic invisible to traditional resource-based explanations of disclosure quality.

2.5 Validation Standards and the Human-Machine Complementarity

Establishing validity for automated textual measures requires careful comparison with human judgment. The literature reveals considerable variation in acceptable agreement levels, reflecting task complexity and domain specificity. Hanley and Hoberg (2010) validated their textual measures by linking informative disclosure to economic outcomes, showing significant predictive power for IPO pricing accuracy and underpricing. Lin et al. (2024) achieve κ statistics of 0.90 for identifying E&S content presence but only 0.52-0.71 for classifying specific subtopics, highlighting how agreement deteriorates with classification granularity.

These moderate agreement levels reflect fundamental differences between human and machine processing rather than methodological failure. Humans form holistic impressions influenced by narrative flow and emphasis, while machines systematically process every term with equal weight. Recent literature increasingly frames this as complementary strengths, with machines providing perfect consistency and unlimited scalability, while humans contributing subtle understanding that no algorithm can fully replicate. As Berg, Koelbel, and Rigobon (2022) and Freiberg, Park, Serafeim, and Zochowski (2021) document, this complementarity becomes essential given the limitations of purely human-based ESG ratings or commercial databases, which suffer from low inter-rater reliability and lack transparency.

2.6 Positioning Our Methodological Contribution

Our methodological approach builds directly on the framework of Lin et al. (2024), who combined expert seeding with Word2Vec embeddings to construct empirically validated dictionaries for E&S disclosure. We diverge in two ways that make our framework uniquely suited for CSRD analysis. First, unlike Lin et al.'s broad E&S categorization, we introduce a

new conceptual axis—strategic versus compliance orientation—that is absent in prior work but important for assessing whether CSRD reporting represents strategic integration or boilerplate adherence. Second, while Lin et al. train on a heterogeneous mix of voluntary and mandatory disclosures from multiple regulatory regimes, we train exclusively on CSRD reports to capture regulation-specific linguistic patterns, such as the translation of “adequate wages” (ESRS terminology) into “competitive compensation” in corporate narrative.

At the same time, our work takes a different path from recent domain-specific transformer models such as ClimateBERT (Webersinke et al., 2021) and FinBERT (Araci, 2019). While these models demonstrate the power of contextual embeddings, they are limited in scope. For example, ClimateBERT covers only a subset of environmental topics and has little to no coverage of social or governance issues, while FinBERT focuses on financial sentiment without sustainability vocabulary. More importantly, transformer models provide only post-hoc interpretability tools (e.g., attention weights, feature importance), whereas our approach produces transparent, dictionary-based word-to-word relationships. This interpretability is essential for regulatory applications where every classification must be traceable and defensible.

3. Methodological Framework: Expert-Seeded, AI-Validated Textual Analysis

3.1 Overview and Data Collection

Our methodological framework addresses the challenge of analyzing thousands of complex sustainability reports through a novel combination of human expertise and machine learning, termed "expert-seeded, AI-validated" textual analysis. The framework consists of four integrated components: (1) systematic corpus preparation isolating relevant sustainability content from integrated annual reports, (2) context-specific language modeling capturing CSRD-specific vocabulary, (3) expert-seeded dictionary construction validated through

machine learning, and (4) automated quality measurement maintaining interpretability while achieving large-scale processing capability.

This framework advances existing textual analysis methods in three ways. First, unlike static dictionaries (Loughran & McDonald, 2011), we dynamically expand terms based on corpus-specific semantic relationships, capturing how CSRD language actually functions rather than imposing predetermined categories. Second, while Lin et al. (2024) measure general E&S disclosure presence, we specifically distinguish strategic from compliance orientation. This is an important distinction for assessing whether CSRD achieves strategic integration or mere compliance. Third, unlike transformer models such as ClimateBERT that operate as black boxes even with post-hoc explainability tools, our dictionary-based approach maintains complete transparency, allowing users to trace every score to specific words, which is essential for regulatory and investment applications.

We assembled a comprehensive corpus from the Sustainability Reports Network (SRN) database, identifying 628 inaugural CSRD reports, which were embedded within integrated annual reports for the 2024 financial year. After excluding 36 reports that were not machine-readable, not in pdf format or not in English, we obtained 592 unique reports representing approximately 60,000 pages of sustainability disclosure. These early adopters span 27 EU member states across all major economic sectors.

For empirical demonstration, we required additional firm-level data to test whether our measures detect meaningful disclosure patterns. Matching our textual corpus with Bloomberg financial and governance data revealed further constraints: 50 firms had zero word counts from failed text extraction, and 96 lacked complete governance data, particularly board composition metrics. Our final regression sample comprises 446 firms with complete information across all variables as provided in Table 1.

[Insert Table 1 about here]

This two-stage approach of using 592 firms for methodology development and 446 for empirical demonstration based on data availability ensures our framework trains on the broadest possible representation while maintaining data quality for statistical analysis.

3.2 Sectional Extraction and Text Processing

Our initial application of textual measures to complete integrated reports revealed an important validity threat: forward-looking statements in financial sections (Management Discussion & Analysis) and strategic language in CEO letters contaminated our sustainability-specific measures. For instance, banks' discussions of expected credit losses registered as forward-looking information, while manufacturers' financial risk policies appeared as compliance language, despite neither relating to sustainability. This contamination meant our measures captured characteristics of integrated reports generally rather than sustainability disclosures specifically.

To isolate sustainability narratives, we developed a two-pronged extraction protocol. Automated extraction succeeded for 249 reports (42%) using hierarchical detection rules: structural markers (section headings containing "Sustainability," "Non-Financial Information," "ESG Report," "Corporate Responsibility," or explicit "ESRS"/"CSRD" references), content indicators (calculating density of sustainability terminology and ESRS-specific vocabulary), minimum length requirements (10+ pages to exclude summary discussions), and boundary detection (formatting changes, transition phrases like "End of Sustainability Statement," narrative style shifts).

The remaining 343 reports (58%) required manual extraction due to deeply integrated structures, non-standard formatting, or visual elements disrupting text flow. For these, the first

author, who has worked extensively on CSRD and ESRS, identified precise boundaries of all ESRS-required environmental, social, and governance content, inserting <<<START>>> and <<<END>>> markers for subsequent processing.

This extraction reduced our corpus by approximately 80%, dramatically improving both measurement validity and computational efficiency, transforming an intractable big data problem requiring cloud computing into manageable local processing. While we didn't formally validate completeness, manual review of 30 automatically-extracted reports confirmed sustainability discussions were indeed concentrated in identified sections with minimal relevant content elsewhere.

Our text processing pipeline employs standard NLP transformations via spaCy: sentence segmentation, tokenization, part-of-speech tagging and lemmatization to base forms (ensuring "reporting," "reports," and "reported" are recognized as one concept). We apply Gensim's Phrases algorithm for data-driven n-gram detection, identifying statistically significant multi-word concepts like "climate_change," "double_materiality," and "transition_plan" that carry specific regulatory meaning in CSRD discourse.

3.3 Corpus-Specific Word Embedding Model

The analytical foundation of our methodology rests on an insight: sustainability reporting under CSRD represents a distinct linguistic domain requiring context-specific language models. Word embedding models learn meaning from context. In CSRD reports, words keep very specific company. "Material" rarely refers to physical substances but signals regulatory significance under double materiality. "Transition" relates to climate pathways rather than organizational change. "Scope" denotes emission boundaries rather than project parameters. These domain-specific semantic relationships cannot be learned from general corpora, necessitating a model trained specifically on CSRD disclosures.

Our approach builds directly on Lin et al. (2024), who demonstrated word embeddings' superiority for analyzing sustainability disclosures. They identified three advantages: handling the multifaceted nature of E&S topics with extensive technical terminology, adapting to rapidly evolving sustainability lexicon, and classifying related terms into categories with ambiguous boundaries. These challenges are amplified in the CSRD context. Unlike Lin et al.'s voluntary reports where firms had discretion over terminology, CSRD mandates compliance with ESRS standards containing hundreds of technical terms with precise legal definitions. "Double materiality" involves specific assessment procedures and documentation standards. "Transition plan" must include elements defined in ESRS E1—emission targets, decarbonization levers, and CapEx allocations. This regulatory specificity creates specialized language requiring context-specific learning.

We implement Word2Vec using skip-gram architecture with negative sampling, following Mikolov et al. (2013). Skip-gram learns word representations by predicting context words given a target word. It is particularly effective for specialized corpora where technical terms appear infrequently but in highly informative contexts. When processing "taxonomy-aligned," the model learns to predict surrounding words like "CapEx," "technical screening criteria," and "DNSH," placing it in semantic space near investment and compliance terminology, distant from generic strategy language.

Our implementation uses 300-dimensional vectors, balancing expressiveness with computational efficiency. The model trains on 592 sustainability sections comprising approximately 8 million tokens. We employ a context window of 10 words which is larger than typical implementations. We do so because sustainability reporting involves complex sentences where related terms are separated by qualifying phrases. Minimum word frequency is set at 5 occurrences; less frequent terms lack sufficient context for reliable vectors.

The decision to train corpus-specific models rather than using pre-trained alternatives requires justification. We evaluated several options:

Pre-trained generic models (Google News Word2Vec, Stanford GloVe) were not considered, as their vocabularies are not designed for regulatory contexts and omit many ESRS-specific terms.

Domain-specific models like ClimateBERT (Webersinke et al., 2021) and FinBERT (Araci, 2019) offered initial promise. ClimateBERT specializes in climate-related text but covers primarily environmental topics, while social and governance vocabulary is limited. FinBERT excels at financial sentiment but was not trained on sustainability terminology. More importantly, while transformer models offer various interpretability methods, they do not provide the direct word-to-word semantic relationships needed for dictionary construction which is a transparency requirement for regulatory applications where classifications must be traceable and defensible.

Topic modeling (e.g., LDA) was unsuitable for our purpose because it is unsupervised, cannot directly incorporate expert-defined ESRS categories, and produces probabilistic topic distributions that require subjective post-hoc interpretation. By contrast, our Word2Vec approach enables expert-seeded categories (strategic vs. compliance, forward-looking) that can be empirically expanded while retaining the interpretability essential for regulatory applications

Our Word2Vec implementation represents a deliberate choice balancing multiple objectives. It provides transparent, interpretable word relationship which enables us to examine which words the model considers similar to seeds and understand why. It enables semi-supervised learning where expert knowledge guides but doesn't predetermine dictionaries. It achieves computational efficiency, training in a few hours on standard hardware rather than requiring

GPU clusters. Most importantly, it learns CSRD-specific semantic patterns, capturing regulatory nuances that generic models miss.

The trained model reveals fascinating semantic structures within CSRD discourse. Strategic terms cluster together—"opportunity" neighbors "potential," "innovation," and "value_creation." Compliance terms form separate clusters—"requirement" associates with "obligation," "mandatory," and "pursuant_to." Forward-looking language creates complex networks—"target" connects to "ambition," "commitment," and "pathway" but also to hedging terms like "expect" and "anticipate." These patterns validate our theoretical distinction between strategic and compliance orientations while revealing nuances like the careful management of forward-looking uncertainty.

This corpus-specific training proves essential for valid measurement. A general language model would miss the specialized meanings embedded in CSRD reports. A voluntary disclosure model would lack the regulatory precision required by ESRS. Our approach captures the unique linguistic environment of mandatory sustainability reporting in the European Union, providing the foundation for dictionaries that meaningfully distinguish strategic integration from compliance orientation.

3.4 Dictionary Construction and Measurement

Our dictionary construction is the bridge between regulatory knowledge and empirical analysis. The CSRD context requires a precise approach: ESRS standards contain hundreds of technical terms that companies translate into varied business language, and distinguishing strategic from compliance orientation demands capturing these subtle linguistic patterns.

We employed a multi-stage process combining expert knowledge with corpus-specific machine learning validation:

Stage 1: Comprehensive Expert Seeding.

We conducted a systematic analysis of ESRS 1–2 (General Requirements/Disclosures), E1–E5 (Environmental), S1–S4 (Social), and G1 (Governance), extracting over 700 terms across four categories. Strategic orientation seeds captured integration language such as “*value creation*,” “*competitive advantage*,” “*market opportunity*,” “*innovation*.” Compliance orientation seeds included regulatory adherence language like “*in accordance with*,” “*pursuant to ESRS*,” “*mandatory disclosure*.” Forward-looking seeds contained temporal markers and commitments such as “*target*,” “*will*,” “*by 2030*.” Topic-specific seeds mapped to individual ESRS requirements (e.g., Scope 1/2/3 emissions, living wage, anti-corruption). This seeding revealed a key insight: much of the precise regulatory terminology (e.g., “*remuneration policies*,” “*collective bargaining coverage*”) rarely appeared verbatim in reports, with firms favoring more business-oriented equivalents.

Stage 2: AI-Driven Corpus Validation.

We validated the seed list using a Word2Vec model trained on our 592-report corpus. A term was considered validated if it appeared at least five times, ensuring stable vector representation. Only 222 (24.3%) of the original seeds “survived” this step, revealing a substantial gap between regulatory and corporate language. Entire ESRS categories were sometimes absent. For instance, no firm used terms like “*adequate wages*” or “*social dialogue*,” preferring “*compensation*,” “*employee engagement*,” “*talent management*.” This stage produced a smaller, empirically grounded core dictionary, forming the basis for the final expansion step.

Stage 3: Contextual Expansion.

We used our corpus-specific Word2Vec model to expand the validated seed list by identifying semantically similar terms. For each seed, we extracted words with cosine similarity > 0.7 which is a conservative threshold chosen after experimentation: lower thresholds (0.5–0.6) introduced noise, while higher ones (0.8+) yielded too few new terms. Strategic seeds expanded

richly (e.g., "*opportunity*" → "*potential*," "*possibility*," "*avenue*"; "*innovation*" → "*technology*," "*solution*," "*transformation*"). Compliance seeds showed limited growth ("*requirement*" → "*obligation*," "*mandate*"). Forward-looking seeds formed broader networks: "*target*" → "*ambition*," "*goal*," "*commitment*" plus hedging language like "*expect*" → "*anticipate*," "*believe*," "*may*," "*might*." Many regulatory terms like "*double materiality*" had no high-similarity matches and remained isolated, highlighting its technical specificity. This selective expansion produced our final dictionaries containing 16 strategic, 16 compliance, and 122 forward-looking terms. This represents a deliberate trade-off: our 154 final terms constitute a small fraction of total corpus vocabulary, but our validation confirms they capture the meaningful variation in disclosure quality.

Stage 4: Human-in-the-Loop Refinement. When S1 yielded zero counts across 592 reports, we augmented seeds with business terminology identified through manual inspection. Re-running expansion with "*compensation*," "*wellbeing*," "*union relations*" successfully captured worker discussions in 87% of reports. This human-in-the-loop approach combines automation with expert judgment, thereby making it neither purely algorithmic nor purely manual.

Disclosure Quality Measures transform counts into meaningful metrics:

$$\text{Strategic_Ratio} = \text{Strategic Word Count} / (\text{Strategic} + \text{Compliance Word Count})$$

This ranges 0 (pure compliance) to 1 (pure strategy), capturing narrative orientation. The ratio design ensures comparability across report lengths. A 200-page report doesn't automatically score higher than 100-page report. By focusing on relative proportions rather than absolute counts, we capture orientation independent of volume.

$$\text{FLI_Score} = \ln(1 + \text{Forward-Looking Word Count}) / \ln(1 + \text{Strategic} + \text{Compliance} + \text{Forward-Looking Word Count})$$

This normalized measure captures forward-looking density. Logarithmic transformation addresses the skewed distribution. Some reports contain extensive future discussion while others minimize forward-looking statements. Adding 1 prevents undefined values for zero forward-looking terms.

These variable measures the forward-looking orientation of the firm's sustainability narrative. It is calculated as the natural logarithm of one plus the raw count of words from our Forward_Looking dictionary, normalized by the natural logarithm of one plus the sum of the word counts from all three of our core linguistic dictionaries (Forward_Looking, Strategic, and Compliance). This ratio captures the extent to which the core sustainability narrative is focused on future plans and commitments, controlling for the overall volume of the measured disclosure.

These measures capture distinct constructs. Reports can be highly strategic (discussing sustainability as opportunity) while minimizing forward-looking commitments. Conversely, compliance-focused reports might include extensive forward-looking information if regulations require specific targets. Our empirical analysis confirms these dimensions are indeed distinct, with different determinants.

The complete dictionary construction process balances competing objectives: regulatory grounding versus empirical prevalence, comprehensiveness versus precision, stability versus evolution. Our approach prioritizes precision, preferring to miss some relevant terms rather than include noise. The conservative similarity threshold, minimum frequency requirements, and manual validation ensure our measures capture meaningful variation rather than random linguistic patterns. This methodological choice means our dictionaries are smaller than exhaustive keyword lists but more reliable for distinguishing genuine strategic integration from sophisticated compliance narratives.

Examples - To illustrate how these measures capture distinct reporting orientations, consider these examples from our corpus:

Strategic orientation: A multinational beverage company states: 'We continue to explore ways to reduce emissions through our commercial strategy and invest in the decarbonization of our operations.' Similarly, a European bank frames its approach as: 'Our climate strategy defines our approach to aligning with the global goal of limiting warming to 1.5°C and supporting the transition to a net-zero economy by 2050.' A technology firm emphasizes innovation: 'GenMatch™ exemplifies our vision for sustainable innovation by combining advanced technology with environmental stewardship'.

Compliance orientation:

An insurance firm states: 'In accordance with the requirements of the ESRS, we define our value chain as the full range of activities, resources, and relationships related to our business model'. A media company notes: 'Use of phase-in provisions in accordance with Appendix C of ESRS 1'. A telecommunications provider specifies: 'The double materiality assessment process was carried out... in accordance with ESRS 1 paragraph 103'.

Forward-looking information: A sportswear manufacturer commits: 'CO2e emissions intensity target per product for 2025 (-15% reduction compared to 2017) in line with our SBTi-approved targets for 2030 and 2050'. A real estate firm targets: '100% of assets in water stressed areas to implement water reuse solutions by 2025, and 100% of our portfolio by 2030'. A food company pledges: 'Reduce salt content by 10% by 2030, relative to 2019'.

3.5 Validation and Scalability

The credibility of our automated measures rests on validation against human judgment. We selected 60 random reports from our 592-report corpus, ensuring representation across countries, industries, and report lengths. Three independent coders evaluated each report: the

first author (a finance professor focusing on AI applications in sustainability), a finance professor specializing in capital markets, and an MBA student specializing in sustainability reporting. While the first author participated as one coder, all three were blind to automated scores and each other's ratings, ensuring independence. This diverse expertise, which spans AI/sustainability, capital markets, and ESRS compliance, provides triangulation for validation.

Our coding protocol operationalized constructs clearly. Strategic orientation: "the extent to which sustainability narrative emphasizes business value and strategic integration versus regulatory compliance" (1=pure compliance to 5=highly strategic). Forward-looking information substance: "the extent of concrete, specific future commitments versus vague/minimal forward-looking discussion" (1=minimal to 5=extensive). After a one-hour calibration session using practice reports, coders rated independently.

Inter-rater reliability showed substantial agreement (Fleiss' $\kappa = 0.68$ for strategic orientation, 0.64 for forward-looking information), indicating strong consensus among coders despite the subjective nature of the constructs. Detailed validation procedures and results are presented in Appendix A.

Correlations between consensus human ratings and automated measures are 0.41 for Strategic_Ratio ($p<0.05$) and 0.39 for FLI_Score ($p<0.05$). These moderate correlations require careful interpretation. They're statistically significant, confirming our measures capture meaningful variance. However, they are not so high as to suggest perfect alignment, which validates that our measures add value beyond simple human coding. The partial alignment reflects different strengths: humans form holistic impressions influenced by narrative flow and emphasis; our measures systematically count every term equally. Humans bring inherent subjectivity despite calibration; our measures provide perfect consistency. Our measures

capture patterns across entire documents that readers might miss focusing on prominent sections.

The validation thus supports complementary human-machine measurement. Human judgment provides the validity anchor confirming construct relevance. Machine measurement provides scalability and consistency impossible for human coding.

Scalability and Efficiency

The framework's transformative advantage lies in its scalability. The initial, one-time setup costs are significant, requiring approximately 60 hours of expert time for dictionary development and 2 hours for the Word2Vec model training. However, once the framework is built, the marginal processing cost is trivial.

Our final pipeline processed the entire 592-report corpus in less than an hour. For context, a conservative estimate for expert human coding is 30 hours per report. Analyzing our 592-report corpus manually would therefore require thousands of hours of sustained, expert-level work. Our framework transforms this impossible task into a tractable analysis.

This efficiency scales linearly. After the initial setup, processing the emerging CSRD universe of 10,000 firms would require approximately 6.5 hours of computation, with no additional human intervention. This enables applications previously considered impossible: real-time disclosure monitoring across entire markets, systematic cross-country and industry comparison, and the automated detection of emerging disclosure trends.

We also maintain complete interpretability despite scale. Unlike neural network 'black boxes,' every score traces to specific words, enabling investors to understand quality assessments and regulators to justify enforcement actions. This transparency extends to methodology refinement. When our S1 dictionary initially failed, we could identify the problem and iterate, something impossible with opaque AI approaches.

The complete dictionaries, code and data available upon request from the corresponding author.

4. Empirical Demonstration: Uncovering Patterns in CSRD Disclosures

4.1 Demonstration Design and Variables

To demonstrate our framework's capability to detect meaningful patterns invisible to manual analysis, we apply it to 446 firms with complete textual and financial data. This demonstration serves to validate that our measures capture meaningful variation, not to test governance theory.

The governance patterns we uncover simply illustrate the types of insights our method enables.

We estimate OLS regressions examining whether firm characteristics influence disclosure quality:

$$\text{DisclosureQuality}_i = \beta_0 + \beta_1 \text{BoardIndependence}_i + \beta_2 \text{FirmAge}_i + \beta_3 \text{FirmSize}_i + \\ \text{Controls}_i + \text{IndustryFE} + \text{CountryFE} + \varepsilon_i$$

Dependent Variables (from our framework):

- **Strategic_Ratio:** Strategic word count divided by strategic plus compliance word count (0=pure compliance, 1=pure strategy)
- **FLI_Score:** $\ln(1 + \text{Forward-Looking Word Count}) / \ln(1 + \text{Strategic} + \text{Compliance} + \text{Forward-Looking Word Count})$

Independent Variables:

- **BoardIndependence:** Percentage of independent directors. These variable tests whether monitoring for quality conflicts with risk management
- **FirmAge:** Log years since founding, which is a proxy for organizational maturity and disclosure experience
- **FirmSize:** Log total assets, which is a traditional proxy for resources

Controls: ROA (profitability), Leverage (debt/assets), plus industry and country-group fixed effects to account for institutional variation.

4.2 Descriptive Evidence and Main Results

Table 2 presents descriptive statistics revealing substantial heterogeneity our framework captures. The mean Strategic_Ratio of 0.39 (SD=0.14) indicates compliance-oriented language dominates, but wide variation exists (range: 0.00-1.00). Initial correlations hint at our key finding: BoardIndependence correlates positively with Strategic_Ratio ($\rho=0.13$) but negatively with FLI_Score ($\rho=-0.05$).

[Insert Table 2 about here]

Table 3 presents regression results revealing a governance trade-off invisible to traditional analysis:

[Insert Table 3 about here]

The R-squared values indicate that while our models capture meaningful variation in disclosure, a considerable portion remains shaped by firm-specific heterogeneity. This underscores both the explanatory contribution of our measures and the inherent complexity of modeling disclosure behavior across diverse firms. Our objective is not to explain total variance but to detect meaningful relationships. The 0.061 coefficient for board independence on Strategic_Ratio represents approximately 6% more strategic language. This effect is economically meaningful because disclosure choices carry legal implications and influence stakeholder perceptions. The statistical significance despite low overall explanatory power actually strengthens our contribution, demonstrating that our method can detect subtle but important patterns that would be invisible to manual analysis.

The Governance Trade-Off: Board independence shows opposing effects on our two measures. Independent directors push for strategic framing ($\beta=0.061$, $p=0.049$) by monitoring management to move beyond boilerplate compliance. Simultaneously, they constrain forward-looking commitments ($\beta=-0.010$, $p=0.001$) by managing legal liability from concrete future promises. This nuanced pattern, where the same governance mechanism pulls disclosure in opposite directions, exemplifies what only systematic textual analysis can reveal.

Challenging Conventional Wisdom: Neither firm size nor age significantly predicts disclosure quality consistently. This challenges resource-based theories dominating disclosure literature, suggesting governance dynamics matter more than traditional resource advantages in the CSRD's highly regulated environment.

4.3 Validation Through Robustness and Extensions

4.3.1 Robustness Tests

Our findings withstand multiple specification changes:

- **Alternative measures:** Raw word counts with length controls yield identical patterns
- **Outlier sensitivity:** Winsorizing at 1%/99% doesn't alter results
- **Extraction method:** Governance trade-off appears in both automatically (249 firms) and manually extracted (197 firms) subsamples
- **Non-linearity:** Quadratic terms leave findings unchanged
- **Sample variations:** Excluding financials or specific countries maintains patterns

4.3.2 Institutional Heterogeneity

The framework automatically detects institutional patterns requiring enormous manual effort to identify. Using Varieties of Capitalism classifications (Hall and Soskice, 2001) reveals:

- **Nordic effect:** Firms in Denmark, Sweden, Norway, Finland show significantly higher Strategic_Ratios (coefficient: 0.082, p<0.01), reflecting stakeholder-oriented governance traditions
- **Sector patterns:** Financial firms exhibit lower strategic orientation (coefficient: -0.047, p<0.05), suggesting regulatory scrutiny drives compliance focus
- **Country clusters:** Continental European firms balance strategic and compliance language more than Anglo or Mediterranean clusters

4.3.3 What Manual Analysis Would Miss

This demonstration highlights our framework's transformative value:

Scale: Detecting the governance trade-off manually would require:

- Thousands of hours of reading and synthesizing
- Multiple coders maintaining perfect consistency on subjective constructs
- Statistical analysis across hundreds of observations

Our framework accomplished this in less than an hour with perfect consistency.

Invisible Patterns: The opposing effects, with more strategic language yet less forward-looking content from board independence, would likely never emerge from qualitative assessment. Human readers are likely to form holistic impressions, missing subtle countervailing patterns our systematic measurement reveals.

Institutional Comparisons: Identifying Nordic firms' systematically different disclosure approach requires comparing hundreds of reports across countries which is feasible only through automation.

Statistical Precision: While human coders achieve moderate agreement ($\kappa=0.68$), our measures provide identical scores for identical text, enabling detection of small but significant effects ($R^2=0.037$) that would disappear in coding noise.

This demonstration proves our framework works, detecting meaningful patterns in real data that validate both our measures and the method's unique value. The governance trade-off, though theoretically plausible but empirically subtle, could only emerge through systematic analysis at scale. As CSRD implementation covers nearly 10,000 firms, such automated yet interpretable analysis becomes not just useful but essential for understanding this new disclosure landscape.

5. Discussion and Implications

5.1 Methodological Contributions

This study introduces a scalable, validated framework for analyzing sustainability disclosures that fundamentally transforms what is possible in the CSRD era. Our "expert-seeded, AI-validated" approach demonstrates that meaningful disclosure quality assessment can be automated without sacrificing the interpretability essential for regulatory and investment decisions.

The methodological contribution extends beyond simple automation. By combining regulatory expertise with machine learning, we create measures more precise than either approach alone. Our seed dictionaries ensure grounding in ESRS requirements, while Word2Vec validation ensures empirical relevance. The moderate but significant correlations with human judgment ($r=0.39-0.41$) confirm our measures capture meaningful constructs while offering perfect consistency impossible for human coders. Most importantly, we maintain complete transparency, as every score traces to specific words and enables users to understand and verify our assessments.

The efficiency gains are substantial. Processing 592 reports in under an hour, compared to the thousands of hours required manually, represents a marked improvement. This is not only faster, but also enables entirely new analytical possibilities. Real-time monitoring across entire markets, systematic cross-jurisdictional comparisons, and early detection of disclosure trends become routine rather than impossible.

Our empirical demonstration, while secondary to the methodological contribution, reveals patterns challenging conventional disclosure theory. The governance trade-off shows that board independence increases strategic framing but decreases forward-looking commitments. This pattern could only be detected through systematic analysis at scale. This finding suggests that in highly regulated environments, governance dynamics may matter more than traditional resource advantages, warranting renewed theoretical attention to how boards navigate competing pressures in mandatory disclosure regimes.

5.2 Implications for Different Stakeholders

For Regulators and Policymakers

Our framework offers regulatory bodies improved monitoring capacity. Rather than sampling reports for manual review, regulators can continuously analyze the entire CSRD universe, identifying patterns requiring intervention. Reports with extremely low strategic ratios might indicate boilerplate compliance warranting closer scrutiny. Sudden shifts in forward-looking information across sectors could signal emerging interpretation issues requiring guidance.

The framework enables evidence-based policy refinement. By analyzing disclosure patterns across countries and industries, regulators can identify where standards create unintended consequences or interpretation difficulties. For instance, if specific ESRS requirements consistently yield compliance-focused language, standards might need clarification to encourage more meaningful disclosure. The ability to track linguistic evolution over time

allows regulators to assess whether the CSRD achieves its strategic integration goals or merely creates sophisticated compliance exercises.

Consider the European Securities and Markets Authority (ESMA) implementing our framework for systematic enforcement. Monthly processing of all CSRD reports could generate alerts for outliers, identify best practices for dissemination, and track implementation progress across member states. The interpretability of our measures means enforcement actions remain defensible, as regulators can point to specific linguistic patterns justifying intervention.

For Investors and Financial Analysts

The framework transforms how investors assess sustainability disclosure quality. Rather than relying on ESG ratings agencies' opaque methodologies, investors can directly evaluate the substance behind corporate narratives. Our strategic ratio distinguishes firms genuinely integrating sustainability from those treating it as compliance burden. The forward-looking measure identifies companies making concrete commitments versus those offering vague aspirations.

Investment strategies could incorporate our measures as screening tools. A sustainability-focused fund might require minimum strategic ratios, ensuring portfolio companies view sustainability as value creation opportunity. Risk-averse investors might scrutinize firms with high forward-looking scores, recognizing these represent future performance commitments. The measures could also identify greenwashing, as firms with high strategic language but minimal forward-looking substance might be engaging in impression management.

The framework enables dynamic monitoring of portfolio companies. Sudden drops in strategic orientation or forward-looking information might signal changing management attitudes toward sustainability, warranting engagement. Comparing firms' scores to industry averages identifies leaders and laggards, informing voting and engagement strategies.

For Reporting Companies

Companies gain powerful benchmarking capabilities. Our measures allow firms to assess their disclosure quality against peers objectively, identifying areas for improvement. A firm discovering its strategic ratio falls below industry average might recognize its reports read as compliance documents despite strategic intent. Low forward-looking scores might indicate excessive legal caution preventing meaningful commitment communication.

The framework guides report development. Communications teams can test different narrative approaches, measuring how language choices affect strategic versus compliance perception. Legal teams can calibrate forward-looking disclosure, balancing transparency with liability management. The ability to analyze competitors' reports systematically reveals emerging best practices and industry disclosure norms.

For Academic Researchers

Beyond our specific findings, the framework provides infrastructure for future CSRD research. Researchers can apply our dictionaries to investigate numerous questions: How do disclosure patterns evolve as firms gain CSRD experience? Do strategic orientations predict future sustainability performance? How do stakeholder pressures influence narrative choices?

The methodology transfers to other regulatory contexts. Researchers studying SEC climate disclosures, TCFD reports, or emerging regulations can adapt our approach by developing regime-specific dictionaries while maintaining the expert-seeded, AI-validated framework. The combination of scalability and interpretability makes previously infeasible research questions tractable.

5.3 Future Development and Extensions

The framework's modular design facilitates multiple enhancement pathways. The framework's modular architecture facilitates adaptation to other mandatory disclosure regimes. Researchers studying SEC climate rules, ISSB standards, or emerging national regulations can maintain our core methodology while developing regime-specific seed dictionaries. This transferability demonstrates the framework's value as generalizable infrastructure rather than a CSRD-specific tool.

Linguistic sophistication could improve through negation detection, allowing distinction between "we have established targets" and "we have not established targets." Prominence weighting could recognize that extensive discussion carries more weight than passing mentions. Sentiment analysis layers could identify whether strategic language conveys genuine enthusiasm or reluctant compliance.

Multi-language capabilities represent an extension for analyzing the full CSRD universe. This requires more than translation, since different languages encode sustainability concepts differently, which in turn necessitates language-specific dictionaries and validation. Collaboration with native-speaking experts across EU member states could develop parallel frameworks maintaining cross-linguistic comparability.

Dynamic dictionary evolution could address the temporal challenge. As sustainability language evolves, dictionaries require updating while maintaining historical comparability. Machine learning could identify emerging terms gaining prominence, flagging them for expert review and potential inclusion. Version-controlled dictionaries could track linguistic evolution explicitly, enabling research on how disclosure language changes over time.

Integration with quantitative metrics could create holistic assessment tools. Combining our textual measures with reported emissions, diversity statistics, and financial performance could

identify alignment or divergence between narrative and numbers. Firms claiming strategic integration but showing no sustainability-linked investments might be engaging in decoupling.

5.4 Broader Implications for AI in Accounting Research

Our study demonstrates AI's potential to transform accounting research while maintaining scholarly rigor. The expert-seeded approach shows how domain knowledge can guide machine learning, avoiding purely algorithmic approaches that might miss regulatory nuances. The validation against human judgment establishes necessary bridges between computational and traditional methods.

The transparency imperative in our design challenges the trend toward black-box AI applications. While neural networks might achieve marginally better prediction, the inability to explain their decisions limits their utility for stakeholders needing defensible assessments. Our framework shows that interpretable AI can achieve meaningful results while maintaining the transparency essential for practical application.

5.5 Methodological Considerations and Limitations

Our framework involves deliberate trade-offs prioritizing scalability and transparency while acknowledging inherent limitations. Dictionary evolution presents the most fundamental challenge, as sustainability language can evolve rapidly. Terms that signal strategic innovation today may become compliance boilerplate tomorrow. "Net-zero" transformed from cutting-edge ambition to standard expectation within a few years. Our dictionaries, trained on 2024 inaugural reports, capture initial linguistic patterns but will require periodic updating. Annual retraining could maintain validity while preserving comparability through core stable terms.

Language and sample constraints create important limitations. Processing only English-language reports likely biases our sample toward Northern and Western European firms with established English reporting traditions and potentially different governance norms than

Southern or Eastern European companies. This may explain the strong 'Nordic effect' we observe and limits generalizability to the full CSRD universe. Future work should develop language-specific models to capture disclosure patterns across all EU member states.

Contextual blindness remains a limitation of our dictionary-based scoring. Negation may lead to false positives (e.g., 'We have not established science-based targets'), and both detailed and passing mentions receive equal weight. Hedging terms ('could,' 'might') are also counted alongside firm commitments ('will,' 'shall').

Measurement interpretation requires careful consideration. Strategic language does not necessarily indicate disclosure quality. It could reflect sophisticated impression management or greenwashing rather than genuine integration. Similarly, limited forward-looking information might indicate either excessive risk aversion or appropriately cautious, evidence-based reporting. Our measures identify linguistic patterns requiring further investigation to establish whether they represent substantive commitment or regulatory gamesmanship.

Sample limitations affect generalizability. Our framework was developed on early adopters, who are typically large, listed companies with sophisticated sustainability management. Linguistic patterns may not extend to smaller firms approaching CSRD as pure compliance. The 2024 reporting year represents a unique transitional moment; patterns may shift as practices standardize. Validation on 60 reports, while significant, leaves substantial unexplained variance in our moderate correlations (0.39-0.41).

Boundary conditions define appropriate application. For research requiring narrative structure analysis or causal reasoning, transformer models may be necessary despite computational costs. For simple compliance checking, rule-based approaches might suffice. Our measures capture linguistic choices, not underlying reality.

Despite limitations, the framework provides robust infrastructure for systematic CSRD analysis. Future developments could integrate negation detection, prominence weighting, or multi-modal analysis. The modular design facilitates such extensions while maintaining our core insight: expert knowledge combined with machine learning creates scalable yet interpretable measures essential for analyzing the expanding universe of mandatory sustainability disclosure.

5.6 Conclusion

The framework's value magnifies over time. As CSRD reports accumulate annually, longitudinal analysis becomes essential. It enables comparison of how firms' disclosures evolve, whether commitments materialize, and how language patterns shift, but performing this task manually becomes extremely difficult. Our automated approach makes tracking disclosure evolution across years as feasible as analyzing a single year. While demonstrated on CSRD reports, our framework provides a template for analyzing mandatory sustainability disclosures globally, requiring only dictionary adaptation for different regulatory vocabularies.

The governance trade-off we uncover illustrates the framework's power to reveal patterns invisible to traditional analysis. But this empirical finding is merely a demonstration, while the true contribution lies in the methodology. We provide tools for regulators to monitor compliance systematically, for investors to assess disclosure quality objectively, and for researchers to investigate previously intractable questions.

The framework is not a finished product but a foundation for community development. As researchers apply, refine, and extend our approach, we envision an ecosystem of interoperable tools enabling comprehensive understanding of the evolving sustainability disclosure landscape. The complete code and dictionaries, available upon request from the corresponding author, invite such collaboration.

The CSRD represents a watershed moment in corporate transparency, demanding new analytical approaches matching its scale and complexity. Our framework offers one such approach, demonstrating that the marriage of human expertise and machine capability can create tools that are both powerful and practical. As sustainability disclosure becomes central to capital markets, such tools transition from academic curiosity to essential infrastructure.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Google AI Studio in order to enhancing final article writing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgements

The authors thank the anonymous reviewers and editors for their valuable comments.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. arXiv preprint arXiv:1908.10063. <https://arxiv.org/abs/1908.10063>
- Berg, F., Koelbel, J. F., & Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1354. <https://doi.org/10.1093/rof/rfac033>
- Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. *Review of Accounting Studies*, 26(3), 1176–1248. <https://doi.org/10.1007/s11142-021-09609-5>
- Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64(2–3), 221–245. <https://doi.org/10.1016/j.jacceco.2017.07.002>
- Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. *The Journal of Law and Economics*, 26(2), 301–325. <https://doi.org/10.1086/467037>
- Freiberg, D., Park, D. G., Serafeim, G., & Zochowski, R. (2021). Corporate environmental impact: Measurement, data and information. *Harvard Business School Working Paper*, 20-098.
- Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine learning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, 2(3), 153–170. <https://doi.org/10.1016/j.jfds.2017.02.001>
- Hall, P. A., & Soskice, D. (Eds.). (2001). *Varieties of capitalism: The institutional foundations of comparative advantage* (1st ed.). Oxford University Press Oxford. <https://doi.org/10.1093/0199247757.001.0001>
- Hanley, K. W., & Hoberg, G. (2010). The information content of IPO prospectuses. *The Review of Financial Studies*, 23(7), 2821–2864. <https://doi.org/10.1093/rfs/hhq024>
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2–3), 146–162. <https://doi.org/10.1080/00437956.1954.11659520>
- Hope, O. K., Hu, D., & Lu, H. (2016). The benefits of specific risk-factor disclosures. *Review of Accounting Studies*, 21(4), 1005–1045. <https://doi.org/10.1007/s11142-016-9371-1>
- Hummel, K., & Schlick, C. (2016). The relationship between sustainability performance and sustainability disclosure – Reconciling voluntary disclosure theory and legitimacy theory. *Journal of Accounting and Public Policy*, 35(5), 455–476. <https://doi.org/10.1016/j.jaccpubpol.2016.06.001>
- Jackson, G., Bartosch, J., Avetisyan, E., Kinderman, D., & Knudsen, J. S. (2020). Mandatory non-financial disclosure and its influence on CSR: An international comparison. *Journal of Business Ethics*, 162(2), 323–342. <https://doi.org/10.1007/s10551-019-04200-0>
- Krueger, P., Sautner, Z., Tang, D. Y., & Zhong, R. (2024). The effects of mandatory ESG disclosure around the world. *European Corporate Governance Institute – Finance Working Paper No. 754/2021*. <https://doi.org/10.2139/ssrn.3832745>

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>

Lang, M., & Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, 60(2–3), 110–135. <https://doi.org/10.1016/j.jacceco.2015.09.002>

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2–3), 221–247. <https://doi.org/10.1016/j.jacceco.2008.02.003>

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7), 3265–3315. <https://doi.org/10.1093/rfs/hhaa079>

Lin, Y., Shen, R., Wang, J., & Yu, Y. J. (2024). Global evolution of environmental and social disclosure in annual reports. *Journal of Accounting Research*, 62(5), 1941–1988. <https://doi.org/10.1111/1475-679X.12575>

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187–1230. <https://doi.org/10.1111/1475-679X.12123>

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv* [arXiv:1301.3781](https://arxiv.org/abs/1301.3781). <https://doi.org/10.48550/arXiv.1301.3781>

Vishnu Nampoothiri, M., Entrop, O., Annamalai, T. R. (2024). Effect of mandatory sustainability performance disclosures on firm value: Evidence from listed European firms. *Corporate Social Responsibility and Environmental Management*, 31(6), 5220–5235. <https://doi.org/10.1002/csr.2860>

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). ClimateBERT: A pretrained language model for climate-related text. *arXiv:2110.12010*. <https://doi.org/10.48550/arXiv.2110.12010>

Abbreviations

AI	Artificial Intelligence
CSRD	Corporate Sustainability Reporting Directive
CapEx	Capital Expenditure
DNSH	Do No Significant Harm
E1–E5	Environmental Standards 1–5 (<i>in ESRS</i>)
EFRAG	European Financial Reporting Advisory Group
ESG	Environmental, Social, and Governance
ESRS	European Sustainability Reporting Standards
EU	European Union
FLI	Forward-Looking Information
FLS	Forward-Looking Statements
GHG	Greenhouse Gas
GOV	Governance (<i>ESRS topical standard category</i>)
G1	Governance Standard 1 (<i>in ESRS</i>)
IRO	Impacts, Risks, and Opportunities
IFRS	International Financial Reporting Standards
κ	Kappa statistic (measure of inter-rater agreement)
KPI	Key Performance Indicator
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
NER	Named Entity Recognition
NFRD	Non-Financial Reporting Directive
MT	Metrics and Targets
RBV	Resource-Based View
SBM	Strategy and Business Model (<i>section in ESRS</i>)
S1–S4	Social Standards 1–4 (<i>in ESRS</i>)
TCFD	Task Force on Climate-related Financial Disclosures
UK	United Kingdom
Word2Vec	Word to Vector (<i>word embedding model</i>)

Appendix A: Human Validation of Automated Disclosure Measures

A.1 Validation Design

To confirm that our automated measures capture meaningful variation in disclosure quality, we compared automated scores against independent human assessments, following established practice in textual analysis validation (Lin et al., 2024).

A.2 Sample and Coders

We randomly selected 60 reports for validation, covering 21 countries, all major industries, and report lengths from 45 to 287 pages. The 60-report sample (10.1% of our 592-report corpus) exceeds typical validation samples in textual analysis studies. Lin et al. (2024) validated on 420 sentences from their 210,000 reports (<0.1%); our 60 complete reports provide more comprehensive validation. Three independent coders evaluated each report:

- The first author, a finance professor focusing on AI applications in sustainability analysis
- A finance professor with expertise in capital markets and disclosure
- An MBA student specializing in sustainability reporting and ESRS compliance

After a one-hour training session using five practice reports, coders independently assessed:

- **Strategic Orientation:** Business value emphasis vs. compliance focus (1–5 scale)
- **Forward-Looking Information:** Concrete commitments vs. vague aspirations (1–5 scale)

A.3 Results

Inter-rater reliability among the three human coders showed substantial agreement, with Fleiss' κ values of 0.68 for strategic orientation and 0.64 for forward-looking information.

These values, which exceeded the 0.60 threshold for substantial agreement (Landis & Koch, 1977), were consistent with average pairwise Cohen's κ values of 0.68 and 0.64 respectively, confirming reliable human assessment of both constructs.

Machine-human agreement analysis revealed statistically significant correlations between our automated measures and consensus human ratings. Strategic_Ratio correlated at 0.41 with human assessments ($p = 0.003$), while FLI_Score showed a correlation of 0.39 ($p = 0.005$). These moderate correlations align with validation benchmarks in the textual analysis literature, where correlations typically range from 0.25 to 0.45. (See Table A.1 for detailed inter-rater reliability statistics and Table A.2 for machine-human agreement correlations in the accompanying Tables document.)

These significant results confirm that our measures capture meaningful variance while offering perfect reproducibility. This feature is important for scaling to 592 reports that would otherwise require thousands of hours of manual expert coding.

Tables

Table 1: Sample Selection

Description	Number of Reports
Annual Reports with initial CSRD reports from SRN database (2024)	628
Less: Not machine-readable or not in English	-36
Textual analysis sample	592
Less: Missing financial/governance data	-146
Regression analysis sample	446

Table 2: Descriptive Statistics - N=446

Variable	Mean	SD	Min	Max
Strategic_Ratio	0.39	0.14	0	1
FLI_Score	0.94	0.05	0.47	0.97
BoardIndependence	0.63	0.23	0	1
FirmSize	9.51	2.08	4.45	17.59
FirmAge	3.42	0.88	0	5.63

Table 3: Regression Results

Variables	(1) Strategic_Ratio	(2) FLI_Score
Board Independence	0.0376 (0.061)	-0.0149 (0.010)
Firm Size	-0.0037 (0.003)	0.0016*** (0.000)
Firm Age	-0.0015 (0.008)	-0.0041** (0.002)
ROA	0.0011 (0.001)	-0.0003** (0.000)
Leverage	0.0001 (0.001)	-0.0002 (0.000)
Industry FE	Yes	Yes
Country FE	Yes	Yes
Observations	274	274
R-squared	0.034	0.043

Notes:

1. Standard errors (in parentheses) are clustered.
2. ***, **, * denote significance at the 1%, 5%, and 10% levels, respectively.
3. Industry and country fixed effects included but not reported for brevity.

A.3 Results

Table A.1: Inter-Rater Reliability

Measure	Fleiss' κ	Avg. Pairwise Cohen's κ
Strategic Orientation	0.68	0.68
Forward-Looking Information	0.64	0.64

Note: κ values of 0.61–0.80 indicate substantial agreement (Landis & Koch, 1977).

Table A.2: Machine–Human Agreement

Automated Measure	Correlation (Human Mean)	p-value
Strategic_Ratio	0.41	0.003
FLI_Score	0.39	0.005