machine learning & knowledge extraction

MDPI

Article

Mapping ESG Trends by Distant Supervision of Neural Language Models

Natraj Raman [1,*,†], Grace Bang [2,†] and Armineh Nourbakhsh [3,†]

1 J.P. Morgan AI Research, London E14 5JP, UK
2 Bloomberg LP, New York, NY 10017, USA; gbang3@bloomberg.net
3 J.P. Morgan AI Research, New York, NY 10179, USA; armineh.nourbakhsh@jpmchase.com
* Correspondence: natraj.raman@jpmorgan.com
† Work done when authors were at S&P Global.

Received: 9 August 2020; Accepted: 6 October 2020; Published: 21 October 2020

Abstract: The integration of Environmental, Social and Governance (ESG) considerations into business decisions and investment strategies have accelerated over the past few years. It is important to quantify the extent to which ESG-related conversations are carried out by companies so that their impact on business operations can be objectively assessed. However, profiling ESG language is challenging due to its multi-faceted nature and the lack of supervised datasets. This research study aims to detect historical trends in ESG discussions by analyzing the transcripts of corporate earning calls. The proposed solution exploits recent advances in neural language modeling to understand the linguistic structure in ESG discourse. In detail, firstly we develop a classification model that categorizes the relevance of a text sentence to ESG. A pre-trained language model is fine-tuned on a small corporate sustainability reports dataset for this purpose. The semantic knowledge encoded in this classification model is then leveraged by applying it to the sentences in the conference transcripts using a novel distant-supervision approach. Extensive empirical evaluations against various pretraining techniques demonstrate the efficacy of the proposed transfer learning framework. Our analysis indicates that in the last 5 years, nearly 15% of the discussions during earnings calls pertained to ESG, implying that ESG factors are integral to business strategy.

Keywords: NLP; pre-trained embeddings; transfer learning; ESG; sustainability reports

1. Introduction

Environmental, Social and Governance (ESG) practices define a company's strategy, business model and conduct, as they relate to sustainability. The three aspects of ESG practices encapsulate a wide range of concepts, including environmental factors, such as renewable energy and waste management, social factors, such as community engagement and labor management, and governance factors, such as business ethics and risk management.

ESG factors have been the topic of a growing body of debates and studies around company performance [1,2], productivity [3], industry trends [4], and impact on sustainable investment strategies [5]. This growing attention has also manifested itself in the emergence and popularity of sustainability reports published by companies, as well as various indices and ratings provided by third-party authorities, such as MSCI's ESG Ratings (https://www.msci.com/esg-ratings) and S&P Global's Green Evaluation (https://www.spglobal.com/ratings/en/products-benefits/products/green-evaluations). Given the overwhelming evidence of ESG factors impacting financial performance [2], we seek to examine how companies have historically considered ESG factors and how they continue to view and treat ESG as part of their business operations. Gaining insight into ESG discourse is essential in order to objectively assess the importance attributed to sustainable business practices over time and our study makes an important contribution in this front.

Existing research on ESG language mainly relies on lexical matches against query words, or simple statistical analyses of term frequencies, which are prone to errors [6,7]. In contrast, we aim to gain semantic understanding of the text by using distributed representations of words in a vector space. These representations encode syntactic and semantic relationships in such a way that allows us to profile ESG language in detail, and scale the study across multiple domains. Furthermore, they allow us to visually explain the relationship between specific expressions and ESG factors without the need for any training or supervision.

Our methodology is carried out in two steps. First, we use these semantic representations to identify ESG discussions in the transcripts of corporate earnings calls. Next, we analyze their trends over time and across sectors and market-cap designations. Earnings calls are selected as a source because of their unique value in bringing together the company's self-reported performance and questions posed by researchers and analysts, which has made them a popular source for studies that aim to mine market-performance signals from text [8–10].

To maintain a granular lens, given a sentence in the transcript of an earnings call, our model is to identify whether the sentence is irrelevant, somewhat relevant, or highly relevant to ESG. By classifying each sentence in this manner, we address the following questions:

- How prominent are ESG-related discussions within earnings call transcripts? How does that differ across different industries and market capitalization designations?
- How has the volume of discussions about ESG evolved over time?
- Have the Environmental, Social, and Governance factors all undergone similar trends, or have certain aspects gained in popularity compared to others?
- When it comes to evaluating the performance of a business, how does ESG fare compared to financial metrics? Do analysts and researchers pay as much attention to ESG factors as they do to financial outcomes?

In order to find answers to the above questions, we need a robust method of classifying sentences into the irrelevant, somewhat relevant, or highly relevant categories. A supervised approach would require significant effort devoted to creating manual labels, and controlling for inter-annotator agreement could further undermine the availability of labeled data. As a result, we propose a distant-learning approach that alleviates this problem. Starting with a pre-trained neural language model, we first fine-tune it on a corpus that includes highly ESG-related language. We then apply the language model to transcripts and measure its performance on a held-out set. Figure 1 provides an overview of our approach.

We target Corporate Sustainability Reports (CSRs) as our distant-training corpus. CSRs are unregulated reports published by companies voluntarily, where ESG factors are discussed in detail and the company's commitment to sustainable practices is corroborated through past activities and future plans. The unregulated and sometimes promotional nature of these reports makes them unreliable candidates for trend analysis, but a potent source for modeling ESG-like language.

We hope that our approach provides inspiration for other studies in applied NLP domains that suffer from scarcity of labels. For example the language provided in technical reports, regulatory disclosures or other specialized domains may be used to profile related topics in other corpora.

During our research, we discovered that, despite the steady growth of ESG-related conversations in earnings calls, the three main aspects of Environmental, Social and Governance factors have not trended similarly. Mainly, social factors have attracted more attention over the past 5 years, as the focus on environmental factors has plateaued. Additionally, we observe that ESG factors are still treated as secondary when company financial performance deteriorates. Section 5 describes these findings in detail.

Our technical and analytical contributions include:

- A method for profiling ESG-related language in business corpora using transfer learning from contextual embeddings.

- A novel method for distant supervision of a classifier that uses CSR corpus to learn ESG-related language and encodes that semantic knowledge to identify ESG language in transcripts of earnings calls.
- A proposed set of standard and augmented features for detecting ESG mentions.
- A comparative study of SOTA language-models and their efficiency in profiling language in business corpora given limited labeled data.
- A deep analysis of historical trends in ESG discussions during earnings calls, divided by factors, industry, and market cap.

The following sections lay out our methodology, data, experiments, and findings.

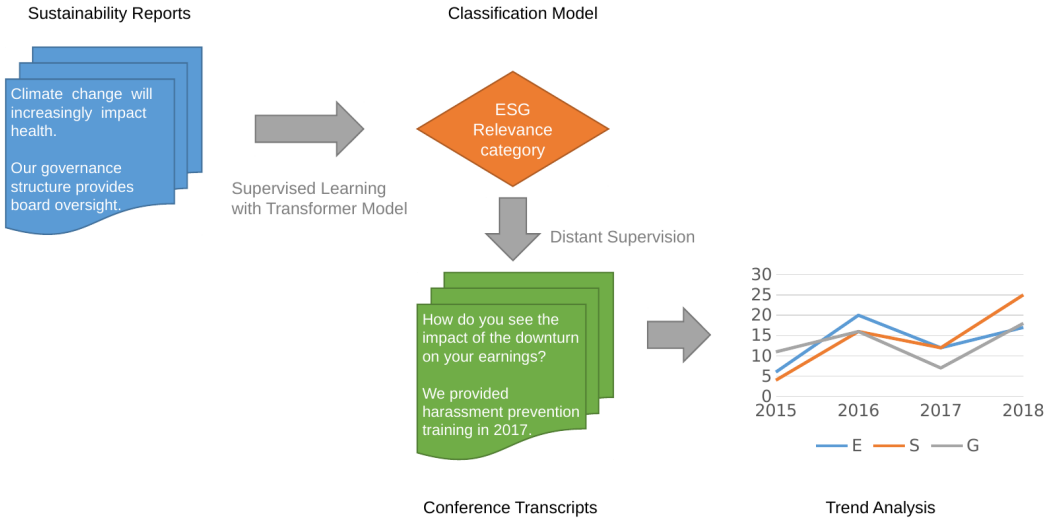

Figure 1. Approach Overview. A classification model is learned from corporate sustainability reports and applied to earnings conference call transcripts for detecting ESG trends.

2. Related Work

Sustainability reporting began in the 1980s but attempts at extracting, analyzing, or predicting ESG-related performance using automated methods have a more recent history. The authors in [11] discus the impact of using machine learning methods on CSR reports in detail. The authors in [6] apply a variety of classifiers including a feed-forward neural network to CSR reports in order to predict the sustainability score. They also apply an ontology extraction method to enhance the performance of their classifiers, but do not explore the generalizability of their models to out-of-domain corpora as we do here.

Studies focused on mining insights from earnings calls have largely focused on drivers of business performance, such as investor sentiment, transparency, and risk-semantics. For example, in [9] the language content in earnings conference calls is examined to predict changes in price recommendation provided by financial analysts. The authors in [8] study the correlation between earnings calls and stock price volatility in order to forecast future financial risk. This paper differs from the above works in two key aspects: (a) a complementary combination of sustainability reports and earnings calls is used here; (b) rather than financial performance, the focus here is on ESG analytics.

Data sources, such as news, social media and stock exchange filings have also been used in the ESG context. The authors of [12] use adverse media coverage to predict companies that are likely to be blacklisted in sustainable investment practices. In [13], real-time event detection techniques is paired with sentiment classification in order to identify Twitter controversies around a given company and associate the controversies with the stock performance of the company. A manually curated dataset is used in [14] to illustrate the theory that companies that score high on ESG create long-term shareholder value. While these diverse data sources carry interesting information, we exclusively focus on the narrative by the company instead of external opinion.

There have been several studies that explore the predictive power of ESG related information on financial markets. CSR scores are used in [15], and the authors find that firms with better scores exhibit cheaper equity financing. The author of [16] analyze whether there is a relationship between companies' stock market performances and their ESG scores. The author proves that ESG metrics indeed predict stock returns in a global investment universe. The recent study in [17] considers the costs originating due to extreme weather events and concludes that risks due to climate change have huge financial implications on investors' portfolio companies. Instead of teh traditional price or return time-series, [18] focuses on the ESG events in financial news in order to forecast equity realized volatility in stock markets. Our work differs from all the above studies in its objective to detect trends in ESG discussions rather than predict market behavior.

Keyword based filtering has been popular in analyzing ESG characteristics of companies [7]. However, relying on lexical matches of query words or simple frequency statistics is prone to errors and successful text analysis require gaining a semantic understanding of the text. Distributed representations of words in a vector space can capture syntactic and semantic word relationships. These representations [19] are learned from large text corpora and the computed word embeddings are often used as features for a downstream task. The authors in [20] extract sentences from sustainability reports by matching the embedding vectors of words in a sentence to a taxonomy of ESG indicators. The sentences are then ranked by their potential importance based on the presence of quantitative indicators and are used as a data-digitization methods for decision-support mechanisms.

Contextualized word embeddings that are trained on Transformer architectures [21] have improved the state-of-the-art for many NLP tasks [22]. Unlike traditional word vectors, which encode only some semantic information, these representations model a rich hierarchy of contextual information and the structure of the language in greater detail. In contrast to RNN- and CNN-based language models, the attention mechanism used in Transformers can directly reference a large number of words in a sentence and thus account for long range dependencies. Transformers are also computationally efficient with significantly more parallelization. We use fine-tuned contextual word embeddings pre-trained using a number of these recent Transformer models, such as BERT [23], XLNet [24] and RoBERTa [25]. We believe our study is the first to exploit these state-of-the-art language representation models to profile the linguistic patterns in ESG discourse.

3. Methodology

3.1. Method Overview

As a starting point, we developed a model that scores the sentences in a CSR document based on its relevance to ESG. The model was framed as a supervised classification task that categorizes text into one of below classes:

- Irrelevant: This applies to sentences that have no ESG content in them, such as "In these last few decades, we've seen technology do amazing things, transform experiences, and improve the lives of millions" [26].
- Quasi-relevant: This applies to sentences that discuss ESG factors in a generic context without explicitly relating them to a concrete business practice. For example, "As data becomes more accessible and meaningful, we must ensure that availability does not come at the expense of privacy" [26].
- Relevant: This applies to specific expressions of ESG-related plans, programs, policies and guidelines that are practices or devised by the target company—e.g., "We also provide ad hoc training to employees on topics such as correct waste sorting on campus" [26].

This level of differentiation allows us to distinguish sentences with concrete evidence of ESG practices from those with a more promotional nature.

The labeled examples used for training the classifier must be annotated manually. In order to make this process efficient, we introduce an unsupervised mechanism that identifies the category labels

automatically. The labels suggested by this unsupervised model are corrected by the human annotator, thereby ensuring that accurate categories are used during training. This procedure is elaborated in Section 3.2.

The classifier is trained by fine-tuning the embeddings provided by a state-of-the-art Transformer model [23]. Specifically, the parameter weights of the deep learning model are adjusted by introducing a custom output layer that uses a loss function based on labeled examples of above categories. Sections 3.3–3.5 discuss the classification model in detail.

The model trained on CSR language is validated on a small held-out set of earnings call transcripts in order to assess its distant supervision performance. The validated model is then applied to a larger corpus of earnings call transcripts to study historical and emerging trends in ESG discussions.

3.2. Labeled Data Curation

Many companies periodically publish the economic, environmental, social and governance impacts caused by their day-to-day activities in CSRs. The reports do not follow any well-defined format and vary in length, style, layout, readability and reporting of performance metrics. We considered about 125 of such publicly available reports and created a small supervised dataset of sentences to train our classifier model.

A naive sampling of the sentences from these reports for annotation would result in an imbalanced dataset. Hence we applied an initial unsupervised mechanism to identify whether a sentence is Relevant, Quasi-Relevant or Irrelevant to ESG factors. Specifically, we performed a vector-space matching of the word embeddings in a sentence with a set of commonly used Key Performance Indicators [27] in order to pre-assess the relevance of a sentence. These indicators are often used to evaluate the sustainability performance of a company and their presence in a sentence is an important measure of relevance. Additionally, the indicators allow for determining the ESG factor discussed in a sentence. Table 5 lists these indicators.

By using pre-trained word embedding vectors to compute the similarity between a sentence and an indicator, we avoided the need to exhaustively specify keywords. For example, in order to capture the indicator *compensation*, it requires the specification of several keywords, such as pay, salary and bonus. However, the pre-trained embedding vectors of these keywords will already be close to the indicator in embedding space and therefore need not be explicitly defined. We employed the cosine similarity metric and used the maximum similarity value of a sentence across different indicators as the aggregated similarity score.

Formally, let $\kappa_1, \ldots, \kappa_k, \ldots, \kappa_K$ be a set of indicators. Each indicator may contain one or more words. Let $\mathbf{e}$ be a function that takes a set of words and returns its corresponding embedding ϕ, as follows:

$$\mathbf{e} : \kappa_k \rightarrow \phi_k, \qquad \phi_k \in \mathbb{R}^{D_e} \tag{1}$$

where D_e is the embedding size. We averaged the Glove [19] vectors corresponding to the words to obtain ϕ. Similarly, let ϕ^* be the embedding vectors obtained for a given sentence. The cosine similarity between an indicator k and the sentence is calculated as

$$sim_k = \frac{\phi_k \,.\, \phi^*}{||\phi_k||\,||\phi^*||} \tag{2}$$

The relevance score r^* of the sentence across different indicators is computed using a max function as follows:

$$r^* = \max_k sim_k \qquad \forall k = 1 \ldots K \tag{3}$$

This score is used as a surrogate measure of relevance for a sentence. The dataset to be used for annotation is now created through a biased sampling of these relevance scores. This procedure ensures that the human annotator is presented with a balanced set of labels corresponding to all the three categories.

3.3. Supervised Text Classification

Our focus here was a sentence level classification task in which given a sequence of words, we aimed to analyze the sequence holistically and assign a class label to it. The preferred solution to this problem was to use word representations from a neural language model that is pre-trained on a large scale unlabeled text corpus with a general-purpose training objective. Although these word representations encode many linguistic regularities and patterns, they do not capture task specific information since they are learned in a generalized context. Hence, it is essential to fine-tune the pre-trained model parameters on the downstream classification task.

The pre-trained models mainly differ in their neural network architecture and training objective. We restrict our discussions to the Transformer [21] architecture that is currently the de facto modeling choice for sequence transduction problems. In the following section, we detail this architecture and present a unified notation for the various models. Figure 2 provides an overview of our model.

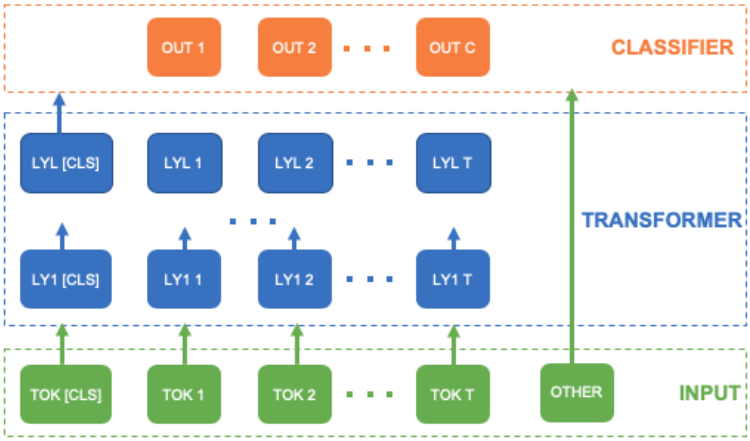

Figure 2. Fine-tune model architecture. Input tokens are converted to contextual representations using a transformer model. The aggregate representation from the last layer L and other input features are used for classification.

3.4. Pre-Training

Let $\mathcal{U}$ be a text corpora with a large number of unlabeled sentences. Let an input sentence $\mathbf{x}$ consist of a sequence of T tokens $(x_1, \ldots, x_T)$ coming from a fixed vocabulary $\mathcal{V}$ of size $|\mathcal{V}| = V$. The tokens are derived after performing preprocessing steps, such as lowercasing, tokenization and out-of-vocabulary resolution on the input sentence. It is common to inject position information about the tokens into the input. Without loss of generality, all the sentences are assumed to be of same length, with shorter sentences being padded. Let x_0 be a special token $[CLS]$ that is augmented to the beginning of a sequence and is used as an aggregate representation for the entire sequence.

The language model aims to learn a function $f : \mathbf{x} \rightarrow \mathbf{z}$ based on $\mathcal{U}$, where $\mathbf{z} = (z_0, \ldots, z_T)$ and $z_t \in \mathbb{R}^D$ is the D dimensional contextualized vector representation of a token. The function f is parameterized as a neural Transformer model with parameters θ.

The joint probabilities $p(x_0, ..., x_T)$ over the token sequence factorizes into a product of conditional probabilities

$$p(\mathbf{x}) = \prod_t p(x_t | x_{\setminus t}; \theta), \tag{4}$$

where $x_{\setminus t} = (x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T)$. Let us introduce a function $\phi(\mathbf{x})$ that returns a set of tuples of the form $\{(x_{t'}, x_{\setminus t'})\}_{t'=1}^{K}$ in order to flexibly define the set of tokens used to model the bi-directional context. For auto-encoding models, such as BERT [23], $x_{t'}$ is a masked token and $x_{\setminus t'}$ are the corresponding unmasked tokens. In the case of auto-regressive models, such as XLNet [24], the tuples are determined through a permutation operation. Using this generalized notation and the factorization in Equation (4), the learning objective function is framed as

$$\max_{\theta} \sum_{t'} \log p(x_{t'} | x_{\setminus t'}); \theta). \tag{5}$$

The conditional probabilities are modeled using a softmax function and the pre-training objective is

$$\max_{\theta} \sum_{t'} \log softmax \left(\mathbf{H}(x_{\setminus t'}); \theta)^T z_{t'} \right), \tag{6}$$

where $\mathbf{H}$ is the context representation produced by the Transformer.

The Transformer uses L identical layers at different levels of abstraction to encode the input into contextual representations. Let $H^0 = (x_0, \ldots, x_T)$ be the input layer and $H^l = (h_0^l, \ldots, h_T^l)$ be the lth hidden layer. Each layer H^l consists of two sublayers: an attention sublayer A^l and a simple feed-forward network F^l. Further, the sub-layers use a residual connection followed by layer normalization.

The attention mechanism in Transformers allow for the modeling of dependencies between arbitrary positions of the input. Specifically, a self-attention mechanism that relates different positions of a single sequence is employed to associate the relevance of one token to another. In fact, a number of such self-attentions are used to attend to information from different representation subspaces at different positions. Formally, let there be M different self-attentions $\{a_m^l\}_{m=1}^{M}$. The attention sublayer A^l is computed as:

$$\begin{gathered} Q = H^{l-1} W_{l,m}^Q, \qquad K = H^{l-1} W_{l,m}^K, \qquad V = H^{l-1} W_{l,m}^V \\ a_m^l = softmax \left(\frac{QK^T}{\sqrt{s}} \right) V, m = 1 \ldots M \\ A^l = (a_1^l \oplus \ldots \oplus a_m^l) W_l^O \end{gathered} \tag{7}$$

where $W_{l,m}^Q, W_{l,m}^K, W_{l,m}^V, W_l^O$ are parameter matrices, $\oplus$ is a concatenate operation and s is a scaling factor. The output of the last layer H^L is used as the context representation in Equation (6).

3.5. Fine-Tuning

Let $\mathcal{S}$ be a supervised dataset with training set pairs of the form $(\mathbf{x}, y)$, where $\mathbf{x}$ is an input sentence comprised of a sequence of tokens as outlined in the previous section and $y \in 1, \ldots, C$ is the class label associated with a sentence.

The word representations learned using the pre-training objective function in Equation (6) models the data distribution in the unlabeled text corpora $\mathcal{U}$. In order to adapt this model to the classification task, it must be fine-tuned using a classification specific objective function based on the supervised training data in $\mathcal{S}$. This is achieved by initializing the Transformer with model parameters θ, and feeding the sentences from $\mathcal{S}$ as inputs into the Transformer. Additionally, a new output layer that takes z_0 (i.e., the representation of special [CLS] token) as input is added. Using cross-entropy loss, the fine-tuning objective function is now written as

$$\max_{\theta^f} \sum_{c} \mathbb{I}(y = c) \log p(y = c), \tag{8}$$

where $\mathbb{I}$ is an indicator function, θ^f is the fine-tuned Transformer parameters and

$$p(y = c) = softmax\left(W^{Y^T}(z_0 \oplus \chi)\right), \tag{9}$$

is the probability of a label given the weight matrix W^Y. Note that it is possible to concatenate the output from the Transformer model with other augmented input features χ and use it as an input to the classification layer. In this work, fine-tuning is performed on a labeled subset of CSRs to determine the ESG-relevance category of a sentence.

4. Results

We first layout the experiment setup and then present the evaluation results for the classifier on the Corporate Sustainability Report (CSR) dataset. Empirical evidence for the classifier's ability to generalize on the Earnings Call Transcript dataset is finally discussed.

4.1. Experiment Setup

As detailed in Section 3, the Transformer architecture can be fine-tuned for a downstream application by the addition of task specific output layers. We used the Tensorflow implementation of Transformers library [28] to fine-tune our classifier on the CSR dataset. The classification models were trained for a maximum of 15 epochs with the training stopping when the F1-Score did not improve for 3 consecutive epochs on a holdout validation set. The model parameters included a learning rate of 3×10^{-5} and an epsilon of 1×10^{-8} for Adam optimizer, a batch size of 32, a dropout probability of 0.1 and a maximum sequence length of 128. These parameters were carefully selected based on a coarse to fine grid search. The training was performed on a single instance of Tesla V100-SXM2-16GB GPU.

For comparison, we evaluated two baseline models. In the first baseline, we develop a non-Transformer model which simply uses the 300 dimensional vector representation of a word [19]. Here, the embeddings do not account for the context of a word in a sentence. In the second baseline, we used a BERT [23] Transformer model with fine-tuning being disabled. More precisely, the second baseline incorporated contextualized embeddings but all the layers except the output layer were frozen so that the Transformer model weights did not update during training.

We also compared against various state-of-the-art general purpose pre-trained language models in order to benchmark their ability to profile the language used in CSRs. The BERT model [23] used masked language modeling with an auto-encoder architecture. The other models evaluated for comparison built on top of BERT. DistilBERT [29] leveraged knowledge distillation to reduce the size of BERT model, while RoBERTa [25] improved the training recipe for BERT by carefully tuning the hyper-parameters. XLNet [24] addressed the pre-train-fine-tune discrepancy in BERT by proposing an auto-regressive training objective. All these models differed in the dataset used for training, the number of parameters and the pre-trained weights. The base version of these models, which does not require significant GPU memory, is evaluated here.

4.2. Evaluation on CSR Dataset

The distribution of sentences across the Irrelevant, Quasi-Relevant and Relevant categories in the CSR dataset is shown in Table 1. A few examples of sentences from these categories, as annotated by subject matter experts, are listed in Table 2.

Table 3 summarizes the classification results on the CSR dataset. The results are reported for 5-fold cross-validation with 80% of the dataset used for training, 20% for testing and a further 20% of the training set serving as a validation set. Following standard procedure, we adopted F1-Score, which is a harmonic mean of the precision and recall, as a measure of performance. The need to adopt contextual word representations is confirmed by the poor performance of the first baseline model. The fine-tuned models significantly outperform the non-fine-tuned baseline model, thereby highlighting the importance of updating the pre-trained weights. Although the

classification performance of all the fine-tuned models remains similar, the original BERT model with uncased text has the best F1-Score of 78.3%. We hypothesize that the small training set size is a likely factor in the lack of improved results for more sophisticated models, such as RoBERTa and XLNet.

Table 1. Dataset Summary.

CSR	Sentences	4935
	Irrelevant	2135
	Quasi-Relevant	1678
	Relevant	1122
Earnings Call Transcript	Companies	499
	Years	9
	Blocks	1.26 (million)
	Sentences	8.49 (million)

Table 2. Manual Annotation - sample sentences and their corresponding relevance.

Relevance	**Sentence**
Irrelevant	Climate change will increasingly impact health and well-being globally.
Irrelevant	At the close of FY15, 24,465 employees had completed the training.
Quasi-Relevant	Industry norms and policies for managing emerging issues related to digital rights and human rights are still being established.
Quasi-Relevant	Our governance structure provides board and management oversight of our risk processes and mitigation plans.
Relevant	In calendar year 2017, 99% of the electricity used at our colocation facilities was powered by renewable energy.
Relevant	The governance of CBRE is supervised by an 11-member Board of Directors, ten of whom are deemed independent.

Table 3. Classifier prediction results for CSRs.

Model	**Parameters**	**Accuracy**	**F1-Score**
Non-contextual Embeddings	400	43.9	26.0
No fine-tuning	2K	57.0	49.8
XLNet [24]	118M	75.8	75.9
RoBERTa [25]	125M	77.8	78.0
DistilBERT [29]	65M	74.6	78.0
BERT Cased [23]	108M	77.8	78.1
BERT Uncased [23]	110M	**78.1**	**78.3**

We also consider fine-tuning the language model itself on the CSR corpus. In this case, the pre-trained weights for the classifier was initialized from the fine-tuned language model rather than the general purpose model. However, we did not find any evidence that this improved the classifier performance. This can again be attributed to the training size: our fine-tuning used a corpus of 20 MB text while the general purpose models were trained on a corpus of 13 GB text. We also experimented with augmenting the input to the classification layer with additional features extracted from a sentence. In particular, a one-hot encoded vector of the performance indicators and the amount of quantitative and qualitative information in a sentence were used. This marginally improved the classification performance. The prediction results for these extensions to the original BERT model reported in Table 4.

Table 4. Classifier prediction results for extensions to BERT model.

Extension	Accuracy	F1-Score
Fine-tune Language Model	77.7	78.1
Augmented Features	**78.2**	**78.4**
Augmented Features + fine-tune LM	78.0	**78.4**

4.3. Distant Supervision of Earnings Call Transcripts Dataset

The financial results and business decisions of companies are shared with investors on a quarterly basis in earnings conference calls. We construct our Earnings Call Transcript dataset from CapitalIQ (http://www.capitaliq.com) with the dataset containing historical reading transcripts of earnings calls of S&P 500 companies. These companies span a wide range of industrial sectors, geographic locations and market size and we focused on the period between 2010 and 2018. Each transcript block in this dataset corresponds to the text spoken by a speaker and we segment the sentences in these blocks and use them in our analysis. Table 1 provides an overview of this dataset's characteristics.

We first evaluated whether, without any updates to the parameters, the classifier that was trained on the CSR dataset could successfully perform predictions on the Earnings Call Transcript dataset. For this purpose, we constructed a small validation set by sampling 500 sentences from this dataset and manually annotated the relevance category. The classifier achieved an F1-Score of **80.3%** and the confusion matrix is shown in Figure 3. It is noteworthy that even though the classifier did not include any sentences from the Earnings Call Transcript dataset during training, it still performed well on this dataset. This demonstrates its ability to generalize and confirms a successful transfer learning framework. We observed that the classifier particularly struggled with sentences that were posed as a question. The CSR dataset does not contain any questions as such, while the question and answer format is typical in earnings calls. We run predictions over the 8 million sentences in this dataset and elaborate our findings in the subsequent section.

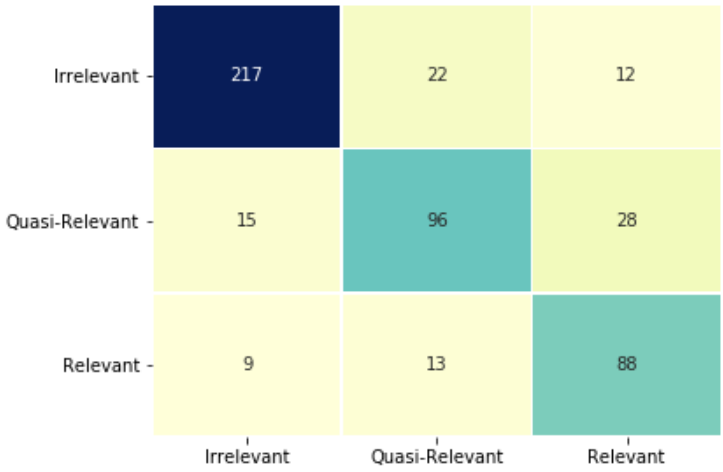

Figure 3. Confusion matrix for classifier predictions on earnings call transcripts (F1 = 80.3).

4.4. Qualitiative Analysis

The attention mechanism used in Transformers improve model interpretability by providing insights into the reasoning behind predictions. The attention weights assigned to each word (token) indirectly quantify the contribution of a word to the classifier decision and, by examining the attention pattern of all the words, the model behavior can be understood in a better manner. In order to compute a unified attention score for a word, we pooled the attention weights of a word across all the different attention heads from the last layer of the Transformer and summed up the weights. The words being attended to corresponded to the special token [CLS].

A visualization of these attention scores for sentences that were classified as Relevant is provided in Figure 4. The highlighted words in this figure possess high attention scores, with a darker color implying more weight. The high importance to words, such as harassment, contributions and philanthropic, reveal that these words help in differentiating between sentences that are not related to ESG. Further, the model's focus on words such as implemented, helps and hsupports suggests that it associates actions taken by an organization as a measure of relevance.

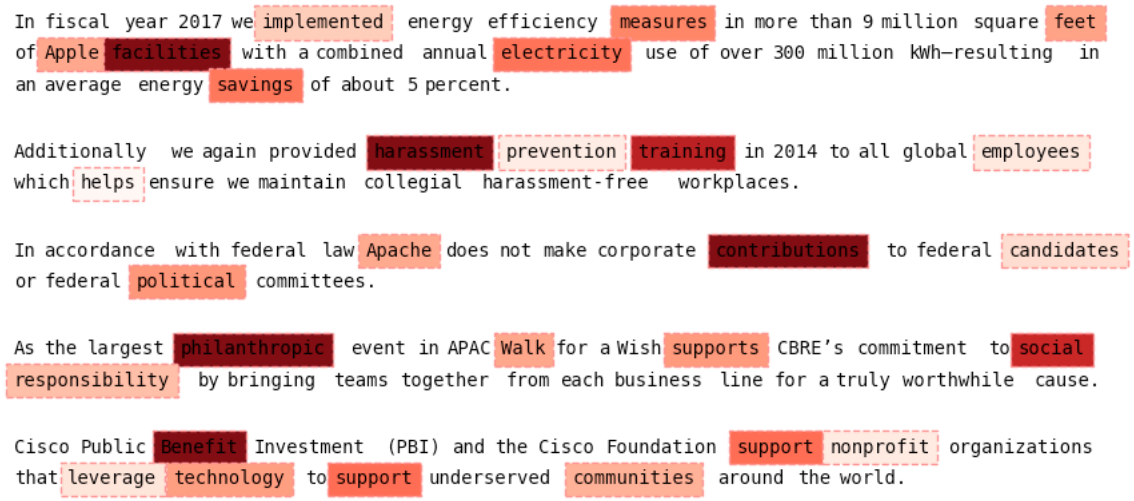

Figure 4. Attention Visualization. The classifier pays more attention to the highlighted words. Darker color implies higher attention weights.

5. Key Findings

The data indicate that, over the past five years, companies have increased their focus on ESG factors and such factors are becoming an integrated aspect of companies' business strategies. Since 2013, about 13% to 18% of the discussions during earnings calls pertained to ESG considerations. Specially, in the past two years, the percentage of discussions on ESG has continuously experienced mid-single-digit to double-digit growth.

Historically, the social component was discussed the least often, while the interest in the environmental component has been growing steadily. However, from 2017 onward, the amount of attention given to the social factor has surpassed that provided to the other two components, as shown in Figure 5. Coincidentally, we observe that factors such as Discrimination and Labor Relations experienced as much as a 30+% increase in attention over the past two years. Meanwhile, factors such as Renewable Energy and Energy Efficiency, which had historically taken up as much as 28% of all discussions related to ESG on earnings calls, show steady decline in mentions from 2016 onward.

Table 5 lists the ESG factors and their usage rates over the past decade. The growth rate trend of various ESG factors for the past 5 years is plotted in Figure 6. Each point in the chart is an ESG factor and the size of the point depends on the number of mentions in the transcripts for the factor corresponding to this point. The points are color coded by their E or S or G category. Several key insights can be gathered from this figure—for example, mentions about Data Privacy has increased a lot recently, while discussion about Renewable Energy have plateaued.

Keen ESG Focus in Business-to-Consumer Industries. Retail investors and consumers are spending more dollars and paying more attention to supporting and investing in sustainable and socially progressive companies [30]. Thus, Figure 7 shows that many companies that operate primarily in the Business-to-Consumer sectors exhibit higher rates of increase in company-specific mentions of ESG factors than those in the Business-to-Business sectors over the past 3 years. For example, companies Consumer Staples, Information Technology, and Health Care demonstrate, on average, over a 20% increase in ESG mentions since 2016.

Table 5. ESG factors and their distribution in the earnings call transcript dataset.

Environmental	**%**	**Social**	**%**	**Governance**	**%**
Biodiversity	4.5	Community Outreach	20.6	Audit Committee	3.6
Climate Change	4.4	Customer Satisfaction	17.3	Board Composition	<1
Ecosystem Change	4.6	Data Privacy	<1	Business Ethics	19.5
Energy Efficiency	32.2	Discrimination	7.4	Corruption Instability	<1
Environmental Policy	7.0	Employee Engagement	5.7	Employee Board Diversity	<1
Environmental Risks	14.3	Human Rights	<1	Executive Compensation	13.7
Greenhouse Gas Emission	2.9	Labor Relations	28.4	Governance Policy	17.1
Pollution	2.5	Labor Standards	1.0	Risk Management	42.0
Renewable Energy	22.1	Political Contributions	1.6	Shareowner Rights	1.4
Resource Depletion	<1	Political Risk	10.4	Tax Transparency	<1
Toxic Chemical Use Disposal	<1	Product Safety Quality	3.2	Voting Risk	<1
Waste Management	5.4	Sexual Harassment	<1		
		Social Policy	3.3		

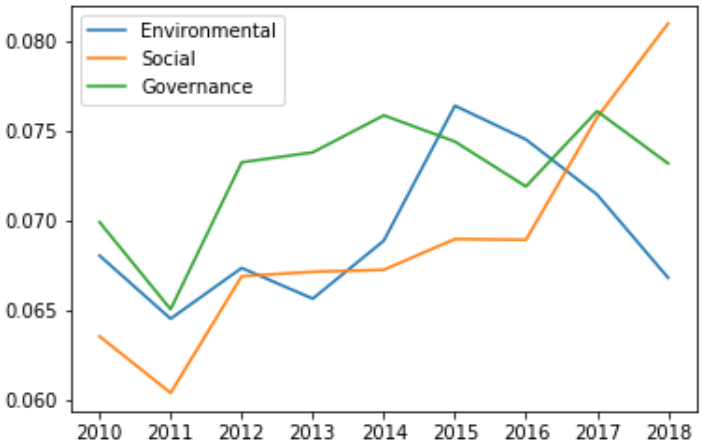

Figure 5. Trend analysis of ESG mentions in earnings calls.

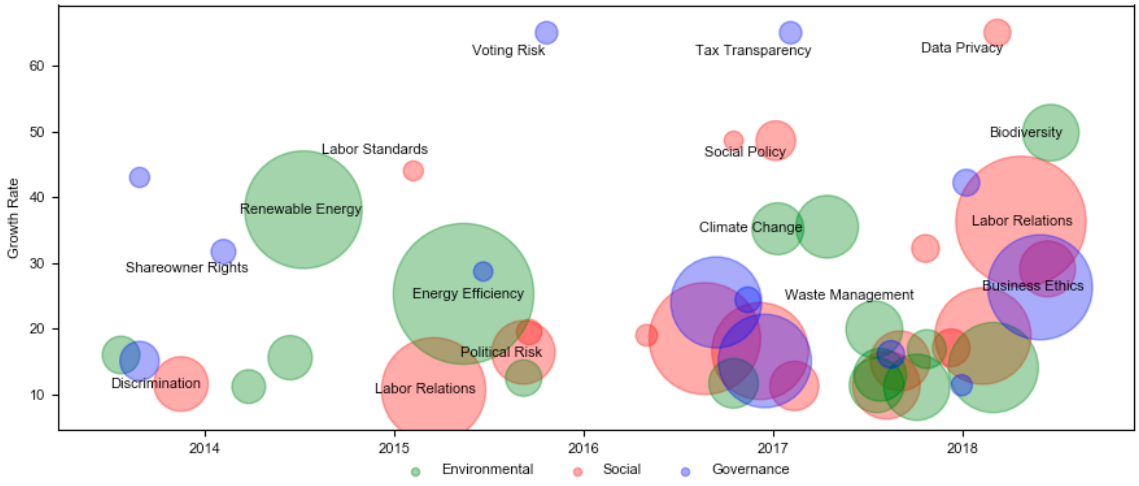

Figure 6. Thematic trend analysis of growth in ESG popularity during earnings calls over the past 5 years. Each bubble correspond to an ESG factor and the size reflects the number of ESG relevant mentions in the transcripts.

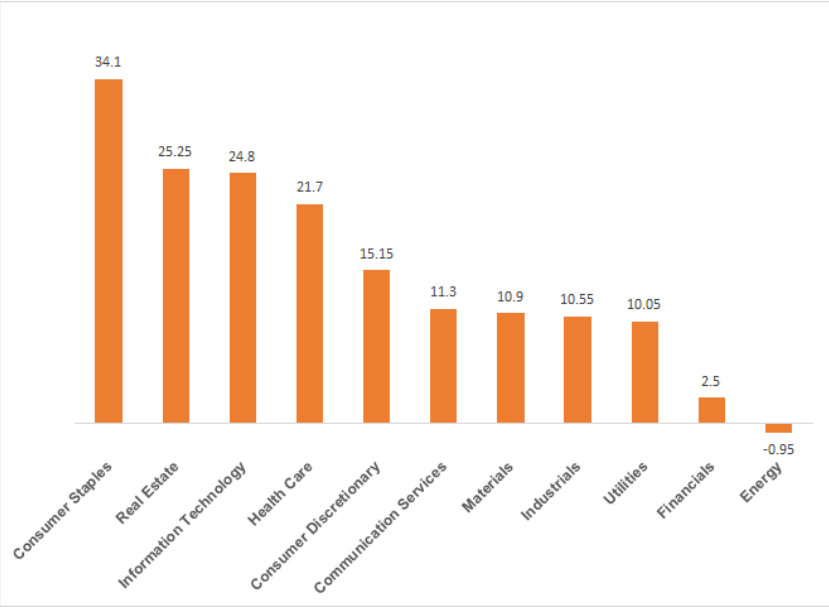

Figure 7. Growth rate of Relevant ESG mentions between 2016 and 2018 for various sectors.

Factors that Impact ESG Considerations. While the increasing trend in ESG considerations for S&P 500 companies is clear, these factors take a backseat when other performance-related concerns are at play in a given industry. In Q1 2016, energy prices fell considerably and hit historical lows. This decline in commodity prices directly impacted companies in the Energy sector and subsequently impacted companies in the Utilities, Industrial, and Materials sectors. We observed that, in 2016 companies in those sectors significantly decreased the time spent discussing ESG factors in their earnings calls by as much as 11%, as most were focused on financial and operational performance.

Large-cap vs. Mid- and Small-cap Companies. Large-cap companies have more resources to dedicate to ESG-related initiatives than small- or mid-cap companies [31] which allows them to develop multi-year ESG strategies and initiatives. On average, large-cap companies have increased or held steady on their ESG mentions over time. However, small- or mid-cap companies demonstrate slightly inconsistent trends in ESG mentions with declines in one year of more than 3% and increases in other years of 20%. Even in industries impacted by the 2016 energy crisis, such as the Industrials sector, large-cap companies remain consistently focused on ESG initiatives while mid-cap companies showed double digit declines in ESG discussions during that time period (Figure 8).

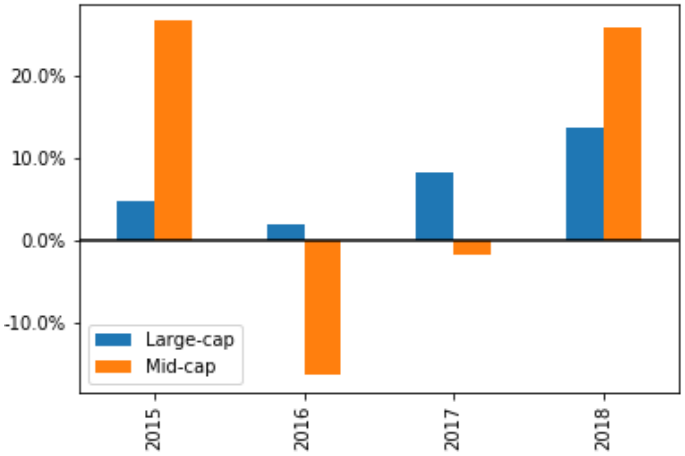

Figure 8. ESG mentions by market-cap designations.

While we observe directionally positive trends in ESG considerations by company management regardless of industry, theme or company size, ESG factors were and remain a meaningful focus for companies and investors alike. It appears that, given the extent of steady and reliable focus on ESG factors, ESG is an integral component of company operations and consequently company analysis by investors. Thus, the desire by many ESG investors for an integrated ESG strategy by companies may have, in fact, already been met in the financial marketplace.

6. Conclusions

ESG has emerged as a focal point for business strategy as investor demand for sustainability grows. It is important to exploit the latest advances in machine learning to analyze ESG related information and our study contributes to this effort. In particular, a novel distant-learning approach to profile the ESG-related language in corporate earnings calls was presented. Furthermore, the results demonstrated the importance of fine-tuning contextual embeddings. The solution proposed here is an important building block for developing a decision support system in the realm of sustainable finance. For example, our model can be applied to automatically derive an ESG relevance score on any corporate discussion. This score can then be used to efficiently filter and rank ESG discourse, thereby providing a mechanism to measure and assess the importance attributed to sustainable business practices. Our study revealed the steady growth in discussion of ESG factors in earnings calls, which suggests that such factors are not viewed in isolation and are integral to company operations.

There are several future research directions arising out of this work. While this study focuses on U.S. companies, it would be interesting to include international companies and multi-lingual conference transcripts. This would require a simple extension to our proposed model to train on foreign languages. Measuring the direct impact of ESG discussions on the financial performance of a company is another promising research direction. It would also be beneficial to combine external data sources, such as news and social media, with company disclosures, such as sustainability reports and earnings calls, when analyzing ESG trends. Finally, we intend to track language pattern changes in the ESG discourse.

Author Contributions: Conceptualization, N.R., G.B. and A.N.; methodology, N.R. and A.N.; software, N.R.; formal analysis, G.B.; data curation, N.R. and G.B.; writing—original draft preparation, N.R., G.B. and A.N.; visualization, N.R. and G.B.; supervision, A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This paper was prepared for information purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. J.P. Morgan makes no representation and warranty whatsoever and disclaims all liability for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

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

References

1. Vincent, O.M. The Impact of Corporate Environmental Responsibility on Financial Performance: Perspective From the Multinational Extractive Sector. Ph.D. Thesis, Brunel University, London, UK, 2012.
2. Friede, G.; Busch, T.; Bassen, A. ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *J. Sustain. Financ. Investig.* **2015**, *5*, 210–233. [CrossRef]
3. Becchetti, L.; Pinnacchio, D.; Giacomo, S.D. The impact of Social Responsibility on productivity and efficiency of US listed companies. *Appl. Econ.* **2007**, *40*, 541–569.

4. Avlonas, N. Sustainability Reporting Trends in North America. Available online: https://www.cse-net.org/wp-content/uploads/documents/Sustainability-Reporting-Trends-in-North%20America%20_RS.pdf (accessed on 7 October 2020).
5. Kwon, S. State of Sustainability and Integrated Reporting 2018. Available online: https://corpgov.law.harvard.edu/2018/12/03/state-of-integrated-and-sustainability-reporting-2018/ (accessed on 7 October 2020).
6. Shahi, A.M.; Issac, B.; Modapothala, J.R. Automatic analysis of corporate sustainability reports and intelligent scoring. *Int. J. Comput. Intell. Appl.* **2014**, *13*, 1450006. [CrossRef]
7. Brown, M. Managing Nature–Business as Usual: Resource Extraction Companies and Their Representations of Natural Landscapes. *Sustainability* **2015**, *7*, 15900–15922. [CrossRef]
8. Wang, W.Y.; Hua, Z. A Semiparametric Gaussian Copula Regression Model for Predicting Financial Risks from Earnings Calls. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, 23–25 June 2014; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 1155–1165. [CrossRef]
9. Keith, K.; Stent, A. Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls. *arXiv* 2019, arXiv:1906.02868. Available online: https://arxiv.org/abs/1906.02868 (accessed on 7 October 2020).
10. Qin, Y.; Yang, Y. What You Say and How You Say It Matters: Predicting Stock Volatility Using Verbal and Vocal Cues. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August; 2019; pp. 390–401. [CrossRef]
11. Napier, E. Technology Enabled Social Responsibility Projects and an Empirical Test of CSR's Impact on Firm Performance. Ph.D. Thesis, Georgia State University, Atlanta, GA, USA, 2019.
12. Hisano, R.; Sornette, D.; Mizuno, T. Prediction of ESG compliance using a heterogeneous information network. *J. Big Data* **2020**, *7*, 1–19. [CrossRef]
13. Nematzadeh, A.; Bang, G.; Liu, X.; Ma, Z. Empirical Study on Detecting Controversy in Social Media. *arXiv* **2019**, arXiv:1909.01093.
14. Ribando, J.M.; Bonne, G. *A New Quality Factor: Finding Alpha With ASSET4 ESG Data*; Starmine Research Note; Thomson Reuters: New York, NY, USA, 2010; Volume 31.
15. El Ghoul, S.; Guedhami, O.; Kwok, C.C.; Mishra, D.R. Does corporate social responsibility affect the cost of capital? *J. Bank. Financ.* **2011**, *35*, 2388–2406. [CrossRef]
16. Khan, M. Corporate Governance, ESG, and Stock Returns around the World. *Financ. Anal. J.* **2019**, *75*, 103–123. [CrossRef]
17. Krueger, P.; Sautner, Z.; Starks, L.T. The importance of climate risks for institutional investors. *Rev. Financ. Stud.* **2020**, *33*, 1067–1111. [CrossRef]
18. Guo, T.; Jamet, N.; Betrix, V.; Piquet, L.A.; Hauptmann, E. ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction. *arXiv* **2020**, arXiv:2005.02527.
19. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
20. Goel, T.; Jain, P.; Verma, I.; Dey, L.; Paliwal, S. Mining Company Sustainability Reports to Aid Financial Decision-Making. Available online: https://www.researchgate.net/publication/343305380 (accessed on 30 July 2020).
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; NIPS: Vancouver, BC, Canada, 2017; pp. 5998–6008.
22. Peters, M.E.; Neumann, M.; Zettlemoyer, L.; Yih, W.t. Dissecting contextual word embeddings: Architecture and representation. *arXiv* **2018**, arXiv:1808.08949.
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
24. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*; NIPS: Vancouver, BC, Canada, 2019; pp. 5753–5763.

25. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
26. Cisco. Corporate Social Resposibility Report. Available online: https://www.cisco.com/c/dam/m/en_us/about/csr/csr-report/2019/_pdf/csr-report-2019.pdf (accessed on 15 March 2019).
27. Rahdari, A.H.; Rostamy, A.A.A. Designing a general set of sustainability indicators at the corporate level. *J. Clean. Prod.* **2015**, *108*, 757–771. [CrossRef]
28. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:1910.03771.
29. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
30. Allianz. Ethics and Investing: How Environmental, Social, and Governance Issues Impact Investor Behavior. 2020. Available online: https://www.allianzlife.com/-/media/files/allianz/pdfs/esg-white-paper.pdf (accessed on 10 February 2020).
31. Polk, D. UN Sustainable Development Goals—The Leading ESG Framework for Public Companies? Available online: https://www.davispolk.com/files/2018-09-20_un_sustainable_development_goals_the_leading_esg_framework_for_large_companies.pdf (accessed on 10 February 2020).