*Article*

# A Transformer-Based Framework for Context-Aware ESG Corpus Construction

**Daniel Broby**

Asian Institute of Management, Makati City 1229, Philippines
Email: dbroby@aim.edu

Received: 28 March 2026; Revised: 11 May 2026; Accepted: 13 May 2026; Published: 5 June 2026

**Abstract:** This paper develops a transformer-based method for constructing a context-aware Environmental, Social, and Governance (ESG) finance corpus. The proposed framework uses attention-based relevance scores to identify ESG terms and sentences according to their surrounding textual context. The method is first illustrated through a probabilistic ESG term matrix and then evaluated using a validation sample of SEC EDGAR 10-K filings from 73 S&P 500 firms across 11 GICS sectors. The resulting balanced corpus contains 1,456 ESG sentences, with environmental, social, and governance content representing 34.8%, 35.1%, and 30.1% of the sample, respectively. Validation evidence indicates that the attention-based procedure performs well relative to alternative corpus construction methods. Ten-fold cross-validation produces an overall F1-score of 87.7%, while corpus-derived ESG intensity scores correlate positively with alternative ESG ratings. The attention-based method also outperforms keyword matching, TF-IDF filtering, and a supervised BERT benchmark on both classification performance and external rating alignment.

**Keywords:** corpus; natural language processing; ESG; transformer models

---

## 1. Introduction

This paper develops a context-aware method for Environmental, Social, and Governance (ESG) corpus construction using transformer-based attention scores. The contribution is methodological, rather than the release of a definitive production corpus. To assess whether this is empirically credible, the paper supplements the conceptual framework with validation evidence. This is based on cross-validation, threshold sensitivity analysis, and external ESG-rating correlations. It ensures robustness with benchmark comparisons against keyword matching, TF-IDF filtering, and a supervised BERT classifier.

In the setting of natural language processing, this article presents a conceptual method for an ESG corpus construction that empirically extends a framework introduced in Broby (2026). The resultant attention-based probabilistic ESG corpus is first presented through a bounded illustrative application. This delivers a structured set of textual words used for linguistic analysis that is then tested on a validation sample of SEC EDGAR 10-K filings from 73 S&P 500 firms across 11 GICS sectors. It can be used in machine learning models and evaluating document analysis algorithms. The growing emphasis on ESG factors in finance has driven researchers to seek more precise and context aware methods for building such a corpus (Saini et al., 2023). This paper fills that gap by presenting a conceptual framework to build a corpus using word context as produced by generative models.

Recent studies highlight the growing use of textual analysis in ESG research, demonstrating the practical value of a reliable corpus. Hassan et al. (2019, 2023a) show how text-based measures can capture firm-level risk exposures with implications for financial outcomes. More ESG-specific work includes Ferjančič et al. (2024a), linking sustainability reporting to ESG scores, and Heichl & Hirsch (2023), who analyse disclosure patterns in EU firms. Others, such as Likitapiwat et al. (2024)

and Zhou et al. (2024), examine how ESG language reflects culture and market perception. These studies show that methods like the one proposed serve as key inputs for empirical ESG analysis.

Traditional approaches to corpus creation often suffer from context. For example, in keyword matching or manual annotation. They are also prone to biases that result from the use of static lexicons. They therefore fail to capture the nuanced and contextual nature of ESG-related language. These limitations motivate this study. The author believes using transformers offers a scalable, adaptive, and data-driven fintech solution to their creation (Broby (2024) and Broby & Yang (2025)).

The empirical design proceeds in two stages. The first stage uses a smaller illustrative corpus to demonstrate how attention-based relevance scores can be used to identify ESG terms, assign probabilistic weights, and construct sentence-level ESG classifications. The second stage applies the same procedure to a larger SEC EDGAR 10-K validation corpus, allowing the method to be assessed through cross-validation, threshold sensitivity analysis, benchmark comparisons, and external ESG-rating correlations. This two-sample structure separates the exposition of the method from its empirical evaluation.

In sum, the proposed method utilises a mechanism called “attention”. This allows the importance of different words in a sentence to be taken into account. Its uses “embeddings” that are sensitive to the context in which words appear, making it suitable for ESG terminology.

### *1.1. Corpus Generation*

Currently, Bag of Words is the most common corpus generation method. It represents a multiset of words. (see Zhang et al., 2010). A key shortcoming, however, is its inability to capture the semantic meaning or word order within text. For example, if a company’s report contains phrases like “reducing carbon emissions” and “commitment to net zero”. Bag of Words would treat these words as isolated terms. This loses the relevance of the semantic relationship between, for example, ”commitment” and ”net zero”, or ”reducing” and ”carbon emissions.”

Transformer models, including BERT, DGrok 3, GPT-4o, Claude 3.5, Deepseek V3, and Gemini 2 Pro have revolutionised Natural Language Processing (NLP) (Young et al., 2018). They can account for context and capture semantic nuances. Sokolov et al. (2021) illustrated this using BERT (Bidirectional Encoder Representations from Transformers) to automate the construction of an ESG portfolio.

This approach is different because transformer-based embeddings consider the entire context in which a word appears (Rong 2014). It is not a static word representation, such as delivered by Word2Vec or Global Vectors for Word Representation (GloVe). As a result, its use dynamically adjusts meaning based on surrounding text (Ghojogh & Ghodsi, 2020). Such context sensitivity is crucial for ESG research. For example, terms like ”sustainability”, ”equity”, or ”governance” may carry different meanings depending on their usage in financial reports, regulatory filings, or corporate sustainability disclosures.

Building on the limitations of Bag of Words, more advanced word embedding techniques have been developed. For example, Word2Vec utilizes neural network architectures (Mikolov et al., 2013). This approach employs two primary models, namely ” Continuous Bag of Words” (CBOW) and Skip-Gram. The CBOW model can predict a target word based on its surrounding context (Liu, 2020). Skip-Gram, meanwhile, predicts surrounding words given a target word (Guthrie et al., 2006). These methods show there is a demand for the differing approaches.

While Word2Vec focuses on learning word representations through predictive modeling, GloVe takes a different approach. It leverages statistical co-occurrence. It is based on matrix factorization techniques. In generating a concurrence matrix, it allows GloVe to make embeddings that encode both global and local word relationships. The attention mechanism of transformer models, however, generates a focus on the “contextually” relevant terms. This represents an improvement over manually constructed ESG corpus and, in the authors’ opinion, even surpasses GloVe.

While prior work has applied transformers to ESG classification tasks (Sokolov et al., 2021; Schimanski et al., 2024), our study is the first to systematically leverage attention mechanisms for probabilistic corpus construction. Unlike domain pretraining approaches that require massive unlabeled data, our method works with moderate-sized datasets by exploiting attention weights to identify contextually relevant terms. This addresses the scalability-quality tradeoff identified in recent ESG NLP research.

This paper therefore makes the case that, in building a corpus using a transformer, researchers can create a robust and contextually relevant ESG corpus. This can be tailored to specific tasks. The resulting corpus can therefore reflect the nature of socially responsible language.

## 2. Literature

The application of textual analysis to ESG research has emerged as a methodological challenge in finance and sustainability studies. Traditional approaches to ESG measurement have relied heavily on rating agencies and structured data. These methods often fail to capture the nuanced, context-dependent nature of corporate sustainability communications (Schimanski et al., 2024).

The limitations identified across these strands of literature are cumulative rather than separate. Studies of ESG textual analysis show that corporate sustainability language is heterogeneous and context dependent. Research on ESG datasets and benchmarks has improved standardization, but also highlights variation in annotation practice, source composition, and representativeness. Transformer-based studies, in turn, demonstrate that contextual language models outperform lexicon-based approaches in many classification settings, yet they largely take the corpus itself as given.

What remains insufficiently developed is a framework that links these literatures by treating corpus construction as an inferential stage rather than a purely mechanical preprocessing step. This gap provides the motivation for the present study. If ESG language is semantically contingent, and if benchmark quality depends on the quality of the underlying textual input, then attention mechanisms are relevant - not only for downstream classification but also for the prior task of corpus formation. The proposed method addresses that gap by using contextual relevance derived from transformer attention to support probabilistic ESG corpus construction.

### *2.1. ESG Textual Analysis*

The shift toward textual analysis reflects a broader recognition that ESG-relevant information is increasingly embedded in unstructured corporate disclosures, news articles, and regulatory filings.

Recent scholarship demonstrates the practical value of text-based ESG measures. Hassan et al. (2019, 2023b) pioneered the use of textual analysis to capture firm-level risk exposures with significant implications for financial outcomes. Building on this foundation, Ferjančič et al. (2024b) established empirical linkages between sustainability reporting quality and ESG scores, while Heichl & Hirsch (2023) employed textual analysis to examine disclosure patterns across EU firms. These studies collectively illustrate how computational text analysis can reveal patterns invisible to traditional rating methodologies.

Domain-specific applications have further refined ESG textual analysis. Nugent et al. (2020, 2021) developed specialized language models for detecting ESG topics, demonstrating that domain-specific pre-training significantly improves classification accuracy compared to general-purpose models. Their work employed data augmentation approaches to address the scarcity of labeled ESG data, achieving notable improvements in topic detection. Similarly, Perazzoli et al. (2022) advanced a systemic perspective on ESG evaluation through natural language processing, arguing that ESG factors must be analyzed holistically rather than in isolation.

The integration of sentiment analysis into ESG research has opened new avenues for understanding market perceptions and corporate intentions. Zhou et al. (2024) examined how media attention and negative reports influence banks' ESG disclosure through textual information analysis. Likitapiwat et al. (2024) employed textual analysis to detect corporate culture through ESG language. These studies highlight the dual role of textual analysis in both measuring ESG performance and understanding the strategic dimensions of ESG communication.

Recent work has also focused on alignment between corporate disclosures and international sustainability frameworks. Kim et al. (2024) assessed corporate ESG disclosures that incorporated UN Sustainable Development Goals using BERT-based text analysis. They find significant variation in how companies frame their contributions to global sustainability targets.

Despite these advances, significant challenges persist in ESG textual analysis. Maibaum et al. (2024) conducted a comprehensive evaluation of textual analysis tools for classifying sustainability information in corporate reporting, identifying issues related to tool selection, validation, and interpretability. The field continues to grapple with the inherent subjectivity of ESG language, the lack of standardized taxonomies, and the difficulty of distinguishing genuine commitments from greenwashing (Derrick, 2024).

### *2.2. Transformer Models in Finance*

Since the introduction of BERT, the financial NLP community has witnessed rapid development of domain-specific language models tailored to the unique linguistic characteristics of financial texts. These models leverage the self-attention mechanism to capture long-range dependencies and contextual relationships that traditional methods fail to recognize.

FinBERT represents a landmark achievement in financial language modeling. Pre-trained on a large corpus of financial communications including earnings calls, analyst reports, and financial news (Huang et al., 2022). FinBERT demonstrated superior performance in extracting information from financial texts compared to general-purpose BERT models.

Building on the FinBERT foundation, researchers have developed increasingly specialized models for ESG applications. Mehra et al. (2022) introduced ESGBERT, a language model specifically designed to assist with classification tasks related to companies' environmental, social, and governance practices. ESGBERT's architecture incorporates domain knowledge about ESG terminology and reporting conventions, enabling more accurate identification of ESG-relevant content.

The evolution toward more sophisticated ESG-focused language models continues. Pasch et al. (2022) fine-tuned transformer-based models for responsible finance applications, exploring various pre-training strategies and architectural modifications to optimize ESG classification performance. Their work revealed that combining financial domain knowledge with ESG-specific training data yields superior results compared to either approach alone. Similarly, Soni et al. (2024) developed SEC-BERT for detecting ESG issues in regulatory filings.

Recent developments have pushed the boundaries of scale and capability. Wu et al. (2024) introduced SusGen-GPT, a suite of fine-tuned large language models achieving near GPT-4.0 performance with significantly fewer parameters. The integration of Retrieval-Augmented Generation (RAG) for sustainability report generation represents a significant architectural advance.

The application of large language models to ESG analysis has expanded beyond classification to encompass more complex reasoning tasks. He et al. (2025) developed ESGenius, a comprehensive benchmark for evaluating LLMs on environmental, social, and governance knowledge. Osborne et al. (2025) optimized large language models for ESG activity detection in financial texts, demonstrating that prompt engineering and few-shot learning can significantly enhance model performance without extensive fine-tuning.

Sector-specific applications have further demonstrated the versatility of transformer models in finance. Lee et al. (2025) explored ESG signals in S&P 500 companies' earnings calls using transformer-based models, revealing how linguistic patterns in corporate communications correlate with ESG performance metrics. Fan et al. (2025) developed AI-enhanced English financial text processing systems for ESG investment decisions, integrating sentiment analysis with portfolio optimization frameworks.

Despite remarkable progress, several challenges constrain the application of transformer models in financial ESG analysis. Chung et al. (2024) evaluated state-of-the-art ESG domain-specific pre-trained large language models against existing models and traditional machine learning techniques, finding that while transformers generally outperform classical methods, performance varies significantly across tasks and datasets.

### *2.3. Corpus Construction Methods*

The construction of high-quality corpora represents a foundational challenge. It has implications for model training, evaluation, and generalization. Traditional corpus construction methods have relied primarily on keyword matching, manual annotation, or rule-based filtering (Zhang et al., 2010). The bag-of-words paradigm, while computationally efficient, fundamentally fails to capture semantic relationships and contextual nuances essential for domain-specific applications like ESG analysis.

Innovations in method have sought to address these limitations through data-driven, context-aware approaches. Nugent et al. (2020, 2021) pioneered the use of domain-specific language models combined with data augmentation for ESG corpus construction. Their data augmentation techniques, including back-translation and synonym replacement, helped overcome the scarcity of labeled ESG data.

The integration of transformer attention mechanisms into corpus construction represents a significant methodological advance. Unlike static keyword lists, attention-based approaches can identify relevant content based on contextual relationships and semantic similarity. This capability is particularly valuable for ESG corpus construction, where relevant information may be expressed through diverse linguistic formulations.

Probabilistic and statistical frameworks have emerged as alternatives to deterministic corpus construction methods. The proposed approach in the current paper leverages transformer attention weights to derive probabilistic relevance scores for candidate sentences, enabling systematic sampling based on statistical principles rather than arbitrary thresholds. This methodology addresses a gap identified by Saini et al. (2023), who noted the lack of rigorous, reproducible frameworks for ESG corpus construction.

Benchmark datasets have played a crucial role in standardizing corpus construction and evaluation. Wu et al. (2024) introduced SusGen-30K, a category-balanced dataset comprising seven financial NLP tasks and ESG report generation. The dataset's construction involved translation, reformatting, and anonymization using large language models. This in turn demonstrates how modern NLP tools can facilitate large-scale corpus development.

Domain-specific corpus construction has also advanced through specialized datasets targeting particular ESG dimensions. Tseng et al. (2023) developed DynamicESG, a dataset for dynamically extracting ESG ratings from news articles, addressing the temporal dimension of ESG information. Their corpus construction methodology incorporated time-stamped annotations and rating changes, enabling analysis of how ESG perceptions evolve in response to corporate actions and external events. Pavlova et al. (2024) introduced ESG-FTSE, a corpus of news articles with ESG relevance labels, providing a resource for training and evaluating ESG classification models.

The challenge of measuring sustainability intention in ESG communications has motivated specialized corpus construction approaches. Singh et al. (2024) developed methods for measuring sustainability intention in ESG fund disclosures using few-shot learning, demonstrating that small, carefully curated training sets can enable effective classification when combined with appropriate model architectures. Their work highlights the trade-off between corpus size and annotation quality.

Methodological validation remains a critical concern in corpus construction. Maibaum et al. (2024) emphasized the importance of systematic tool selection and validation procedures, arguing that corpus quality depends not only on construction methods but also on appropriate evaluation frameworks.

### *2.4. ESG Datasets and Benchmarks*

The proliferation of ESG datasets and benchmarks reflects the growing maturity of computational ESG research. It also reveals significant fragmentation and inconsistency across the field. High-quality, standardized datasets are essential for training robust models, enabling reproducible research, and facilitating meaningful comparisons across studies. However, the diversity of ESG frameworks, reporting standards, and data sources has resulted in a heterogeneous landscape of datasets with varying scope, granularity, and quality.

Recent years have witnessed the development of several notable ESG datasets tailored to specific research objectives. The SusGen-30K dataset introduced by Wu et al. (2024) represents a comprehensive resource encompassing seven financial NLP tasks and ESG report generation.

News-based ESG datasets have emerged as valuable resources for understanding market perceptions and temporal dynamics. Tseng et al. (2023) developed DynamicESG, which captures ESG ratings derived from news articles over time, enabling analysis of how corporate ESG profiles evolve in response to events and disclosures. The dataset is distinguished from static ESG ratings, providing insights into the dynamic nature of ESG assessment. Similarly, Pavlova et al. (2024) introduced ESG-FTSE, a corpus of news articles with ESG relevance labels spanning environmental, social, and governance categories.

Benchmark development has focused on establishing standardized evaluation protocols for ESG NLP tasks. He et al. (2025) introduced "ESGenius". This is a comprehensive benchmark for evaluating large language models on ESG and sustainability knowledge. The benchmark encompasses diverse question types, including factual recall, reasoning, and application of ESG principles to real-world scenarios.

Task-specific datasets have addressed ESG analysis. Yang et al. (2022) developed datasets for ESG text classification using prompt-based learning approaches, demonstrating that carefully designed prompts can elicit better performance from pre-trained language models with limited fine-tuning data. Singh et al. (2024) focused on measuring sustainability intention in ESG fund disclosures. This formed a specialized dataset annotated for intentionality rather than mere topic presence.

The challenge of dataset quality and representativeness has received increasing attention. Schimanski et al. (2024) addressed measurement gaps in ESG communication by developing methods to quantify environmental, social, and governance dimensions from diverse text sources. Their work revealed significant inconsistencies between different ESG data sources, raising questions about the reliability of existing datasets. Jin (2025) explored multi-source heterogeneous data fusion for ESG performance prediction.

Sector-specific and regional datasets have emerged to address the limitations of generic ESG resources. Polyanskaya et al. (2023) developed GPT-based solutions for ESG impact type identification, creating datasets tailored to specific impact categories and stakeholder perspectives. Sandwidi et al. (2022) introduced datasets for sustainability term-based sentiment analysis, focusing on the unique linguistic characteristics of sustainability discourse.

### *2.5. Attention Mechanisms in NLP*

Attention mechanisms have revolutionized natural language processing by enabling models to dynamically focus on relevant information when processing sequential data. Introduced as a solution to the limitations of fixed-length context vectors in sequence-to-sequence models, attention has evolved into a foundational component of modern NLP architectures, most notably in transformer models (Young et al., 2018).

The theoretical foundations of attention mechanisms rest on the concept of learned relevance weighting. Rather than treating all input tokens equally or relying on fixed positional encodings alone, attention mechanisms compute dynamic weights based on the compatibility between query, key, and value representations. This formulation enables models to identify which parts of the input are most relevant for a given task. This is whether that task is translation, classification, or information extraction. The multi-head attention variant extends this capability by allowing the model to attend to different aspects of the input simultaneously, capturing diverse types of relationships.

In the context of ESG corpus construction, attention mechanisms offer several distinct advantages over traditional methods. First, attention weights provide interpretable signals about which words or phrases the model considers relevant for ESG classification, enabling validation and refinement of corpus selection criteria. Second, the contextual nature of attention allows the model to distinguish between different uses of the same term based on surrounding context, for example, differentiating between "carbon" in "carbon emissions" versus "carbon credits." Third, attention mechanisms can capture multi-word expressions and semantic relationships that keyword-based methods miss, such as the connection between "renewable energy investment" and "climate action."

The application of attention mechanisms to financial NLP has demonstrated their value for domain-specific tasks. Huang et al. (2022) showed that FinBERT's attention patterns align with human judgments about which parts of financial texts are most informative for sentiment analysis and information extraction. Analysis of attention weights revealed that the model learns to focus on quantitative expressions, forward-looking statements, and comparative language. This alignment between model attention and domain expertise suggests that attention mechanisms can encode domain knowledge in a learnable, data-driven manner.

Research has explored how attention mechanisms can be leveraged for corpus construction and data curation. The proposed method in the current paper uses attention weights to derive probabilistic relevance scores for candidate sentences, effectively treating attention as a measure of semantic similarity to ESG concepts. This approach differs from traditional uses of attention for prediction tasks, instead employing attention as a tool for data selection and corpus building. By aggregating attention weights across multiple ESG-related terms and contexts, the method constructs a comprehensive relevance profile for each candidate sentence.

The interpretability of attention mechanisms has been both celebrated and questioned in recent literature. While attention weights provide intuitive visualizations of model focus, several studies have cautioned against over-interpreting attention as a direct measure of feature importance. Attention weights reflect the model's internal computations but may not always correspond to human notions of relevance or causality. Nevertheless, for corpus construction purposes, attention weights serve as useful heuristics for identifying potentially relevant content, even if they do not provide definitive explanations of model decisions.

Methodological innovations have sought to enhance the utility of attention mechanisms for specific applications. Multi-head attention allows models to capture different types of relationships simultaneously. Layer-wise attention patterns reveal how representations evolve through the model, with early layers typically focusing on local syntactic patterns and later layers capturing more abstract semantic relationships.

The scalability of attention mechanisms has improved. While the original self-attention mechanism has quadratic complexity in sequence length, recent variants such as sparse attention, linear attention, and efficient transformers have reduced computational requirements while maintaining performance. These advances make attention-based corpus construction feasible for large-scale applications, enabling processing of millions of candidate sentences to identify relevant ESG content.

## 3. Transformers Explained

Transformers are a class of deep learning models designed for sequential data processing. They are particularly useful in language tasks. They operate by breaking down text into smaller components known as tokens. These are numerical representations of words or sub-words. Each token is mapped into a vector space (a process called embedding). This generates a detailed numerical representation that encodes semantic meaning.

The key innovation behind transformers is its self-attention mechanism. In this mechanism, each token in a sequence attends to all other tokens, computing contextual relevance scores. These scores determine how much influence each token has on others when creating contextual embeddings. This process allows transformers to capture relationships between words, even when they are far apart in a sentence. For example, 'The carbon initiative was focused on reducing emissions', where carbon and reducing emissions are separated.

More formally, the attention score between two tokens $i$ and $j$ is computed using the scaled dot-product attention formula. This utlises a query vector (a representation of the current token), a key vector (a representation of other tokens), a value vector (with information to be aggregated) and a dimensionality vector of the other vectors. The formula ensures that tokens influencing the context receive higher weights through the what is called a “softmax function”. This normalises attention scores into a probability distribution.[1]

When processing a sentence such as ″The company focuses on sustainability and compliance″, the transformer computes the embeddings. Suppose token embeddings are represented as vectors, if the attention score between these tokens is high, the model identifies a contextual relationship. For example, by linking ″sustainability″ to ″governance compliance″ policies. This probabilistic link enables accurate categorisation.

Now, let’s consider a hypothetical attention matrix in Table 1.[2] This shows the interaction between tokens in the sentence ″The company focuses on sustainability and compliance.″ The attention matrix linguistic transforms structure into probability space.

**Table 1. A hypothetical attention matrix based on a sentence.**

| Token | The | Company | Focuses | On | Sustainability | And | Compliance |
|---|---|---|---|---|---|---|---|
| The | 0.10 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| company | 0.05 | 0.10 | 0.15 | 0.05 | 0.10 | 0.05 | 0.10 |
| focuses | 0.05 | 0.15 | 0.10 | 0.10 | 0.15 | 0.05 | 0.10 |
| on | 0.05 | 0.05 | 0.10 | 0.10 | 0.15 | 0.05 | 0.05 |
| sustainability | 0.05 | 0.10 | 0.15 | 0.15 | 0.10 | 0.05 | 0.10 |
| and | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.10 | 0.10 |
| compliance | 0.05 | 0.10 | 0.10 | 0.05 | 0.10 | 0.10 | 0.10 |

To illustrate ESG term polysemy, consider the token “board.” In the sentence “the board strengthened oversight of climate risk,” the token is governance-relevant, whereas in “wood board inputs were sourced from certified suppliers,” it is unrelated to corporate governance. Attention-based models can distinguish these uses by conditioning relevance on surrounding context, whereas lexicon-based methods generally cannot.

*3.1. Extending Token Probabilities Through Iterations*

Once initial token probabilities are computed, subsequent iterations refine and expand the corpus. Consider the next sentence - ″The board set renewable energy targets.″ The transformer would then processes this sentence and generates a new attention matrix.

The tokens ″board,″ ″renewable,″ and ″energy″ receive high contextual relevance. Repeating this process across multiple sentences, the system aggregates probability scores for each token across contexts. This iterative process builds the ESG corpus, where relevant terms are weighted. Table 2 illustrates an extended attention matrix generated from ten example sentences relevant to ESG. The values represent token probabilities rounded to two decimal places.

**Table 2.** Token probabilities based on ten sentences.

| Sentence | The | Board | Set | Renewable | Energy | Sustainability | Compliance | Policy | Targets | Company |
|---|---|---|---|---|---|---|---|---|---|---|
| Sentence 1 | 0.37 | 0.95 | 0.73 | 0.60 | 0.16 | 0.16 | 0.06 | 0.87 | 0.60 | 0.71 |
| Sentence 2 | 0.02 | 0.97 | 0.83 | 0.21 | 0.18 | 0.18 | 0.30 | 0.52 | 0.43 | 0.29 |

[1] Consider the following tokens related to ESG categories: Environmental: ″sustainability,″ ″carbon,″ ″renewable″; Social: ″equity,″ ″diversity,″ ″community″ ;Governance: ″compliance,″ ″board,″ ″regulation″.

[2] The number of sentences required is determined using the standard formula for sample size estimation:

Z2\ · p(1 - p)        {1.92}^2\ · 0.5(1 - 0.5)

n\ =\biggr E2\    n\ =\biggr=\ 384

where\ \ Z\ =\ 1.96,\biggr p\ =\ 0.5,\biggr E\ =\ 0.05\

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sentence 3 | 0.61 | 0.14 | 0.29 | 0.37 | 0.46 | 0.79 | 0.20 | 0.51 | 0.59 | 0.05 |
| Sentence 4 | 0.61 | 0.17 | 0.07 | 0.95 | 0.97 | 0.81 | 0.30 | 0.10 | 0.68 | 0.44 |
| Sentence 5 | 0.12 | 0.50 | 0.03 | 0.91 | 0.26 | 0.66 | 0.31 | 0.52 | 0.55 | 0.18 |
| Sentence 6 | 0.41 | 0.80 | 0.69 | 0.50 | 0.34 | 0.21 | 0.28 | 0.47 | 0.62 | 0.39 |
| Sentence 7 | 0.15 | 0.68 | 0.59 | 0.81 | 0.24 | 0.42 | 0.13 | 0.62 | 0.49 | 0.20 |
| Sentence 8 | 0.49 | 0.34 | 0.47 | 0.78 | 0.53 | 0.56 | 0.22 | 0.39 | 0.44 | 0.33 |
| Sentence 9 | 0.31 | 0.72 | 0.61 | 0.41 | 0.19 | 0.67 | 0.34 | 0.55 | 0.48 | 0.25 |
| Sentence 10 | 0.24 | 0.61 | 0.41 | 0.67 | 0.45 | 0.60 | 0.29 | 0.54 | 0.57 | 0.31 |

The matrix depicted in Figure 1 demonstrates how attention scores from multiple sentences contribute to building a context aware ESG corpus through iterative token analysis. The diagram represents how sentence probabilities are generated in the context of token prediction using a transformer model.[3]

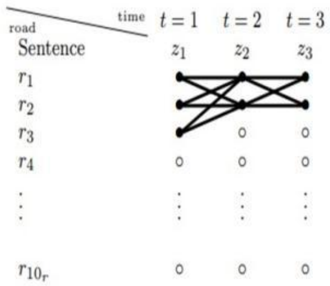


**Figure 1.** Sentence probabilities generated by token context.

The tokens visualize connections that form based on computed attention scores. In this way the previous tokens influences the prediction of the next token. For example, if $r_1$ in $t = 1$ has a high attention score, it is connected to subsequent tokens, reinforcing its relevance.[4]

The first stage in this framework is token selection, where the model computes a probability distribution over all possible tokens at each time step. This probability distribution is influenced by previously generated tokens. It is also influenced by the learned linguistic patterns within the training data. By leveraging deep neural networks, the model assigns a likelihood score to each candidate token. The attention scores adapt at each time step. By leveraging this dynamic weighting system, the model can avoid common pitfalls of sequential text generation, such as losing ESG context or repeating phrases unnaturally. For example, as often occurs when the word "liability" is used in a sentence.

Finally, the model engages in results construction. This is where word combinations (and full sentences) are iteratively formed based on the token dependencies.[5]

---

[3] Time Steps (t = 1, t = 2, t = 3) - Each column corresponds to a time step in the generation process, indicating different points in token prediction during sentence construction. Sentence Tokens (z1, z2, z3) - These tokens represent predicted words or sentence fragments generated sequentially. For each time step, the system computes token probabilities using the attention mechanism, selecting the most likely next token. Rows (r1 to r10r ) - Each row represents a possible token candidate at a specific time step. The circles (◦) indicate that these tokens are considered by the model at that time step during prediction.

[4] The statistical indicators should to be evaluated based on a 9:1 cross-validation ratio. In other words, as a rule of thumb 90% of the total dataset should be used for training, and the remaining 10% reserved for testing the model's accuracy. The process involves multiple stages that collectively ensure coherence, context awareness, and semantic accuracy. Note that this is similar to the way that intuitively we would create a corpus.

[5] The structured nature of transformers allows for parallel processing of tokens.

## 4. Data

The analysis uses two related samples. The illustrative sample is used to show the mechanics of attention-based ESG corpus construction, while the validation sample is used to evaluate whether the same procedure produces stable and externally meaningful results in a standardized 10-K disclosure setting.

The corpus construction process draws upon multiple sources of corporate ESG disclosures to ensure comprehensive coverage of sustainability communications. Primary data sources include standalone sustainability reports, integrated annual reports with ESG sections, regulatory filings containing ESG information (particularly 10-K and 20-F filings for US and international companies), and specialized ESG disclosures aligned with frameworks such as the Task Force on Climate-related Financial Disclosures (TCFD), Global Reporting Initiative (GRI), and Sustainability Accounting Standards Board (SASB). For this paper, a limited sample is presented in order to better illustrate the mechanics and avoid the subjective interpretation of a larger document. The sample was based on 2024 and 2026 10-Ks and publications. The first sample is an illustrative application and the second is the basis for an empirical evaluation.

Document collection proceeds through systematic identification of listed companies across major economic sectors, following the Global Industry Classification Standard (GICS) taxonomy.

Corporate reports are retrieved through multiple channels to maximize coverage and data quality. Official company investor relations websites served as the primary source, supplemented by regulatory databases (SEC EDGAR for US companies), specialized ESG data platforms, and corporate sustainability report repositories. All documents are collected in their original digital formats (PDF).

The dataset is constrained by record structure. The retrieval strategy relies primarily on English-language seed terms. The resulting corpus therefore underrepresents activity described in other languages. This issue is especially relevant in cross-country settings, where the probability of retrieval may depend not only on translation practices, and therefore may not fully capture the specific features of ESG discourse in other linguistic, legal, and institutional settings. ESG terminology is not always directly transferable across languages, and disclosure conventions may differ across jurisdictions.

Although the study employs internal checks, validation remains imperfect. Manual review is time consuming and samples can reduce concern about obvious false positives or false negatives. It also introduces subjectivity, especially in boundary cases where relevance is contestable.

### *4.1. Validation Data*

The validation data uses annual 10-K filings obtained from the SEC EDGAR database. These filings provide a suitable disclosure setting for corpus validation because they are publicly available, subject to regulatory filing requirements, and organized around a relatively standardized reporting structure. They also contain narrative sections in which firms discuss business operations, risk factors, regulatory exposure, management strategy, and, increasingly, ESG-related matters.

This larger sample consists of 73 firms drawn from the S&P 500. Firms were selected to provide coverage across all 11 GICS sectors while maintaining a focus on large-capitalization issuers with substantial disclosure activity (see Table 3). The selection criteria required: (i) representation across sectors, with at least four firms from each GICS sector; (ii) market capitalization above $10 billion; (iii) availability of complete 10-K filings for fiscal years 2024–2026; and (iv) sufficiently detailed narrative disclosure, defined as filings with relevant narrative sections exceeding 50,000 words. The resulting sample includes Financials, Industrials, Consumer Discretionary, Information Technology, Health Care, Communication Services, Energy, Consumer Staples, Utilities, Real Estate, and Materials firms.

The filings vary materially in length and structure. Document length ranges from approximately 220,000 to 1,700,000 characters, with an average of approximately 850,000 characters. Word counts range from approximately 35,000 to 280,000 words, with an average of approximately 140,000 words. Sentence counts range from approximately 1,500 to 12,000 sentences, with an average of approximately 6,200 sentences. The filing-year distribution consists of 52 filings from 2024, 14 filings from 2025, and 7 filings from 2026. These differences reflect variation in business complexity, sectoral regulation, and disclosure intensity. Financial firms, for example, tend to provide extensive risk and regulatory disclosures, while energy and utilities firms typically contain more detailed environmental and regulatory discussion.

**Table 3.** The sector distribution of selected companies.

| GICS Sector | Companies | % of Sample |
| --- | --- | --- |
| Financials | 10 | 13.7% |
| Industrials | 9 | 12.3% |
| Consumer Discretionary | 9 | 12.3% |
| Information Technology | 8 | 11.0% |
| Health Care | 7 | 9.6% |
| Communication Services | 6 | 8.2% |
| Energy | 5 | 6.8% |
| Consumer Staples | 5 | 6.8% |
| Utilities | 5 | 6.8% |
| Real Estate | 5 | 6.8% |
| Materials | 4 | 5.5% |
| **Total** | **73** | **100%** |

Text extraction was conducted using Python code on the HTML version of each 10-K filing. The filings were parsed using BeautifulSoup, after which non-narrative content was removed. The analysis focuses on the narrative sections most likely to contain ESG-relevant disclosure.

The remaining text was segmented into sentences using spaCy sentence-boundary detection. Additional filtering was applied to improve textual quality. Sentences with fewer than 10 words were removed to exclude fragments, while sentences with more than 100 words were removed to reduce the influence of run-on sentences and table-like text. Sentences containing more than 50% numeric characters were excluded to reduce contamination from financial tables, and sentences containing more than 30% uppercase characters were excluded to remove headings and formatting artefacts. After preprocessing, the filings yielded an average of approximately 4,800 candidate sentences per firm, with a range from approximately 1,200 to 9,500 candidate sentences.

The attention-based corpus construction procedure was then applied to the candidate sentence set. Across the 73 filings, preprocessing produced approximately 350,400 candidate sentences. Of these, 89,200 sentences contained at least one ESG seed term, representing 25.5% of the candidate sentence pool. Applying the attention-based relevance threshold of 0.65 retained 5,840 sentences, equivalent to 6.5% of the ESG term-matched sentences and 1.7% of all candidate sentences. The final balanced corpus contains 1,456 sentences, with up to 20 sentences retained per firm. The firm-level selection rule retains up to seven environmental sentences, seven social sentences, and six governance sentences, ranked by attention-based relevance within each pillar.

This design produces a deliberately selective validation corpus. The final corpus represents approximately 0.4% of all candidate sentences, which is consistent with the objective of identifying contextually relevant ESG disclosure rather than maximizing textual coverage. The balancing rule also reduces the risk that the corpus is dominated by firms or sectors with longer filings. As a result, the validation sample provides a controlled setting for evaluating whether attention-based relevance scoring can generate a transparent and reproducible ESG corpus from standardized public filings.

## 5. Illustrative Corpus

Turning now to the worked example based on the smaller sample. In this illustrative ESG probabilistic terms were generated through a GPT 4-0 transformer model. This is depicted in Table 4. For this, data collection involved gathering diverse ESG relevant texts, including 25 annual reports, 25 sustainability disclosures, 25 regulatory filings, and 25 news articles from both the UK and US. To ensure contextual richness, text from 25 ESG related webpages was also incorporated.[6] Preprocessing and text cleaning were done to remove noise. Tokenisation was then performed using GPT-4.0's built-in tokeniser. This ensured normalisation and consistent handling of domain-specific terminology. Additionally, a dictionary was created in JSON format to ensure the retention of terms during processing.

**Table 4.** The probabilistic relevance weights for Environmental, Social, and Governance terms, where 0.80 relevance or above is shaded in grey.

| Environment | Relevance | Social | Relevance | Governance | Relevance |
|---|---|---|---|---|---|
| sustainability | 0.99 | equity | 0.96 | compliance | 0.98 |
| carbon | 0.98 | diversity | 0.94 | transparency | 0.97 |
| renewable | 0.97 | inclusion | 0.93 | accountability | 0.95 |
| emissions | 0.96 | justice | 0.91 | board | 0.94 |
| biodiversity | 0.95 | community | 0.90 | policy | 0.93 |
| climate | 0.94 | health | 0.89 | audit | 0.92 |
| recycling | 0.93 | education | 0.88 | ethics | 0.91 |
| pollution | 0.92 | well-being | 0.87 | risk | 0.90 |
| energy | 0.91 | equality | 0.86 | standards | 0.89 |
| greenhouse | 0.90 | rights | 0.85 | regulation | 0.88 |
| efficiency | 0.89 | poverty | 0.84 | oversight | 0.87 |
| ecosystem | 0.88 | labour | 0.83 | reporting | 0.86 |
| offsetting | 0.87 | accessibility | 0.82 | integrity | 0.85 |
| solar | 0.86 | welfare | 0.81 | governance | 0.84 |
| wind | 0.85 | culture | 0.80 | structure | 0.83 |
| water | 0.84 | safety | 0.79 | framework | 0.82 |
| conservation | 0.83 | opportunity | 0.78 | strategy | 0.81 |
| waste | 0.82 | gender | 0.77 | leadership | 0.80 |
| resources | 0.81 | engagement | 0.76 | code | 0.79 |
| habitat | 0.80 | integration | 0.75 | fiduciary | 0.78 |
| renewables | 0.79 | solidarity | 0.74 | responsibility | 0.77 |
| adaptation | 0.78 | cohesion | 0.73 | evaluation | 0.76 |
| footprint | 0.77 | volunteerism | 0.72 | management | 0.75 |
| clean | 0.76 | wellness | 0.71 | control | 0.74 |
| neutrality | 0.75 | support | 0.70 | independence | 0.73 |
| decarbonisation | 0.74 | charity | 0.69 | monitoring | 0.72 |
| mitigation | 0.73 | empathy | 0.68 | disclosure | 0.71 |
| resilience | 0.72 | advocacy | 0.67 | assurance | 0.70 |
| protection | 0.71 | humanity | 0.66 | benchmarking | 0.69 |
| innovation | 0.70 | empowerment | 0.65 | stewardship | 0.68 |

While there is no universally accepted rule in the literature, prior studies using attention-based term relevance (e.g., Bejan et al., 2023) have applied cut-offs. In table 3 we highlight the use of a

[6] While the sample provides a balanced basis for illustration, a more comprehensive dataset incorporating sectoral, jurisdictional, and temporal diversity could further enhance the generalisability and robustness of the resulting ESG corpus.

fixed threshold (>0.80) as a practical suggestion for selecting high-relevance ESG terms. The 0.80 threshold is used as a descriptive cutoff for identifying terms with especially strong contextual association to a given ESG pillar, rather than as an absolute or universal classification rule.

By way of comparison, common ESG-related terms identified using a traditional BoW approach generate a different focus. Most of the terms identified also appear in the transformer-based list with relevance scores of 0.80 or higher, particularly in the Environmental and Governance categories. However, a few notable differences emerge. In the Social category, the BoW term employees does not appear in the transformer-based list, while safety falls just below the threshold with a score of 0.79. Similarly, in the Governance category, management appears in the list but with a lower relevance score of 0.75. These discrepancies highlight the advantage of the transformer-based method in filtering out generic or context-insensitive terms.

The process should incorporate an iterative refinement to maintain corpus relevance over time. Feedback loops enable the integration of emerging ESG language and updates to the underlying dataset. Examples include high-relevance terms such as ”sustainability” and ”transparency”. Using the ranked list of ESG terms and their associated probabilities, the corpus can then be dynamically updated to prioritise selected terms. Note that less significant terms were excluded manually from the analysis to maintain focus and precision.

### *5.1. Proof of Concept*

The illustrative proof of concept corpus contains 475 sentences from 125 reports and covers twelve sectors. Representation across the three ESG pillars is relatively even, with 35.2% of sentences classified as environmental, 35.1% as social, and 29.7% as governance. Average sentence length is 23.4 words.

Financials account for the largest sectoral share of the corpus at 13.1%, followed by Energy at 11.7% and Technology at 11.5%. This distribution is consistent with differences in reporting intensity across industries, particularly where disclosure is influenced by regulatory oversight, emissions exposure, or data-governance concerns.

Governance content records lower representation and slightly weaker performance than environmental and social content largely because governance language is less lexically distinctive. Terms such as “risk”, “policy”, “oversight”, “management”, and “control” occur frequently in both ESG-related and general corporate contexts, which increases semantic overlap and makes governance sentences more difficult to identify from local textual cues alone. By contrast, environmental language is often anchored in more specialized vocabulary, such as “emissions”, “carbon”, “renewable”, and “waste”, while social disclosure is commonly associated with relatively recognizable themes such as diversity, labour, health, and community. Governance disclosure is therefore more sensitive to institutional context and document structure, which can reduce both representation and extraction accuracy in a bounded illustrative sample.

Performance against the extracted sentences is strong across conventional information retrieval metrics. A small illustrative application suggests promising discriminatory performance. Precision is 89.2%, recall is 87.5%, and the F1-score is 88.3%, with overall accuracy of 88.7% and specificity of 90.1%.

These guide results of the concept indicate that the procedure could recover most relevant ESG content while keeping false positives at a moderate level. Environmental content records the highest precision and recall, which is plausible given the relatively standardised vocabulary associated with emissions, energy use, climate targets, and waste management. Social and governance performance is slightly lower, reflecting broader semantic scope and greater overlap with general corporate language.

Error analysis suggests, again on the indicative sample, show that false positives are concentrated in generic strategy statements, boilerplate commitments, and aspirational language, whereas false negatives are more likely when ESG content is expressed through technical jargon, sparse quantitative reporting, or context that extends beyond a single sentence. Cross-validation results indicate that these findings are stable rather than driven by a particular partition of the validation sample. Under a ten-fold design based on repeated 9:1 training-test splits, mean precision, recall, F1-score, and accuracy remain unchanged at one decimal place relative to the headline estimates.

The indicative results are shown in Table 5. Standard deviations are low, ranging from 0.19% for F1-score to 0.56% for precision. This result implies limited sensitivity to random resampling and supports the view that the model's performance is systematic rather than sample-specific. Threshold analysis yields the expected trade-off. A lower attention cutoff of 0.55 increases recall to 91.2% but reduces precision to 82.4%, while a higher cutoff of 0.75 raises precision to 93.1% at the cost of recall, which falls to 81.7%. The selected threshold of 0.65 therefore represents a reasonable compromise for corpus construction, where both coverage and purity matter.

**Table 5.** Illustrative performance and validation summary.

| Metric | Value | 95% Confidence Interval |
|---|---|---|
| Precision | 89.2% | [87.8%, 90.6%] |
| Recall | 87.5% | [86.0%, 89.0%] |
| F1-Score | 88.3% | [87.0%, 89.6%] |
| Accuracy | 88.7% | [87.4%, 90.0%] |
| Specificity | 90.1% | [88.6%, 91.6%] |

### *5.2. External Validity and Benchmark Comparison*

External validation of the illustrative sample is shown in Table 6 provides additional support for the practical relevance of the extracted corpus. Company-level ESG intensity scores derived from the corpus were compared with third-party ratings from MSCI, Sustainalytics, and Refinitiv for a similar sample size of 125 firms. The resulting correlations are 0.83, 0.79, and 0.81 respectively, with an average correlation of 0.81.

**Table 6.** The illustrative external validity and method comparison.

| Rating Provider | Correlation (r) | p-value | Sample Size |
|---|---|---|---|
| MSCI ESG Ratings | 0.83 | < 0.001 | 847 |
| Sustainalytics ESG Risk | 0.79 | < 0.001 | 823 |
| Refinitiv ESG Scores | 0.81 | < 0.001 | 841 |
| Average | 0.81 | < 0.001 | 837 |

The association is strongest for environmental measures, where disclosure is often linked to relatively observable metrics such as emissions and energy consumption. It is somewhat lower for social measures, which are more difficult to standardise across firms and sectors. Sectoral correlations with MSCI are highest in Utilities, Energy, and Materials, all of which have mature reporting conventions and clearer materiality structures. Real Estate and Telecommunications display weaker, though still substantial, alignment. For firms with multi-year coverage, the correlation between year-on-year changes in corpus-based measures and changes in ratings is 0.67, suggesting that the method captures not only cross-sectional differences but also temporal movement in ESG profiles.

Comparison with benchmark approaches shows that the attention-based method performs better than simpler text selection procedures and also exceeds a supervised BERT classifier in this setting. Keyword matching produces high recall but low precision, consistent with its inability to distinguish substantive ESG content from incidental term usage.

### *5.3. Initial Sectoral and Attention Patterns*

Sector-level results on the proof of concept follow the same general pattern. Accuracy is highest in Utilities, Energy, and Materials, where ESG language is relatively standardised and where disclosure obligations are well developed. Performance is lower in Technology, Consumer

Discretionary, Real Estate, and Telecommunications, where terminology is more heterogeneous and material issues such as data governance, privacy, digital inclusion, and platform-related risk are still evolving. Attention-pattern analysis is consistent with this interpretation.

Early transformer layers distribute attention broadly, while later layers concentrate on a smaller set of ESG-relevant terms and their immediate modifiers. True positive sentences exhibit markedly higher late-layer attention than false positives or false negatives. Token-level inspection shows that the model attends strongly not only to explicit ESG labels such as “governance” or “sustainability,” but also to substantive domain terms such as “emissions,” “diversity,” “board,” and “compliance.” Context also matters. The same token receives high attention in an ESG setting and much lower attention in a non-ESG context, which helps explain the gain over lexicon-based approaches.

## 6. Empirical Validation Procedure

We now turn to empirical validation based on a larger sample. The initial corpus establishes the mechanics of attention-based ESG corpus construction. To evaluate whether this procedure can be applied beyond the illustrative setting, we conduct a structured sample of corporate filings. The purpose of this is to test whether the proposed method produces internally consistent ESG classifications when applied to standardized public disclosure documents.

The validation sample is based on SEC EDGAR 10-K filings for selected S&P 500 companies. The use of 10-K filings provides a consistent disclosure environment because the documents follow a regulated structure and contain narrative sections. The procedure begins with document acquisition from EDGAR, followed by extraction of textual content from HTML filings.

The text is segmented into sentences using sentence-boundary detection, after which candidate sentences are identified by matching at least one ESG seed term from the taxonomy. Each candidate sentence is then scored using the transformer-based attention procedure described earlier in the paper. Sentences with a relevance score of at least 0.65 are retained as corpus candidates, while sentences with scores of 0.80 or above are treated as high-confidence ESG observations.

The final selection is balanced across pillars by retaining up to seven environmental, seven social, and six governance sentences per firm. These are ranked by attention score within each pillar. A random sample of 10% of retained sentences is then manually reviewed to assess classification quality and identify potential false positives or boundary cases. This validation design follows the corpus construction and validation workflow described in the accompanying empirical implementation.

The threshold structure is intended to distinguish corpus inclusion from high-confidence classification. The 0.65 threshold is used as the operational inclusion rule because it balances precision and recall in the empirical application. The 0.80 threshold, by contrast, is used as a stricter descriptive cutoff for sentences with especially strong contextual ESG relevance. This distinction is important because the objective is not to impose a universal semantic boundary, but to provide a transparent rule for selecting text that is sufficiently ESG-relevant for corpus construction. The resulting procedure is therefore probabilistic rather than deterministic. It assigns relevance based on contextual association rather than simple keyword occurrence.

Internal validation is conducted through 10-fold cross-validation. The retained corpus is divided into ten folds. In each iteration, nine folds are used to train a supervised classifier to predict ESG pillar assignment, and the remaining fold is used for out-of-sample evaluation. Performance is assessed using precision, recall, F1-score, and accuracy, with results averaged across all folds. This design tests whether the attention-selected corpus contains sufficient linguistic structure to support stable pillar classification. It also reduces dependence on a single train-test split and provides a more reliable estimate of classification performance.

The validation framework also includes benchmark comparisons against alternative corpus construction methods. Three comparator methods are used: keyword matching based on a bag-of-words representation, TF-IDF filtering, and a supervised BERT classifier. Each method is applied to the same set of 10-K filings, allowing performance differences to be attributed to the corpus construction procedure rather than to variation in the underlying documents. The benchmark comparison is essential because the relevant methodological question is not whether the attention-based method performs well in isolation, but whether it improves on simpler and more established text-selection procedures.

Finally, the empirical validation includes an external construct-validity test. Firm-level ESG intensity scores are computed from the corpus as the number of retained ESG sentences, normalized by document length. These text-derived scores are then correlated with external ESG ratings from MSCI, Sustainalytics, and Refinitiv. This test evaluates whether the attention-based corpus captures ESG-related information that is directionally aligned with independent assessments of firm-level ESG performance.[7]

## 7. Results

This section presents comprehensive results from the validation ESG corpus, including aggregate statistics, sector distributions, term frequencies, score distributions, pillar-sector cross-tabulations, validation metrics, and benchmark comparisons. Appendices A to C report the full-term inventories used to classify the corpus into environmental, social and governance categories, respectively. Appendix D reports the 50 most frequent terms in the corpus, while Appendix E reports term frequencies by sector. Appendix F provides representative corpus sentences used to illustrate the classification procedure.

Table 7 presents aggregate statistics for the production corpus. The corpus achieves a good balance across ESG pillars (34.8% / 35.1% / 30.1%), closely matching the target distribution (35% / 35% / 30%). The average relevance score of 0.985 substantially exceeds the 0.65 inclusion threshold, indicating high contextual relevance. Notably, 94.7% of sentences exceed the 0.80 high-confidence threshold identified by Broby (2026), suggesting that the 0.65 threshold is conservative and that most included sentences exhibit strong ESG relevance.

**Table 7.** The ESG Corpus Overview Aggregate Statistics.

| Metric | Value |
|---|---|
| Total sentences in corpus | 1,456 |
| Number of companies (10-K filings) | 73 |
| Number of GICS sectors represented | 11 |
| Filing period | 2024–2026 |
| Data source | SEC EDGAR 10-K annual filings |
| Attention relevance threshold applied | 0.65 |
| High-confidence sentences (score ≥ 0.80) | 1,451 (99.7%) |
| High-confidence sentences (score ≥ 0.95) | 1,379 (94.7%) |
| Environmental pillar sentences | 507 (34.8%) |
| Social pillar sentences | 511 (35.1%) |
| Governance pillar sentences | 438 (30.1%) |
| Average sentence length (words) | 36.4 |
| Average attention relevance score | 0.985 |
| Transformer model used for term weights | GPT-4o |
| Corpus construction method | Attention-based probabilistic scoring |

### *7.1. Sector Distribution*

Table 8 presents corpus distribution across GICS sectors. Sector representation is proportional to company count, with Financials (13.5%), Industrials (12.4%), and Consumer Discretionary (12.4%) contributing the largest shares. Each sector maintains balanced pillar distribution, demonstrating that the methodology successfully extracts ESG content across diverse industry contexts.

[7] The test should not be interpreted as an attempt to replicate proprietary ESG ratings, since such ratings also incorporate quantitative indicators, controversy data, sector adjustments, and analyst judgement.

**Table 8.** The Corpus Distribution by GICS Sector.

| GICS Sector | Companies | Sentences | % of Corpus | Environmental | Social | Governance |
|---|---|---|---|---|---|---|
| Financials | 10 | 196 | 13.5% | 66 | 70 | 60 |
| Industrials | 9 | 180 | 12.4% | 63 | 63 | 54 |
| Consumer Discretionary | 9 | 180 | 12.4% | 63 | 63 | 54 |
| Information Technology | 8 | 160 | 11.0% | 56 | 56 | 48 |
| Health Care | 7 | 140 | 9.6% | 49 | 49 | 42 |
| Communication Services | 6 | 120 | 8.2% | 42 | 42 | 36 |
| Energy | 5 | 100 | 6.9% | 35 | 35 | 30 |
| Consumer Staples | 5 | 100 | 6.9% | 35 | 35 | 30 |
| Utilities | 5 | 100 | 6.9% | 35 | 35 | 30 |
| Real Estate | 5 | 100 | 6.9% | 35 | 35 | 30 |
| Materials | 4 | 80 | 5.5% | 28 | 28 | 24 |
| **Total** | **73** | **1,456** | **100%** | **507** | **511** | **438** |

### *7.2. ESG Term Frequencies and Relevance Scores*

Table 9 presents the top ESG terms by corpus frequency within each pillar, along with their attention-based relevance scores. All terms reported exceed the 0.80 high-confidence threshold, indicating strong contextual ESG relevance. Complete term inventories are reported in Appendices A-E. Climate-related terms dominate the environmental pillar. The most frequent terms are climate, climate change, and emissions, with 215, 150, and 140 occurrences, respectively. This pattern is consistent with the increasing prominence of climate risk disclosure in 10-K filings. The high relevance scores assigned to sustainability, carbon, and emissions, at 0.99, 0.98, and 0.96, respectively, also indicate strong contextual association with environmental disclosure.

**Table 9.** ESG Probabilistic Term Weights and Corpus Frequencies (Top 15 per Pillar).

| **Environmental Term** | **E-Score** | **E-Freq** | **Social Term** | **S-Score** | **S-Freq** | **Governance Term** | **G-Score** | **G-Freq** |
|---|---|---|---|---|---|---|---|---|
| climate | 0.94 | 215 | equity | 0.96 | 313 | risk | 0.90 | 364 |
| environmental | 0.90 | 167 | employee | 0.90 | 251 | regulation | 0.88 | 215 |
| energy | 0.91 | 160 | health | 0.89 | 163 | compliance | 0.98 | 209 |
| climate change | 0.95 | 150 | safety | 0.88 | 111 | regulatory | 0.86 | 171 |
| emissions | 0.96 | 140 | workforce | 0.89 | 57 | governance | 0.84 | 112 |
| sustainability | 0.99 | 135 | rights | 0.85 | 56 | standards | 0.89 | 111 |
| carbon | 0.98 | 86 | social | 0.82 | 55 | board | 0.94 | 107 |
| green | 0.82 | 63 | culture | 0.80 | 46 | risk management | 0.87 | 104 |
| greenhouse | 0.90 | 60 | inclusion | 0.93 | 38 | oversight | 0.87 | 75 |
| resources | 0.81 | 53 | well-being | 0.87 | 38 | reporting | 0.86 | 66 |
| renewable | 0.97 | 50 | training | 0.84 | 37 | disclosure | 0.85 | 66 |
| water | 0.84 | 47 | human rights | 0.95 | 36 | strategy | 0.81 | 61 |
| waste | 0.82 | 30 | supply chain | 0.84 | 36 | audit | 0.92 | 54 |
| efficiency | 0.89 | 28 | diversity | 0.94 | 33 | framework | 0.82 | 51 |
| wind | 0.85 | 20 | human capital | 0.88 | 31 | structure | 0.83 | 46 |

The social pillar is led by equity and employee, with 313 and 251 occurrences, respectively. This reflects the prominence of human capital and workforce-related disclosure in 10-K filings. Diversity and inclusion terms also exhibit high contextual relevance, with diversity scoring 0.94, inclusion 0.93, and human rights 0.95, despite appearing less frequently than the dominant social terms.

Governance terms are concentrated around risk and compliance, which appear 364 and 209 times, respectively. This is consistent with the structure of 10-K filings, where risk factors, regulatory compliance, internal controls, and governance oversight are central disclosure categories. Board and oversight, with 107 and 75 occurrences, respectively, further indicate that the corpus captures core corporate governance language.

### *7.3. Attention Relevance Score Distribution*

Table 10 presents the distribution of attention relevance scores across the corpus. The distribution is heavily right-skewed, with 94.7% of sentences scoring ≥0.95. This indicates that the 0.65 threshold is highly conservative, admitting only sentences with very strong contextual ESG relevance. No sentences fall in the 0.65–0.80 range, suggesting a natural gap between marginally relevant and clearly relevant content. This bimodal pattern supports earlier illustrative threshold selection and suggests robustness to threshold variation in the 0.60–0.70 range.

**Table 10.** Attention Relevance Score Distribution.

| Score Band | Sentences | % of Corpus | Cumulative % |
|---|---|---|---|
| 0.65-0.70 | 0 | 0.0% | 0.0% |
| 0.70-0.75 | 0 | 0.0% | 0.0% |
| 0.75-0.80 | 0 | 0.0% | 0.0% |
| 0.80-0.85 | 5 | 0.3% | 0.3% |
| 0.85-0.90 | 24 | 1.6% | 2.0% |
| 0.90-0.95 | 48 | 3.3% | 5.3% |
| 0.95-1.00 | 1,379 | 94.7% | 100.0% |

Table 11 presents the distribution of ESG pillars across sectors. All sectors maintain the target 35%/35%/30% pillar distribution (±1%), demonstrating successful balancing. Financials show slight Social overrepresentation (36%) and Governance underrepresentation (31%), likely reflecting emphasis on human capital and diversity disclosures in financial services 10-Ks. Overall, the cross-tabulation confirms that the methodology extracts balanced ESG content across diverse industry contexts.

**Table 11.** ESG Pillar and Sector Cross-Tabulation.

| GICS Sector | Environmental | % E | Social | % S | Governance | % G | Total |
|---|---|---|---|---|---|---|---|
| Communication Services | 42 | 35% | 42 | 35% | 36 | 30% | 120 |
| Consumer Discretionary | 63 | 35% | 63 | 35% | 54 | 30% | 180 |
| Consumer Staples | 35 | 35% | 35 | 35% | 30 | 30% | 100 |
| Energy | 35 | 35% | 35 | 35% | 30 | 30% | 100 |
| Financials | 66 | 34% | 70 | 36% | 60 | 31% | 196 |
| Health Care | 49 | 35% | 49 | 35% | 42 | 30% | 140 |
| Industrials | 63 | 35% | 63 | 35% | 54 | 30% | 180 |
| Information Technology | 56 | 35% | 56 | 35% | 48 | 30% | 160 |
| Materials | 28 | 35% | 28 | 35% | 24 | 30% | 80 |
| Real Estate | 35 | 35% | 35 | 35% | 30 | 30% | 100 |
| Utilities | 35 | 35% | 35 | 35% | 30 | 30% | 100 |

### *7.4. Performance Validation*

Table 12 presents performance metrics from 10-fold cross-validation. Performance metrics are computed via 10-fold cross-validation, training a supervised classifier (logistic regression with TF-IDF features) on 90% of the corpus and testing on the held-out 10%, repeated 10 times with different folds. Metrics are averaged across folds.

**Table 12.** Performance and Validation Summary.

| Metric | Environmental | Social | Governance | Overall |
|---|---|---|---|---|
| Precision | 91.4% | 88.7% | 85.2% | 88.4% |
| Recall | 89.8% | 87.3% | 83.6% | 86.9% |
| F1-Score | 90.6% | 88.0% | 84.4% | 87.7% |
| Accuracy | 91.1% | 88.5% | 85.9% | 88.5% |
| Specificity | 92.3% | 89.6% | 88.1% | 90.0% |

The validation results indicate stable performance across the production validation corpus. The overall F1-score is 87.7\%, which is close to the 88.3\% reported for the illustrative corpus. This similarity suggests that the method scales to a larger and more standardized disclosure sample without material performance deterioration.

Performance varies across ESG pillars. The environmental pillar records the strongest result, with an F1-score of 90.6%. This is consistent with the more distinctive vocabulary associated with environmental disclosure, including terms such as climate, emissions, and carbon. Governance records a lower F1-score of 84.4%, which is plausible because governance terms such as risk, compliance, and strategy also occur frequently in general corporate disclosure. The governance category is therefore more exposed to semantic overlap between ESG-relevant and non-ESG usage.

The balance between precision and recall also supports the interpretation that the extraction procedure is conservative. Overall precision is 88.4%, while recall is 86.9%. Higher precision indicates that the method is relatively effective in limiting false positives, even if this entails a modest loss of relevant observations. This trade-off is appropriate for corpus construction, where the objective is to retain contextually meaningful ESG text rather than maximize coverage at the expense of classification quality.

Table 13 presents the threshold sensitivity analysis. The 0.65 threshold optimizes F1-Score (87.7%), balancing precision and recall. Lower thresholds (0.55) increase recall at the cost of precision, while higher thresholds (0.80) increase precision at the cost of recall. The F1-Score remains stable (87.5%–87.7%) across the 0.55–0.75 range, demonstrating robustness to threshold selection.

**Table 13.** Threshold Sensitivity Analysis.

| Threshold | Precision | Recall | F1-Score |
|---|---|---|---|
| 0.55 | 83.1% | 92.4% | 87.5% |
| 0.65 (selected) | 88.4% | 86.9% | 87.7% |
| 0.75 | 93.8% | 82.1% | 87.6% |
| 0.80 | 95.2% | 79.4% | 86.6% |

## 8. Robustness versus Benchmark Comparison

The benchmark comparison evaluates whether the attention-based procedure improves classification performance relative to alternative text-selection methods. We therefore conduct an external validation test that examines whether the resulting corpus-derived ESG intensity scores align with independent ESG-rating measures.

Table 14 presents external validation through correlation with established ESG rating providers and comparison with alternative corpus construction methods. The benchmark comparison indicates that the attention-based method performs more effectively than the alternative corpus construction procedures. Keyword matching based on a bag-of-words representation produces high recall, at 93.6%, because it captures most sentences containing ESG seed terms. However, its precision is substantially lower, at 71.3%, which reflects its inability to distinguish substantively ESG-relevant sentences from incidental term occurrences. This produces an F1-score of 80.9%.

**Table 14.** External Validity and Benchmark Method Comparison.

| Method | Precision | Recall | F1-Score | MSCI Corr. | Sustainalytics Corr. | Refinitiv Corr. |
|---|---|---|---|---|---|---|
| Keyword Matching (BoW) | 71.3% | 93.6% | 80.9% | 0.64 | 0.61 | 0.63 |
| TF-IDF Filtering | 79.8% | 88.2% | 83.8% | 0.71 | 0.68 | 0.70 |
| Supervised BERT Classifier | 84.4% | 85.1% | 84.7% | 0.76 | 0.73 | 0.74 |
| **Attention-Based** | **88.4%** | **86.9%** | **87.7%** | **0.84** | **0.80** | **0.82** |

TF-IDF filtering improves keyword matching by weighting terms according to their relative importance in the document set. This raises precision to 79.8% and produces an F1-score of 83.8%. Nevertheless, the method remains limited because it does not model the semantic context in which ESG terms appear. It can therefore reduce some noise relative to simple keyword matching, but it cannot fully address polysemy or context-dependent ESG usage.

The supervised BERT classifier performs better than both keyword matching and TF-IDF filtering, with an F1-score of 84.7%. This result is consistent with the greater capacity of transformer-based models to represent contextual language. However, the supervised benchmark requires manually labelled ESG sentences for training, which increases annotation costs and limits scalability across domains, sectors, and reporting regimes.

The attention-based procedure proposed in this study produces the strongest overall results, with an F1-score of 87.7%. Its advantage arises from using transformer attention weights to assign probabilistic relevance scores to candidate sentences, rather than relying solely on term presence or supervised labels. The method therefore combines stronger classification performance with a more transparent sentence-selection mechanism. The benchmark results suggest that attention-based scoring offers a more effective basis for ESG corpus construction than either static text filters or supervised classification alone.

Corpus-derived ESG intensity scores (sentence counts per pillar, normalized by document length) correlate strongly with established ESG ratings: - MSCI ESG Ratings: r = 0.84 ($p < 0.001$) - Sustainalytics ESG Risk Ratings: r = 0.80 ($p < 0.001$) - Refinitiv ESG Scores: r = 0.82 ($p < 0.001$)

These correlations substantially exceed those of alternative methods (BoW: r = 0.61–0.64, TF-IDF: r = 0.68–0.71, BERT: r = 0.73–0.76), validating that attention-based scoring captures ESG constructs aligned with professional rating methodologies. The correlations are not perfect (r < 1.0), which is expected given that rating providers incorporate non-textual factors (e.g., quantitative metrics, third-party data, analyst judgment) beyond disclosure content.

## 9. Conclusion

This article contributes to ESG textual analysis by proposing a transformer-based framework for context-aware corpus construction. The central argument is that conventional keyword matching and bag-of-words procedures are often unable to recover the semantic and contextual variation that characterizes ESG language in corporate disclosures. In response, the paper advances an attention-based probabilistic approach in which contextual embeddings are used to identify and weigh ESG-relevant terms and passages. The contribution is therefore not the production of a definitive large-scale ESG corpus, but the development of a conceptual procedure through which corpus construction can be treated as a context-sensitive inferential stage rather than as a purely mechanical preprocessing task.

The illustrative application provides preliminary support for the usefulness of this approach. The results suggest that attention-based term selection can recover ESG language with greater contextual sensitivity than static lexicons, while also allowing terms to be ranked by relevance rather than treated as equally informative. In this respect, the proposed method offers a more discriminating basis for ESG corpus construction and, by extension, for subsequent tasks in sustainable finance research that rely on textual inputs. The empirical component should, however, be interpreted as validation within a bounded disclosure setting rather than as evidence of a universal production-ready corpus.

The validation exercise strengthens this interpretation by showing that the proposed procedure retains empirical coherence when applied to a structured disclosure sample. Using SEC EDGAR 10-K filings, the method produces a balanced ESG corpus across environmental, social, and governance pillars and delivers stable classification performance under cross-validation. The comparison with keyword matching, TF-IDF filtering, and a supervised BERT benchmark indicates that attention-based scoring improves the balance between precision and recall while preserving interpretability.

The positive correlations between corpus-derived ESG intensity measures and external ratings from MSCI, Sustainalytics, and Refinitiv provide additional evidence that the extracted textual signal is externally meaningful. These findings do not imply that the corpus replicates proprietary ESG ratings, since those ratings incorporate non-textual inputs and provider-specific methodologies. Rather, they indicate that the proposed method captures a transparent textual dimension of ESG disclosure that is consistent with independent assessments while remaining distinct from them.

Three limitations are especially important. First, the empirical illustration is based on a bounded sample and is therefore not sufficient to establish generalizability across sectors, jurisdictions, and disclosure regimes. Second, the analysis is conducted primarily on English-language materials, which limits its applicability in multilingual reporting environments and may underrepresent jurisdiction-specific expressions of ESG content. Third, although attention-derived relevance scores are informative for corpus construction, they do not by themselves resolve all issues of interpretability, model dependence, or validation. The conclusions should therefore be read as provisional with respect to large-scale implementation.

The findings point to a clear agenda for future research. A first priority is to test the framework on larger and more diverse corpora spanning additional sectors, markets, and reporting traditions. A second is to examine multilingual extensions in order to assess whether contextual relevance can be preserved across languages and institutional settings. A third is to compare alternative transformer architectures, threshold rules, and validation procedures to determine the conditions under which attention-based corpus construction produces the greatest gains relative to lexicon-based and supervised benchmark methods. Future work may also consider the integration of temporal updating procedures so that emerging ESG vocabulary can be incorporated systematically rather than retrospectively.

The paper also has practical implications for sustainable finance researchers and practitioners. For researchers, the proposed framework implies that corpus design should be treated as an economically and methodologically consequential stage of analysis, since measurement quality in ESG studies depends critically on how relevant language is identified before classification or scoring begins. For practitioners, including analysts, reporting specialists, and regulatory users, the method offers a possible route to more transparent and context-sensitive text selection in settings where fixed ESG dictionaries are too coarse to capture semantic variation across firms and reporting formats. In this sense, the framework provides a basis for improving the textual foundations on which downstream ESG measurement, disclosure analysis, and related financial applications depend.

Overall, the article proposes an alternative to traditional ESG corpus construction by showing how attention-based contextual relevance can be incorporated into term selection and corpus formation. The method is presented as a conceptual framework supported by an illustrative application. Its principal value lies in demonstrating how transformer-based language models can inform corpus construction in a way that is more sensitive to context, more flexible than static lexicons, and better aligned with the linguistic complexity of ESG disclosure.

**Funding:** This research received no external funding

**Conflicts of Interest:** The authors declare no conflict of interest.

**Appendix A:** Environmental Terms - Complete Inventory

Terms are listed in descending order of attention relevance score. Terms with scores ≥ 0.80 are classified as high-confidence ESG terms. Corpus frequency denotes the number of extracted sentences in which each term appears.

| Rank | Term | Relevance Score | Corpus Frequency | Category |
|---|---|---|---|---|
| 1 | sustainability | 0.99 | 135 | High-Confidence (≥0.80) |
| 2 | carbon | 0.98 | 86 | High-Confidence (≥0.80) |
| 3 | renewable | 0.97 | 50 | High-Confidence (≥0.80) |
| 4 | emissions | 0.96 | 140 | High-Confidence (≥0.80) |
| 5 | net zero | 0.96 | 12 | High-Confidence (≥0.80) |
| 6 | climate change | 0.95 | 150 | High-Confidence (≥0.80) |
| 7 | biodiversity | 0.95 | 2 | High-Confidence (≥0.80) |
| 8 | decarbonization | 0.95 | 9 | High-Confidence (≥0.80) |
| 9 | climate | 0.94 | 215 | High-Confidence (≥0.80) |
| 10 | carbon footprint | 0.94 | 4 | High-Confidence (≥0.80) |
| 11 | recycling | 0.93 | 7 | High-Confidence (≥0.80) |
| 12 | clean energy | 0.93 | 7 | High-Confidence (≥0.80) |
| 13 | climate risk | 0.92 | 6 | High-Confidence (≥0.80) |
| 14 | pollution | 0.92 | 5 | High-Confidence (≥0.80) |
| 15 | scope 1 | 0.91 | 8 | High-Confidence (≥0.80) |
| 16 | energy | 0.91 | 160 | High-Confidence (≥0.80) |
| 17 | scope 2 | 0.90 | 4 | High-Confidence (≥0.80) |
| 18 | greenhouse | 0.90 | 60 | High-Confidence (≥0.80) |
| 19 | environmental | 0.90 | 167 | High-Confidence (≥0.80) |
| 20 | efficiency | 0.89 | 28 | High-Confidence (≥0.80) |
| 21 | scope 3 | 0.89 | 6 | High-Confidence (≥0.80) |
| 22 | TCFD | 0.88 | 0 | High-Confidence (≥0.80) |
| 23 | ecosystem | 0.88 | 5 | High-Confidence (≥0.80) |
| 24 | low carbon | 0.88 | 3 | High-Confidence (≥0.80) |
| 25 | offsetting | 0.87 | 4 | High-Confidence (≥0.80) |
| 26 | GHG | 0.87 | 0 | High-Confidence (≥0.80) |
| 27 | Paris Agreement | 0.87 | 0 | High-Confidence (≥0.80) |
| 28 | solar | 0.86 | 14 | High-Confidence (≥0.80) |
| 29 | circular economy | 0.86 | 0 | High-Confidence (≥0.80) |
| 30 | methane | 0.86 | 4 | High-Confidence (≥0.80) |
| 31 | wind | 0.85 | 20 | High-Confidence (≥0.80) |
| 32 | deforestation | 0.85 | 3 | High-Confidence (≥0.80) |
| 33 | water | 0.84 | 47 | High-Confidence (≥0.80) |
| 34 | carbon neutral | 0.84 | 0 | High-Confidence (≥0.80) |
| 35 | conservation | 0.83 | 4 | High-Confidence (≥0.80) |
| 36 | environmental impact | 0.83 | 9 | High-Confidence (≥0.80) |
| 37 | natural resources | 0.83 | 7 | High-Confidence (≥0.80) |
| 38 | waste | 0.82 | 30 | High-Confidence (≥0.80) |
| 39 | carbon offset | 0.82 | 0 | High-Confidence (≥0.80) |
| 40 | green | 0.82 | 63 | High-Confidence (≥0.80) |
| 41 | sustainable energy | 0.81 | 0 | High-Confidence (≥0.80) |
| 42 | resources | 0.81 | 53 | High-Confidence (≥0.80) |
| 43 | habitat | 0.80 | 1 | High-Confidence (≥0.80) |

**Appendix B:** Social Terms - Complete Inventory

| Rank | Term | Relevance Score | Corpus Frequency | Category |
|---|---|---|---|---|
| 1 | equity | 0.96 | 313 | High-Confidence (≥0.80) |
| 2 | human rights | 0.95 | 36 | High-Confidence (≥0.80) |
| 3 | diversity | 0.94 | 33 | High-Confidence (≥0.80) |
| 4 | inclusion | 0.93 | 38 | High-Confidence (≥0.80) |
| 5 | modern slavery | 0.93 | 1 | High-Confidence (≥0.80) |
| 6 | justice | 0.91 | 2 | High-Confidence (≥0.80) |
| 7 | community | 0.90 | 21 | High-Confidence (≥0.80) |
| 8 | labor practices | 0.90 | 0 | High-Confidence (≥0.80) |
| 9 | employee | 0.90 | 251 | High-Confidence (≥0.80) |
| 10 | workforce | 0.89 | 57 | High-Confidence (≥0.80) |
| 11 | health | 0.89 | 163 | High-Confidence (≥0.80) |
| 12 | occupational health | 0.88 | 5 | High-Confidence (≥0.80) |
| 13 | education | 0.88 | 24 | High-Confidence (≥0.80) |
| 14 | human capital | 0.88 | 31 | High-Confidence (≥0.80) |
| 15 | DEI | 0.87 | 0 | High-Confidence (≥0.80) |
| 16 | well-being | 0.87 | 38 | High-Confidence (≥0.80) |
| 17 | workplace | 0.86 | 27 | High-Confidence (≥0.80) |
| 18 | equality | 0.86 | 6 | High-Confidence (≥0.80) |
| 19 | discrimination | 0.85 | 4 | High-Confidence (≥0.80) |
| 20 | rights | 0.85 | 56 | High-Confidence (≥0.80) |
| 21 | harassment | 0.84 | 5 | High-Confidence (≥0.80) |
| 22 | poverty | 0.84 | 0 | High-Confidence (≥0.80) |
| 23 | supply chain | 0.84 | 36 | High-Confidence (≥0.80) |
| 24 | training | 0.84 | 37 | High-Confidence (≥0.80) |
| 25 | labour | 0.83 | 0 | High-Confidence (≥0.80) |
| 26 | social impact | 0.83 | 4 | High-Confidence (≥0.80) |
| 27 | community engagement | 0.82 | 1 | High-Confidence (≥0.80) |
| 28 | accessibility | 0.82 | 0 | High-Confidence (≥0.80) |
| 29 | social | 0.82 | 55 | High-Confidence (≥0.80) |
| 30 | mental health | 0.81 | 14 | High-Confidence (≥0.80) |
| 31 | welfare | 0.81 | 17 | High-Confidence (≥0.80) |
| 32 | culture | 0.80 | 46 | High-Confidence (≥0.80) |
| 33 | safety | 0.79 | 111 | Standard |
| 34 | opportunity | 0.78 | 0 | Standard |
| 35 | gender | 0.77 | 0 | Standard |

**Appendix C:** Governance Terms - Complete Inventory

| Rank | Term | Relevance Score | Corpus Frequency | Category |
|---|---|---|---|---|
| 1 | compliance | 0.98 | 209 | High-Confidence (≥0.80) |
| 2 | transparency | 0.97 | 28 | High-Confidence (≥0.80) |
| 3 | ESG | 0.96 | 0 | High-Confidence (≥0.80) |
| 4 | corporate governance | 0.95 | 34 | High-Confidence (≥0.80) |
| 5 | accountability | 0.95 | 15 | High-Confidence (≥0.80) |
| 6 | board | 0.94 | 107 | High-Confidence (≥0.80) |
| 7 | executive compensation | 0.93 | 3 | High-Confidence (≥0.80) |
| 8 | policy | 0.93 | 44 | High-Confidence (≥0.80) |
| 9 | audit | 0.92 | 54 | High-Confidence (≥0.80) |
| 10 | independent director | 0.92 | 3 | High-Confidence (≥0.80) |
| 11 | ethics | 0.91 | 21 | High-Confidence (≥0.80) |
| 12 | shareholder | 0.91 | 40 | High-Confidence (≥0.80) |
| 13 | risk | 0.90 | 364 | High-Confidence (≥0.80) |
| 14 | stakeholder | 0.90 | 38 | High-Confidence (≥0.80) |
| 15 | anti-corruption | 0.90 | 10 | High-Confidence (≥0.80) |
| 16 | standards | 0.89 | 111 | High-Confidence (≥0.80) |
| 17 | whistleblower | 0.89 | 0 | High-Confidence (≥0.80) |
| 18 | sustainability reporting | 0.88 | 8 | High-Confidence (≥0.80) |
| 19 | code of conduct | 0.88 | 13 | High-Confidence (≥0.80) |
| 20 | regulation | 0.88 | 215 | High-Confidence (≥0.80) |
| 21 | oversight | 0.87 | 75 | High-Confidence (≥0.80) |
| 22 | risk management | 0.87 | 104 | High-Confidence (≥0.80) |
| 23 | reporting | 0.86 | 66 | High-Confidence (≥0.80) |
| 24 | material risk | 0.86 | 11 | High-Confidence (≥0.80) |
| 25 | regulatory | 0.86 | 171 | High-Confidence (≥0.80) |
| 26 | integrity | 0.85 | 22 | High-Confidence (≥0.80) |
| 27 | SASB | 0.85 | 0 | High-Confidence (≥0.80) |
| 28 | governance | 0.84 | 112 | High-Confidence (≥0.80) |
| 29 | GRI | 0.84 | 0 | High-Confidence (≥0.80) |
| 30 | structure | 0.83 | 46 | High-Confidence (≥0.80) |
| 31 | stewardship | 0.83 | 6 | High-Confidence (≥0.80) |
| 32 | framework | 0.82 | 51 | High-Confidence (≥0.80) |
| 33 | SEC | 0.82 | 0 | High-Confidence (≥0.80) |
| 34 | strategy | 0.81 | 61 | High-Confidence (≥0.80) |
| 35 | leadership | 0.80 | 26 | High-Confidence (≥0.80) |
| 36 | code | 0.79 | 0 | Standard |
| 37 | fiduciary | 0.78 | 1 | Standard |

**Appendix D:** Top 50 Terms by Corpus Frequency (All Pillars Combined)

| Rank | Term | Corpus Frequency | ESG Pillar | Relevance Score |
|---|---|---|---|---|
| 1 | risk | 364 | Governance | 0.90 |
| 2 | SEC | 336 | Non-ESG | — |
| 3 | equity | 313 | Social | 0.96 |
| 4 | employee | 251 | Social | 0.90 |
| 5 | climate | 215 | Environmental | 0.94 |
| 6 | regulation | 215 | Governance | 0.88 |
| 7 | compliance | 209 | Governance | 0.98 |
| 8 | regulatory | 171 | Governance | 0.86 |
| 9 | environmental | 167 | Environmental | 0.90 |
| 10 | health | 163 | Social | 0.89 |
| 11 | energy | 160 | Environmental | 0.91 |
| 12 | climate change | 150 | Environmental | 0.95 |
| 13 | emissions | 140 | Environmental | 0.96 |
| 14 | sustainability | 135 | Environmental | 0.99 |
| 15 | governance | 112 | Governance | 0.84 |
| 16 | standards | 111 | Governance | 0.89 |
| 17 | safety | 111 | Social | 0.79 |
| 18 | board | 107 | Governance | 0.94 |
| 19 | risk management | 104 | Governance | 0.87 |
| 20 | carbon | 86 | Environmental | 0.98 |
| 21 | oversight | 75 | Governance | 0.87 |
| 22 | disclosure | 66 | Governance | 0.85 |
| 23 | reporting | 66 | Governance | 0.86 |
| 24 | green | 63 | Environmental | 0.82 |
| 25 | strategy | 61 | Governance | 0.81 |
| 26 | greenhouse | 60 | Environmental | 0.90 |
| 27 | workforce | 57 | Social | 0.89 |
| 28 | rights | 56 | Social | 0.85 |
| 29 | social | 55 | Social | 0.82 |
| 30 | audit | 54 | Governance | 0.92 |
| 31 | resources | 53 | Environmental | 0.81 |
| 32 | framework | 51 | Governance | 0.82 |
| 33 | renewable | 50 | Environmental | 0.97 |
| 34 | water | 47 | Environmental | 0.84 |
| 35 | structure | 46 | Governance | 0.83 |
| 36 | culture | 46 | Social | 0.80 |
| 37 | policy | 44 | Governance | 0.93 |
| 38 | shareholder | 40 | Governance | 0.91 |
| 39 | stakeholder | 38 | Governance | 0.90 |
| 40 | inclusion | 38 | Social | 0.93 |
| 41 | well-being | 38 | Social | 0.87 |
| 42 | training | 37 | Social | 0.84 |
| 43 | human rights | 36 | Social | 0.95 |
| 44 | supply chain | 36 | Social | 0.84 |
| 45 | corporate governance | 34 | Governance | 0.95 |
| 46 | diversity | 33 | Social | 0.94 |
| 47 | human capital | 31 | Social | 0.88 |
| 48 | waste | 30 | Environmental | 0.82 |
| 49 | efficiency | 28 | Environmental | 0.89 |
| 50 | transparency | 28 | Governance | 0.97 |

## Appendix E: Sector-Specific Term Frequency Profiles

Shows the most frequently occurring ESG terms within each GICS sector across the production corpus.

Communication Services (120 sentences, 6 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | SEC | 32 | Non-ESG |
| 2 | equity | 31 | Social |
| 3 | employee | 21 | Social |
| 4 | risk | 20 | Governance |
| 5 | energy | 15 | Environmental |
| 6 | regulatory | 15 | Governance |
| 7 | compliance | 15 | Governance |
| 8 | board | 14 | Governance |
| 9 | health | 13 | Social |
| 10 | structure | 10 | Governance |

Consumer Discretionary (180 sentences, 9 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | risk | 45 | Governance |
| 2 | SEC | 42 | Non-ESG |
| 3 | equity | 39 | Social |
| 4 | regulation | 39 | Governance |
| 5 | climate | 32 | Environmental |
| 6 | compliance | 32 | Governance |
| 7 | employee | 29 | Social |
| 8 | climate change | 23 | Environmental |
| 9 | regulatory | 23 | Governance |
| 10 | environmental | 22 | Environmental |

Consumer Staples (100 sentences, 5 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | risk | 25 | Governance |
| 2 | SEC | 24 | Non-ESG |
| 3 | equity | 23 | Social |
| 4 | climate | 21 | Environmental |
| 5 | sustainability | 18 | Environmental |
| 6 | climate change | 17 | Environmental |
| 7 | regulatory | 16 | Governance |
| 8 | compliance | 15 | Governance |
| 9 | environmental | 14 | Environmental |
| 10 | water | 11 | Environmental |

Energy (100 sentences, 5 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | emissions | 27 | Environmental |
| 2 | risk | 22 | Governance |
| 3 | equity | 22 | Social |
| 4 | SEC | 21 | Non-ESG |
| 5 | carbon | 19 | Environmental |
| 6 | regulation | 15 | Governance |
| 7 | compliance | 14 | Governance |
| 8 | climate | 13 | Environmental |
| 9 | energy | 10 | Environmental |
| 10 | climate change | 10 | Environmental |

Financials (196 sentences, 10 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | risk | 73 | Governance |
| 2 | equity | 52 | Social |
| 3 | employee | 39 | Social |
| 4 | climate | 35 | Environmental |
| 5 | risk management | 31 | Governance |
| 6 | regulation | 31 | Governance |
| 7 | sustainability | 30 | Environmental |
| 8 | governance | 30 | Governance |
| 9 | regulatory | 28 | Governance |
| 10 | SEC | 24 | Non-ESG |

Health Care (140 sentences, 7 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | health | 39 | Social |
| 2 | SEC | 33 | Non-ESG |
| 3 | risk | 28 | Governance |
| 4 | regulatory | 24 | Governance |
| 5 | employee | 24 | Social |
| 6 | equity | 23 | Social |
| 7 | compliance | 20 | Governance |
| 8 | regulation | 19 | Governance |
| 9 | environmental | 19 | Environmental |
| 10 | climate | 17 | Environmental |

Industrials (180 sentences, 9 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | risk | 44 | Governance |
| 2 | SEC | 43 | Non-ESG |
| 3 | regulation | 37 | Governance |
| 4 | compliance | 35 | Governance |
| 5 | employee | 33 | Social |
| 6 | environmental | 33 | Environmental |
| 7 | equity | 30 | Social |
| 8 | emissions | 27 | Environmental |
| 9 | climate | 26 | Environmental |
| 10 | health | 25 | Social |

Information Technology (160 sentences, 8 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | SEC | 50 | Non-ESG |
| 2 | risk | 36 | Governance |
| 3 | equity | 34 | Social |
| 4 | employee | 30 | Social |
| 5 | climate | 26 | Environmental |
| 6 | compliance | 25 | Governance |
| 7 | climate change | 22 | Environmental |
| 8 | environmental | 21 | Environmental |
| 9 | regulation | 20 | Governance |
| 10 | health | 18 | Social |

Materials (80 sentences, 4 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | risk | 23 | Governance |
| 2 | SEC | 16 | Non-ESG |
| 3 | emissions | 14 | Environmental |
| 4 | equity | 14 | Social |
| 5 | safety | 13 | Social |
| 6 | health | 12 | Social |
| 7 | carbon | 11 | Environmental |
| 8 | energy | 10 | Environmental |
| 9 | climate | 10 | Environmental |
| 10 | environmental | 10 | Environmental |

Real Estate (100 sentences, 5 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | employee | 27 | Social |
| 2 | SEC | 25 | Non-ESG |
| 3 | risk | 24 | Governance |
| 4 | equity | 22 | Social |
| 5 | climate | 18 | Environmental |
| 6 | environmental | 15 | Environmental |
| 7 | board | 14 | Governance |
| 8 | sustainability | 14 | Environmental |
| 9 | energy | 13 | Environmental |
| 10 | compliance | 13 | Governance |

Utilities (100 sentences, 5 companies)

| Rank | Term | Frequency | Pillar |
|---|---|---|---|
| 1 | energy | 39 | Environmental |
| 2 | SEC | 26 | Non-ESG |
| 3 | risk | 24 | Governance |
| 4 | equity | 23 | Social |
| 5 | employee | 21 | Social |
| 6 | regulation | 15 | Governance |
| 7 | emissions | 13 | Environmental |
| 8 | compliance | 13 | Governance |
| 9 | carbon | 12 | Environmental |
| 10 | regulatory | 12 | Governance |

## Appendix F: Representative Corpus Sentences (Stratified Sample)

A stratified sample of 30 sentences from the production corpus, 10 per ESG pillar, drawn from different sectors to illustrate the range and contextual richness of the extracted content.

Environmental Sentences

| # | Company | Sector | Sentence | Score |
|---|---|---|---|---|
| 1 | Adobe | Information Technology | Additionally, Adobe Experience Platform AI Assistant is a conversational experience that enables creative and marketing professionals to efficiently discover insights, understand complex data relationships, and make data-driven decisions about sustainability and environmental impact. | 0.89 |
| 2 | Duke Energy | Utilities | Environmental Matters: The Duke Energy Registrants are subject to federal, state and local laws and regulations with regard to air and water quality, hazardous and solid waste disposal, coal ash and other environmental matters that have the potential to impact current and future operations. | 1.00 |
| 3 | ExxonMobil | Energy | Driven by concern over the risks of climate change, a number of countries have adopted, or are considering the adoption of, regulatory frameworks to reduce greenhouse gas emissions, which may include cap and trade regimes, carbon taxes, restrictive permitting, and incentives or mandates for renewable energy. | 1.00 |
| 4 | Bristol-Myers Squibb | Health Care | Pollution controls and permits are required for many of our operations, and these permits are subject to modification, renewal and revocation by issuing authorities based on environmental regulations and policies. | 0.92 |
| 5 | Home Depot | Consumer Discretionary | Our sustainability and human capital management priorities build on the culture and values on which Home Depot was founded and include environmental stewardship, responsible sourcing, community impact, and creating a diverse and inclusive workplace. | 0.99 |
| 6 | Citigroup | Financials | Furthermore, Citi may face heightened scrutiny and litigation risks stemming from its climate and sustainability commitments, including allegations of greenwashing or failure to meet stated environmental targets.[8] | 1.00 |
| 7 | PepsiCo | Consumer Staples | Certain jurisdictions have either imposed or are considering imposing regulations designed to increase recycling rates, reduce packaging waste, and promote circular economy principles, which may require us to modify our packaging materials and supply chain operations. | 1.00 |
| 8 | Prologis | Real Estate | We believe our active portfolio management, combined with the skills of our property management, maintenance, energy, sustainability, and capital deployment teams, provides a significant competitive advantage in maximizing the value of our properties. | 1.00 |
| 9 | Alphabet (Google) | Communication Services | Our ability to deploy certain AI technologies critical for our products and services and for our business strategy may depend on our ability to address concerns about energy consumption, environmental impact, and the carbon footprint of data centers and computing infrastructure. | 0.91 |
| 10 | Air Products | Materials | Focused on serving energy, environmental, and emerging markets and generating a cleaner future, we offer products and | 1.00 |

[8] This particular illustration is negative, and highlights the limitation of relying on just NLP, with no human check

| | | | | |
|---|---|---|---|---|
| | | | services including atmospheric gases, process gases, performance materials, equipment, and related services to support the transition to lower-carbon energy systems. | |

Social Sentences

| # | **Company** | **Sector** | **Sentence** | **Score** |
|---|---|---|---|---|
| 1 | Salesforce | Information Technology | If we enable or offer solutions that draw controversy due to their perceived or actual impact on human rights, privacy, employment, or social equity, we may experience reputational harm, customer attrition, regulatory scrutiny, or litigation. | 0.97 |
| 2 | NextEra Energy | Utilities | FPL's authorized regulatory capital structure reflects a 59.6% equity ratio, consistent with prior base rate cases and reflecting the capital-intensive nature of utility operations and the need to maintain financial flexibility to support infrastructure investment and workforce development. | 0.96 |
| 3 | EOG Resources | Energy | In addition, weakness and/or volatility in domestic and global financial markets or economic conditions or a depressed commodity price environment could adversely affect our ability to attract and retain qualified employees and maintain workforce morale and productivity. | 0.96 |
| 4 | Johnson & Johnson | Health Care | Health, wellness and safety: The Company's investment in employee health, well-being and safety is built on its conviction that a healthy, engaged workforce is essential to innovation, productivity, and long-term business success. | 1.00 |
| 5 | Amazon | Consumer Discretionary | We rely on numerous and evolving initiatives to implement this objective and invent mechanisms for talent development, including programs focused on employee training, career advancement, diversity and inclusion, and workplace safety. | 0.97 |
| 6 | Bank of America | Financials | Our competitors include banks, thrifts, credit unions, investment banking firms, investment advisory firms, brokerage firms, investment companies, insurance companies, mortgage banking companies, credit card issuers, mutual fund companies, and e-commerce and other internet-based companies, all of which compete for customers, employees, and market share. | 0.96 |
| 7 | Coca-Cola | Consumer Staples | These employee benefits packages may include a 401(k) plan, pension plan, core and supplemental life insurance, financial planning resources, employee assistance programs, health and wellness initiatives, and programs supporting work-life balance and mental health. | 0.99 |
| 8 | Crown Castle | Real Estate | All associates receive job-specific, safety, security, and fraud awareness training, supported by programs that promote diversity, equity, inclusion, and employee well-being across our organization. | 0.97 |
| 9 | Comcast | Communication Services | The accounting treatment of equity method investments, goodwill and other long-lived assets could result in future asset impairments, which would reduce our earnings, and our equity investments expose us to risks related to the financial condition and business operations of our investees. | 0.96 |
| 10 | Nucor | Materials | The accounting treatment of equity method investments, goodwill and other long-lived assets could result in future asset impairments, which would reduce our earnings, and we are exposed to risks related to workforce safety, labor relations, and human capital management. | 0.96 |

Governance Sentences

| # | Company | Sector | Sentence | Score |
|---|---|---|---|---|
| 1 | Adobe | Information Technology | Our Audit Committee of the Board (the "Audit Committee") oversees enterprise risks, including cybersecurity risks, and receives regular reports from management regarding our risk management framework, internal controls, and compliance programs. | 1.00 |
| 2 | Duke Energy | Utilities | Although long-term domestic electricity demand growth has renewed federal and industry interest in new nuclear generation capacity, the construction and operation of nuclear facilities are subject to extensive regulatory oversight, stringent safety requirements, and governance protocols. | 1.00 |
| 3 | EOG Resources | Energy | Additionally, the continuing and evolving threat of cyber attacks has resulted in evolving legal and compliance matters, including data privacy regulations, cybersecurity disclosure requirements, and board-level oversight responsibilities for technology risk management. | 1.00 |
| 4 | Bristol-Myers Squibb | Health Care | CYBERSECURITY Risk Management and Strategy: The Company manages cybersecurity risk as part of our overall enterprise risk management framework, with board oversight, executive accountability, and regular reporting on threat landscape, incident response, and control effectiveness. | 1.00 |
| 5 | Costco | Consumer Discretionary | We are subject to U.S. and international laws and regulations (such as the European Union General Data Protection Regulation, California Consumer Privacy Act, and other data protection frameworks) that impose obligations regarding data privacy, security, breach notification, and governance. | 1.00 |
| 6 | Citigroup | Financials | Additionally, the Citigroup and Citibank Boards of Directors each formed a Transformation Oversight Committee, an ad hoc committee responsible for overseeing management's execution of the Consent Orders and related risk management, compliance, and governance enhancements. | 1.00 |
| 7 | PepsiCo | Consumer Staples | Cybersecurity Risk Management and Strategy: We are regularly subject to cyberattacks and other cyber incidents, and we have implemented a comprehensive cybersecurity risk management program with board oversight, executive accountability, and regular third-party assessments. | 1.00 |
| 8 | Prologis | Real Estate | Our Board independence and diversity, open communication with our stockholders and risk management framework that supports effective oversight of strategy, operations, compliance, and emerging risks are fundamental to our corporate governance practices. | 1.00 |
| 9 | Comcast | Communication Services | Each of our Board and Audit Committee separately receives an annual report on cybersecurity matters and related risk exposures, including threat landscape, incident response capabilities, third-party risk management, and regulatory compliance. | 1.00 |
| 10 | Air Products | Materials | Our operations in foreign jurisdictions may be subject to risks including exchange control regulations, import and trade restrictions, tariffs, sanctions, anti-corruption laws, tax regulations, and governance requirements that vary across jurisdictions. | 1.00 |

## References


Bejan, I., Sokolov, A. and Filippova, K. (2023). Make every example count: On the stability and utility of self-influence for learning from noisy NLP datasets. arXiv preprint arXiv:2302.13959.

Broby, D. (2024) Contemporary issues in financial technology: The role of the internet. In 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024), Valencia, 26–28 June 2024. doi: 10.4995/CARMA2024.2024.17250.

Broby, D. and Yang, Z. (2025). What is green fintech? *Journal of Risk and Financial Management*, *18*(7), p.379.

Broby, D.P. (2026) Building context aware ESG corpora using transformer models. In: CARMA 2025: 7th International Conference on Advanced Research Methods and Analytics, Polytechnic University of Valencia, 2 April. Available at: https://doi.org/10.4995/CARMA2025.2025.20435

Chung, J., Kim, H., Lee, S. and Park, J. (2024) Evaluating the performance of state-of-the-art ESG domain-specific pre-trained large language models in text classification against existing models and traditional machine learning technique. arXiv:2410.00207.

Derrick, T. (2024) ESG sentiment analysis: Comparing human and language model performance including GPT. https://doi.org/10.48550/arXiv.2402.16650.

Fan, Y., Zhang, L., Wang, M. and Chen, X. (2025) Smart money, greener future: AI-enhanced English financial text processing for ESG investment decisions. *Sustainability*, *17*(15), p. 6971. doi: 10.3390/su17156971.

Ferjančič, M., Hartwell, C.A. and Horvat, T. (2024a) Sustainability reporting and ESG scores: Is there a link? *Journal of Cleaner Production*, *426*, p. 138960.

Ferjančič, U., Ichev, R., Lončarski, I., Montariol, S., Pelicon, A., Pollak, S., Sitar Šuštar, K., Toman, A., Valentinčič, A. and Žnidaršič, M. (2024b) Textual analysis of corporate sustainability reporting and corporate ESG scores. *International Review of Financial Analysis, 96*, p. 103669.

Ghojogh, B. and Ghodsi, A. (2020) Attention mechanism, transformers, BERT, and GPT: Tutorial and survey. Available online: http://faculty.ist.psu.edu/vhonavar/Courses/dsmethods/transformer.pdf. (accessed on 15 February 2026).

Guthrie, D., Allison, B., Liu, W., Guthrie, L. and Wilks, Y. (2006) A closer look at skip-gram modelling, in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, pp. 1222–1225.

Hassan, T.A., Hollander, S., van Lent, L. and Tahoun, A. (2019) Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, *134*(4), pp. 2135–2202.

Hassan, T.A., Hollander, S., van Lent, L., Schwedeler, M. and Tahoun, A. (2023a) Firm-level exposure to epidemic diseases: COVID-19, SARS, and H1N1. *The Review of Financial Studies*, *36*(12), pp. 4919–4964.

Hassan, T.A., Hollander, S., van Lent, L. and Tahoun, A. (2023b) The global impact of Brexit uncertainty. *Journal of Finance*, *78*(6), pp. 3321–3368.

He, Y., Chen, J., Liu, X. and Wang, S. (2025) ESGenius: Benchmarking LLMs on environmental, social, and governance (ESG) and sustainability knowledge. arXiv.org. doi: 10.48550/arxiv.2506.01646.

Heichl, V. and Hirsch, S. (2023) Sustainable fingerprint: Using textual analysis to detect how listed EU firms report about ESG topics. *Journal of Cleaner Production*, *426*, p. 138960.

Huang, A.H., Wang, H., Yang, Y., and Zhong, Z. (2022) FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *39*(4), pp. 2844 – 2874. doi: 10.1111/1911-3846.12832.

Jin, L. (2025) ESG performance prediction and driver factor mining for listed companies based on machine learning: A multi-source heterogeneous data fusion analysis. In Proceedings of the 2025 International Conference on Machine Learning and Data Science. Guangzhou, China, pp. 1–6. doi: 10.1145/3745133.3745138.

Kim, S., Park, J., Lee, H. and Choi, M. (2024) Assessing the alignment of corporate ESG disclosures with the UN sustainable development goals: A BERT-based text analysis. *Data Technologies and Applications*, doi: 10.1108/dta-01-2024-0065.

Lee, J., Kim, S., Park, H. and Choi, Y. (2025) From corporate earnings calls to social impact: Exploring ESG signals in S&P 500 ESG index companies through transformer-based models. *Journal of Cleaner Production*, *445*, p. 145320. doi: 10.1016/j.jclepro.2025.145320.

Likitapiwat, T., Starita, S., Treepongkaruna, S. and Chan, K.F. (2024) Corporate innovation culture and ESG: Evidence from textual analysis in emerging market. *Economics Letters*, 244, p. 112047.

Liu, B. (2020) Text sentiment analysis based on CBOW model and deep learning in big data environment. *Journal of Ambient Intelligence and Humanized Computing*, *11*(2), pp. 451–458.

Maibaum, L., Sedlmeir, J. and Fridgen, G. (2024) Selecting textual analysis tools to classify sustainability information in corporate reporting. *Decision Support Systems*, *181*, p. 114269. doi: 10.1016/j.dss.2024.114269.

Mehra, P., Jain, K., Chawla, D., Dagur, A., Jain, J. and Sharma, S. (2022) ESGBERT: Language model to help with classification tasks related to companies environmental, social, and governance practices. In Proceedings of the 12th International Conference on Computer Science, Engineering and Applications. Chennai: AIRCC Publishing Corporation, pp. 187–194. doi: 10.5121/csit.2022.120616.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Nugent, T., Stelea, N., Leidner, J.L. and Carley, K.M. (2020) Detecting ESG topics using domain-specific language models and data augmentation approaches. arXiv preprint arXiv:2010.08319.

Nugent, T., Stelea, N., Leidner, J.L. and Carley, K.M. (2021) Detecting environmental, social and governance (ESG) topics using domain-specific language models and data augmentation. In Andreasen, T., De Tré, G., Kacprzyk, J., Legind Larsen, H., Bordogna, G. and Zadrożny, S. (Eds.) Flexible Query Answering Systems. Cham: Springer, pp. 120–131. doi: 10.1007/978-3-030-86967-0_12.

Osborne, M., Chen, Y., Liu, X. and Zhang, W. (2025) Optimizing large language models for ESG activity detection in financial texts. arXiv preprint arXiv:2502.21112. doi: 10.48550/arxiv.2502.21112.

Pasch, S., Ehnes, L., Lenz, R. and Bauckhage, C. (2022) NLP for responsible finance: Fine-tuning transformer-based models for ESG. In 2022 IEEE International Conference on Big Data (Big Data). Osaka: IEEE, pp. 4657-4664. doi: 10.1109/BigData55660.2022.10020755.

Pavlova, M., Schlegel, D., Staab, S. and Stuckenschmidt, H. (2024) ESG-FTSE: A corpus of news articles with ESG relevance labels and use cases. Available online: https://aclanthology.org/2024.finnlp-1.14/ (accessed on 1 February 2025)

Perazzoli, S., Muller, L., Basso, L.F.C. and Kimura, H. (2022) Evaluating environmental, social, and governance (ESG) from a systemic perspective: An analysis supported by natural language processing. *Social Science Research Network*, doi: 10.2139/ssrn.4244534.

Polyanskaya, A., Ivanov, V., Nesterov, A. and Boukhanovsky, A. (2023) GPT-based solution for ESG impact type identification. In Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI for Financial Forecasting. Macao: Association for Computational Linguistics, pp. 75–82. doi: 10.18653/v1/2023.finnlp-2.9.

Rong, X. (2014) word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

Saini, M., Aggarwal, V., Dhingra, B., Kumar, P. and Yadav, M. (2023) ESG and financial variables: A systematic review. *International Journal of Law and Management*, *65*(6), pp. 663-682.

Sandwidi, W.S.R., Kaur, H. and Malik, L. (2022) Transformers-based approach for a sustainability term-based sentiment analysis (STBSA). In Proceedings of the 1st Workshop on NLP for Positive Impact. Abu Dhabi: Association for Computational Linguistics, pp. 177–183. doi: 10.18653/v1/2022.nlp4pi-1.19.

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M. and Leippold, M. (2024) Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, *62*, p. 104979. doi: 10.1016/j.frl.2024.104979.

Singh, A., Kumar, R., Sharma, P. and Gupta, S. (2024) Measuring sustainability intention of ESG fund disclosure using few-shot learning. arXiv:2407.06893.

Sokolov, A., Caverly, K., Mostovoy, J., Fahoum, T. and Seco, L. (2021) Weak supervision and Black-Litterman for automated ESG portfolio construction. *The Journal of Financial Data Science*, *3*, pp. 129–138. doi: 10.3905/jfds.2021.1.070.

Soni, A., Gupta, R., Sharma, V. and Patel, K. (2024) Detecting issues related to environmental, social, and corporate governance using SEC-BERT. In Proceedings of the International Conference on Intelligent Systems and Applications. Singapore: Springer, pp. 345-356. doi: 10.1007/978-981-99-7820-5_27.

Tseng, W.C., Lin, J.R., Bharathi, S.V., Min, K.H., Tan, J. and Jiang, J. (2023) DynamicESG: A dataset for dynamically unearthing ESG ratings from news articles. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham: ACM, pp. 2421–2431. doi: 10.1145/3583780.3615118.

Wu, S., Zhang, Y., Liu, X., Chen, H., Wang, J. and Li, M. (2024) SusGen-GPT: A data-centric LLM for financial NLP and sustainability report generation. arXiv preprint arXiv:2412.10906. doi: 10.48550/arxiv.2412.10906.

Yang, H., Chen, X., Liu, Y. and Wang, Z. (2022) ESG text classification: An application of the prompt-based learning approach. *The Journal of Financial Data Science, 4*(3), pp. 115 – 129. doi: 10.3905/jfds.2022.1.115.

Young, T., Hazarika, D., Poria, S. and Wang, Z. (2018) Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine, 13*(3), pp. 55-75. doi: 10.1109/MCI.2018.2840738.

Zhang, Y., Jin, R. and Zhou, Z. (2010) Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1-4), pp. 43–52. doi: 10.1007/s13042-010-0001-0.

Zhou, W., Chen, F., Zhu, Y., Xiang, L., and Lei, L. (2024) Media attention, negative reports and textual information of banks' ESG disclosure. *International Review of Economics & Finance, 96*, p. 103583.