NLP for SDGs: Measuring Corporate Alignment with the Sustainable Development Goals

Mike Chen, George Mussalli, Amir Amel-Zadeh, and Michael Oliver Weinberg

Mike Chen
is the head of sustainable investments at PanAgora Asset Management in Boston, MA.
mchen@panagora.com

George Mussalli
is the chief investment officer and director of research and equity at PanAgora Asset Management in Boston, MA.
gmussalli@panagora.com

Amir Amel-Zadeh
is an associate professor of accounting at Saïd Business School, University of Oxford in Oxford, UK.
amir.amelzadeh@sbs.ox.ac.uk

Michael Oliver Weinberg
is an adjunct professor at Columbia Business School in New York, NY.
mow5@columbia.edu

KEY FINDINGS

- The authors use modern NLP methods to identify companies that contribute to the UN SDGs.
- Combining NLP with machine learning methods for classification allows scalability in measuring SDG contribution of public companies with reasonably high accuracy.
- Using Doc2Vec embeddings with support vector machine classifiers results in the highest predictive performance.

ABSTRACT

This article uses advanced natural language processing (NLP) methods to identify companies that are aligned with the UN Sustainable Development Goals (SDGs) based on the text in their sustainability disclosures. Using the Corporate Social Responsibility (CSR) reports of Russell 1000 companies between 2010–2019, we apply a logistic classifier, support vector machines (SVM), and a fully-connected neural network to predict alignment with the SDGs. Specifically, we use word embeddings to augment dictionary-based input features, as well as the embeddings as features themselves, based on Word2Vec and Doc2Vec models to classify companies' alignment with the SDGs over time. Notably, the Doc2Vec embedding inputs to the SVM classifier result in high average accuracy of above 80% for predicting alignment.

Asset owners and asset managers are increasingly concerned about the environmental and social impact of their investments. For example, the Global Sustainable Investment Alliance (GSIA) reports that, at the beginning of 2020, sustainable assets stood at more than $35 trillion (GSIA 2020), and over 4,000 asset managers and owners managing more than $100 trillion in assets have signed the Principles of Responsible Investing (PRI).[1] This trend is largely driven by the demands of their constituents to not only consider financial metrics when making asset allocation decisions, but also to assess the long-term impact of their investments on the planet and society (Amel-Zadeh and Serafeim 2018). This shift has led to massive growth of so-called sustainable investment products that screen companies based on their sustainability or environmental, social and governance (ESG) attributes.[2]

[1] https://www.unpri.org/pri/about-the-pri.
[2] See, for example, WEF (2019) chapter 6.

Recent academic evidence also suggests that institutional and wealthy investors increasingly tilt their portfolios to more sustainable investments and engage with portfolio companies on sustainability issues (Dimson, Karakas, and Li 2020; Gibson and Krüger 2018; Dyck et al. 2019; He, Kahraman, and Lowry 2020; Krüger, Sautner, and Starks 2020; Amel-Zadeh, Lustermans, and Pieterse-Bloem 2020).

When incorporating sustainability considerations in their asset allocation decisions, investors face difficulty in measuring the multiple dimensions of ESG given the paucity of universally agreed upon ESG reporting standards and the resulting lack of comparability for ESG disclosures. Naturally, investors have turned to third-party ESG ratings providers that use proprietary technology to summarize the various ESG attributes into one score. These ratings aggregate a mixture of quantitative and qualitative information, and often measure policies and targets rather than outcomes. Moreover, recent research finds low correlations across various ratings providers, suggesting that disagreement exists even in quantitative ESG data and ultimately about whether a particular company is actually sustainable or not (Berg, Kölbel, and Rigobon 2020; Dimson, Marsh, and Staunton 2020).

Furthermore, ESG ratings do not always allow the sustainability-minded investor to understand how a company's actions contribute the UN Sustainable Development Goals (UN SDGs), which have become the overarching benchmark against which companies' ESG efforts are being measured. The SDGs were adopted by the UN in 2015 following the Millennium Development Goals. They comprise 17 high-level goals to be achieved by 2030 that are further broken down into 169 targets aimed at, for example, eradicating poverty (SDG 1), good health (SDG 3), education (SDG 4), economic growth (SDG 8), and action to combat climate change (SDG 13). The aim of the SDGs is to serve as a blueprint for collective global action by private and public sector entities to reduce social inequalities and environmental damage. In recent years, companies increasingly reference the SDGs when reporting on their ESG activities.[3]

Moreover, companies increasingly report on how their core business is aligned with the SDGs to highlight the positive impact their business has on people and the planet. In contrast, ESG disclosures often predominantly focus on how companies reduce or avoid a negative impact. It is therefore difficult to measure a public company's contribution to the SDGs using ESG ratings alone, as these generally do not directly map onto the 17 goals.

In addition, public companies are often encouraged to focus their sustainability-related disclosures in financial reports on a narrower set of financially material ESG issues that might only include a subset of the targets of the SDGs. Hence, at present, it is difficult for asset managers and owners to understand and track to which of the 17 Goals their investee companies contribute, and how much. Some have resorted to using analyst judgment to tag companies as either positively or negatively contributing to the SDGs (Schramade 2017). While it is feasible to get an understanding of a company's contribution to the SDGs from analyzing its business model and value chain, such methods do not scale easily to a larger universe of stocks.

In this study, we propose the use of machine learning (ML) and natural language processing (NLP) techniques to measure whether companies are aligned with the UN SDGs in their activities. Specifically, we use a large sample of text from corporate sustainability reports for Russell 1000 companies from 2010–2019 to train various ML algorithms using state-of-the art NLP processing methods based on distributed word representations to quantify and map the narrative descriptions in corporate sustainability reports (CSR) reports to the 17 UN SDGs. The application of machine learning and NLP enables investors to automate their SDG analysis of portfolio companies, potentially enabling index and quantitative funds to deploy capital at scale. This would

[3] KPMG reports that around 40% of the 250 largest companies report on the SDGs (Blasco, King, and Jayaram 2018).

further help impact-aligned investors, that is, those seeking positive impact on the SDGs, achieve scale.

To train the machine learning algorithms and assess the accuracy of the SDG assignments, we utilize human hand-coded mappings of Refinitiv Asset4 (Datastream) ESG metrics to the SDGs. Using a logistic classifier, support vector machines (SVM), and a dense shallow neural network with input features constructed from word embeddings based on Word2Vec and Doc2Vec models, we find reasonably high out-of-sample accuracy in predicting SDG alignment. For example, an SVM with Doc2Vec achieves up to 83.5% accuracy with a F1-score of 78%. Our other model implementations achieve similarly high prediction accuracy. That is, we show that it is possible to identify which companies are aligned with which SDGs through their textual disclosures in an automated and scalable fashion.

The purpose of this article is to provide initial proof of concept to show how investors can apply machine learning and natural language processing techniques to overcome challenges in measuring companies' contribution to the SDGs. We also highlight several limitations and suggest avenues for future research. One limitation is that, in its current form, our method only measures a company's alignment with a specific SDG, that is, a binary outcome, and not the extent of the company's (positive or negative) contribution, that is, an assessment on a continuous scale.[4]

Furthermore, our approach relies on the companies' own descriptions of their environmental and social activities in their CSR reports. The text we use as the basis to develop our measures thus captures the management's personal views of how their company's operations relate to several ESG dimensions. As firms tend to describe as many ESG activities in their CSR reports as possible, our measures likely represent an upper bound with which and how many SDGs the cross-section of the Russell 1000 is aligned.[5]

Notwithstanding potential "greenwashing," the text nevertheless allows us to map what companies describe in their CSR reports to descriptions of the UN SDGs, without having to rely on companies themselves attributing their actions to specific SDGs. Moreover, as we use mappings from ESG scores to the SDGs provided by Refinitiv to train the algorithms, we rely on a third party to measure alignment.[6] While we focus the task on identifying which companies contribute to which SDGs, simple extensions to our models would allow investors to also measure the extent to which the identified companies contribute to the SDGs. We provide several suggestions for future extensions.

In parallel, there have also been other industry-initiated efforts to apply machine learning to measure SDG alignment. For example, the Sustainable Development Investments Asset Owner Platform (SDI AOP) was established and is backed by asset owners to enable investors to assess their global capital markets' portfolios on their contribution to the SDGs and to report to their clients and external stakeholders

[4] In other words, as companies likely remain silent about their potential negative impacts on particular SDGs, our method will simply find no alignment. That is, our measure has a lower bound of zero, and is potentially, in its current form, an overestimate of a company's "true" overall alignment. We discuss these limitations and provide suggestions for extensions to the models to overcome these in the section, Limitation & Future Research.

[5] We also only limit training the algorithms based on the text in CSR reports. One might potentially achieve further refinement of the predictions by also including annual reports (and potentially other disclosures). A cursory look at several 10-K reports, however, reveals little discussion of topics related to the SDGs in 10-Ks. We therefore leave it to future research to assess the additional informativeness of 10-Ks for our analysis.

[6] As such, we rely on Refinitiv's ESG scores largely being unaffected by a company's greenwashing efforts.

EXHIBIT 1

Sample of CSR Reports

Year	Number of Reports	Average Report Length (number of words)
2010	299	8704
2011	374	12967
2012	336	10532
2013	383	11003
2014	384	10215
2015	416	10443
2016	435	9689
2017	565	7748
2018	597	8164
2019	654	8989
Total	4779	9845

NOTE: This exhibit shows the number and average length of CSR the reports in our sample by year.

transparently and consistently, using a common and auditable standard (see https://sdi-aop.org/).

Overall, this study shows that modern machine learning techniques can aid investors in measuring investee companies' alignment with the SDGs. This article provides initial evidence on the feasibility of using natural language processing in this context, serving as a first step for possible further research and practical applications in this area.

SAMPLE AND DATA

Sample

Annual sustainability reports are obtained from Bloomberg for Russell 1000 companies for 2010–2019. Not all Russell 1000 companies provide sustainability reports annually, and for some firms, no reports are available throughout the sample period. Exhibit 1 represents the number of reports in our sample per year. The number of firms issuing sustainability reports increases over the years, leaving us with a corpus of 4,779 sustainability reports for which the text can be processed.

We apply standard text data cleaning routines. First, all numbers and stop words are removed, as they do not contribute to the meaning of the text for our purposes. Stop words include prepositions and articles such as "our," "in," "the," etc. We then transform all words to lower case and their lemmas. Lemmatization removes inflectional endings from words and returns them to their base form. Finally, each cleaned report is tokenized into the remaining lemmas. Exhibit A1 in the Appendix shows an example of the result of the cleaning and lemmatization process.

SDG DICTIONARY AND ESG SCORES

To construct the seed dictionary for the word embeddings, text descriptions of each goal are scraped from the UN SDG website (https://sdgs.un.org/goals). The text contains a description for each of the 17 goals and their indicators and a short report on current progress toward each goal. The same text cleaning processes as above are applied, and the descriptions are tokenized into unigrams and bigrams. We then sort the tokens by frequency and select the most descriptive uni- and bigrams for the seed dictionary for the respective SDG goal. Each goal has between 17 and 28 seed tokens in the seed dictionary. Exhibit 2 summarizes the final seed words for each of the SDG goals. In the section Word2Vec embeddings, we further discuss how we use the seed words in our Word2Vec model to augment the SDG dictionary used in the classification models.

The next step is to establish which company is aligned with which SDG in the training sample, the so-called "*ground truth*" in machine learning lingo. To label the training sample, we can either map companies to SDGs manually using our judgment, or use ESG metrics that are already mapped to the SDGs to determine the company's SDG alignment. We chose the latter approach for greater objectivity.

EXHIBIT 2

SDG Seed Dictionary

SDG	Goal	Words Associated with SDG
SDG 1:	No Poverty	Social, end poverty, poverty dimension, poverty, social protection, poor, unemployed person, poverty line, protection, cash benefit, extreme poverty, poor vulnerable, humanitarian, vulnerable
SDG 2:	Zero Hunger	Malnutrition, hunger, food producer, underweight, hunger malnutrition, undernutrition, famine, food insecurity, agricultural productivity, agricultural, extreme hunger, agriculture, prevalence undernourishment, nutritional need, food
SDG 3:	Good Health and Well-Being	Life expectancy, mental health, air pollution, medicine vaccine, infectious disease, good health, respiratory disease, reproductive health, mortality, healthcare, disease diabetes, disease, health coverage, health, maternal mortality, death preventable, cardiovascular disease
SDG 4:	Quality Education	Teacher, secondary school, proficiency level, primary school, inclusive, literacy, literacy numeracy, higher education, quality education, school, effective learn, vocational train, level proficiency, minimum proficiency, technical vocational
SDG 5:	Gender Equality	Domestic work, right, sexual violence, woman, girl, discrimination, reproductive health, managerial position, woman girl, marriage, woman representation, gender equality, gender parity, child marriage, gender
SDG 6:	Clean Water and Sanitation	Water sanitation, drink water, sanitation, basic drink, wastewater, water scarcity, hygiene, water, sanitation service, sanitation hygiene, supply freshwater, hand wash facility, resource management, water stress
SDG 7:	Affordable and Clean Energy	Energy, technology, renewable energy, infrastructure, electricity, cheap energy, solar, solar power, wind power, thermal power, energy productivity, energy efficiency, greenhouse gases, greenhouse, fossil fuels, pollution, energy standards, energy access, energy consumption, access electricity, without electricity, fuel technology, fossil fuel
SDG 8:	Decent Work and Economic Growth	Labor, employment, gdp, job, unemployed, economic growth, productivity, job creation, slavery, forced labor, labor force, women participation, labor organization, human right, informal employment, growth rate, labor productivity, decent work, secure work, global economic, gender pay, crisis level, rate real, decent work, education employment, slavery human, child labor, youth employment
SDG 9:	Industry, Innovation, and Infrastructure	Research development, development, industry, infrastructure, transport, technological progress, communication technology, sustainable development, sustainable industries, innovation, entrepreneurship, access information, access internet, material footprint, develop country, least develop, economic infrastructure, infrastructure support, global manufacture, scientific research, resilient infrastructure, research innovation
SDG 10:	Reduced Inequalities	Income, population, income inequality, economic inclusion, safe migration, economic inequality, reduce inequality, wealth share, indigenous rights, migrant worker, migrant, official development, income inequality, migration mobility, global wealth, least develop, development assistance
SDG 11:	Sustainable Cities and Communities	City, urban, urban population, public, disability, disaster, sustainable city, affordable housing, housing access, resilient societies, public transport, public spaces, urban planning, inclusive, business opportunities, sustainable development, person disability, green public, sustainable resilient, sustainable urbanization, population convenient, convenient access
SDG 12:	Responsible Consumption and Production	Responsible consumption, sustainable development, resources, consumption, production, development, reduce waste, efficient, efficient economy, energy consumption, energy efficient, supporting developing, material footprint, natural resource, recycle, sustainable consumption, domestic material, consumption production, food waste
SDG 13:	Climate Action	Climate, develop, disaster, local, emissions reductions, global warm, climate change, climate system, greenhouse gas, emissions, co2 emissions, low carbon, disaster risk, sustainable management, natural resource, sea levels, sustainable energy, Paris Agreement, sustainable energy, climate relate, green climate, disaster risk
SDG 14:	Life Below Water	Marine, ocean, fish, sea, water, fishery, overfishing, coastal biodiversity, coastal ecosystems, ocean resources, marine biodiversity, fish stocks, marine pollution, ocean acidification, depleted fisheries, unregulated fish, fish stock, fishery management, marine coastal, fishery subsidy, conservation sustainable, marine technology
SDG 15:	Life on Land	Land degradation, terrestrial freshwater, ecosystem, deforestation, species animal, forest management, biodiversity, conservation, protect area, forest, terrestrial, species, wildlife, protect, agriculture, land, area
SDG 16:	Peace, Justice, and Strong Institutions	Law, human right, right violation, insecurity, institution, violence, exploitation, global governance, transparency, rule, corruption bribery, access justice, corruption, justice, peace, conflict violence, international, victim human, conflict
SDG 17:	Partnerships for the Goals	Sustainable development, innovation, enhance international, official development, capacity build, development, coordinate policy, develop country, assistance, cooperation, international support, international cooperation, partnership, international, policy coherence, development assistance, country

NOTE: Exhibit 2 shows the seed words for the SDG seed dictionaries.

EXHIBIT 3

SDG Datastream ESG Metric Mapping

UN SDG	Datastream ESG Metrics
SDG 1:	Product Access Low Price, Community Lending, and Investments
SDG 2:	Obesity Risk
SDG 3:	Policy Employee Health & Safety, Policy Supply Chain Health & Safety, Health & Safety Training, Supply Chain Health & Safety Training, HIV-AIDS Program
SDG 4:	TRDIR People Development Score, Policy Skills Training, Policy Career Development, Employees with Disabilities, Training Costs Per Employee
SDG 5:	TRDIR Diversity Score, TRDIR Inclusion Score, Policy Diversity and Opportunity, Targets Diversity and Opportunity, Human Rights Policy
SDG 6:	Policy Water Efficiency, Toxic Chemicals Reduction, Targets Water Efficiency, Water Use To Revenues USD, Waste Reduction Initiatives, Waste Recycling Ratio, Biodiversity Impact Reduction, Water Technologies
SDG 7:	Renewable Energy Use, Policy Energy Efficiency, Renewable/Clean Energy Products, Product Environmental Responsible Use
SDG 8:	Human Rights Policy, Policy Child Labor, Policy Forced Labor, Employees with Disabilities, Policy Human Rights
SDG 9:	Environmental Innovation Score, Community Lending & Investments, Product Sales at Discount to Emerging Markets
SDG 10:	TRDIR Diversity Score, TRDIR Inclusion Score
SDG 11:	Product Access Low Price
SDG 12:	Environmental Materials Sourcing, Policy Water Efficiency, Policy Energy Efficiency, Policy Sustainable Packaging, Policy Environmental Supply Chain, Take-back and Recycling Initiatives, Waste Recycling Ratio, Total Waste Reduction, CSR Sustainability Reporting, Resource Use Score, Emissions Score
SDG 13:	Climate Change Commercial Risks & Opportunities
SDG 14:	n/a
SDG 15:	Biodiversity Impact Reduction, Environmental Project Financing
SDG 16:	Human Rights Policy, Policy Child Labor, Fundamental Human Rights ILO UN, Policy Bribery and Corruption
SDG 17:	Product Access Low Price

NOTE: This exhibit shows the SDG to Datastream ESG metric mappings.

The ESG metrics we use are from Refinitiv Asset4 (Datastream).[7] The advantage is that Refinitiv provides a comprehensive set of ESG metrics for Russell 1000 companies which are already mapped to the UN SDGs. For example, as part of the social score, the ESG category "Product responsibility: access to low price products" is mapped to SDG 1 as it relates to a company's offering of low-priced products specifically designed for lower income customers; or, as a part of the environmental score, the category "water use," is mapped to SDG 6. In total, 61 ESG metrics (33 social, 27 environmental, and 1 governance metric) are mapped to 16 SDGs. There is currently no mapping to SDG 14: Life below Water. Exhibit 3 summarizes the mappings.

The Datastream SDG mappings are used to construct a binary variable that represents alignment with each SDG. A company is designated as aligned with the respective SDG depending on its score on the ESG metrics mapped to the SDG. If only one or two ESG metrics are mapped to the SDG, alignment requires a score on at least one, whereas if more than two metrics are mapped to the SDG, alignment requires a score for at least two. For example, a company is aligned with SDG 1 if

[7] Other commercial ESG metric vendors were considered, but we found Refinitv's data to be the most complete and comprehensive.

EXHIBIT 4

Fraction of Sample Aligned with SDGs by Year

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
SDG1	8.78%	8.83%	8.72%	7.95%	6.60%	6.37%	6.88%	7.48%	7.73%	4.05%
SDG2	2.35%	2.54%	2.47%	2.05%	1.65%	1.30%	1.09%	0.99%	1.18%	0.73%
SDG3	15.95%	18.26%	19.67%	18.75%	20.13%	23.22%	26.75%	30.80%	33.48%	18.50%
SDG4	2.84%	2.78%	3.06%	2.84%	2.53%	5.18%	40.50%	57.10%	57.94%	27.44%
SDG5	0.00%	0.00%	0.00%	0.00%	0.00%	1.51%	35.26%	54.24%	57.62%	26.82%
SDG6	13.47%	15.60%	16.61%	18.07%	18.59%	19.98%	21.18%	23.21%	22.64%	10.60%
SDG7	21.14%	23.82%	26.50%	25.91%	24.53%	24.73%	26.75%	28.27%	28.33%	14.97%
SDG8	6.30%	8.46%	10.25%	13.07%	17.27%	22.89%	32.64%	41.58%	48.82%	32.33%
SDG9	28.68%	28.90%	28.03%	26.93%	25.85%	29.05%	29.26%	29.15%	28.54%	17.98%
SDG10	0.00%	0.00%	0.00%	0.00%	0.00%	1.84%	30.68%	41.36%	39.91%	22.14%
SDG11	8.53%	8.71%	8.60%	7.84%	6.27%	6.16%	6.66%	7.37%	7.51%	3.74%
SDG12	35.35%	34.58%	34.51%	33.07%	32.45%	36.50%	36.90%	37.07%	36.37%	22.35%
SDG13	39.93%	42.44%	42.52%	39.77%	36.96%	34.23%	38.10%	40.59%	42.17%	28.48%
SDG15	15.20%	15.96%	17.08%	15.80%	14.96%	14.58%	15.72%	17.60%	18.24%	10.08%
SDG16	13.35%	15.48%	17.20%	18.98%	22.33%	27.54%	35.70%	43.01%	48.82%	31.39%
SDG17	8.53%	8.71%	8.60%	7.84%	6.27%	6.16%	6.66%	7.37%	7.51%	3.74%

NOTE: This exhibit shows the percentage of the sample firms by year that are predicted to be aligned with the particular SDG in the table row.

it scores on one of the two ESG metrics (product access and community lending) mapped to SDG 1.[8]

Exhibit 4 provides the summary statistic for the percentage of sample firms aligned with SDG 1 to 17 (except for SDG 14 as above). Overall, Exhibit 4 shows a large variation in SDG alignment in the cross-section and over time. Generally, a larger fraction of firms are aligned with SDG 7 (affordable and clean energy), SDG 9 (innovation), SDG 12 (responsible consumption), and SDG 13 (climate action). The table further documents an increasing fraction of firms over time aligned with SDGs that have become highly topical in recent years such as, for example, SDG 5 (gender equality) and SDG 10 (reduced inequalities). While before 2015 no firm was aligned with the two inequality goals, the number jumps to between 40–60% of the sample throughout 2017–18, likely in response to heightened societal attention on inequality, such as the emergence of the "Me-Too" movement during that time. The increasing fraction of companies that align with a broader cross-section of the SDGs also likely reflects companies' increasing awareness of heightened investor and other stakeholders' attention to the SDGs. This potentially incentivizes companies to increase referencing matters related to the goals in their CSR reports.[9]

[8] The thresholds for alignment are chosen to have a sufficiently large number of firms available in each SDG bucket. The threshold itself can be considered a hyperparameter of the model adjustable to investor preferences. Setting a higher threshold provides more stringent alignment criteria reducing the number of eligible firms, setting a lower threshold relaxes the alignment criteria producing a wider investable universe.

[9] We discuss related caveats for our methodology with respect to potential "SDG washing" in SDG reports in the section, Limitations and Future Research.

RESEARCH DESIGN

Feature Engineering

Word2Vec Embeddings. To create a more comprehensive list of words associated with each SDG, we train a Word2Vec embedding model using all pre-processed and cleaned reports over a five-year rolling window as input corpus to the model. We describe the model and training in more detail in Appendix A1. We train the model with negative sampling over 10 epochs using a window size of 5 and an embedding size of 50, and ignoring words with a frequency lower than 10. The outputs of the model are word vectors of length 50 containing word embeddings for each word in the cleaned CSR reports. The word vectors are a distributed representation of each word in a multi-dimensional vector space, here with 50 dimensions, where semantically similar words are located closer to each other.

Prior research has shown that the accuracy of word embeddings increases when word vectors that have been trained on large text corpora are used to initialize the parameters for training, a process known as transfer learning. We use Google's Word2Vec model that has been trained on Google's news dataset containing about 100 billion words to augment our model, increasing the embedding size to 300. We then augment our SDG dictionary with semantically similar terms in the CSR reports found through the Word2Vec model. The augmented dictionaries are constructed using the initial collection of uni- and bigrams associated with each SDG from the UN website as seed tokens in the Word2Vec model. Semantically similar terms are found by extracting the top 20 closest word vectors by cosine similarity to the vector representations of the seed words. This procedure results in an augmented dictionary of 200–300 similar words. The augmented SDG uni-and bigram dictionary is then used to calculate frequencies based on how often each token in the dictionary appears in a CSR report in a given year.

Narrowing the Set of Features. In a final step to construct the uni-and bigram features for the subsequent classification task, the augmented dictionaries are cleaned and purged of less important words. Important features are identified in the training sample through LASSO regressions and random regression forests.[10] The regressions use the ESG scores associated with each SDG as dependent variables, and the frequencies of the words from the augmented dictionaries as independent variables. The LASSO regression penalizes the regression weights to avoid overfitting, where weights for less predictive words are pushed toward zero. The final dictionaries retain only tokens with LASSO coefficients larger than zero. Similarly, the random forest enables identifying features that are of particular importance in reducing the prediction error at every node in the regression tree. We choose the top 50 features with the highest importance metric averaged over all regression trees of the random forest.

Predicting SDG Alignment

Word2Vec. After we train the Word2Vec embedding model for every five rolling years from 2010 to 2019, and augment our seed dictionaries associated with each SDG, we use the frequency of the identified important feature tokens from the previous section, Narrowing the Set of Features, and feed them into our classifier models. We train three different classifiers: logistic regression, SVMs, and a fully-connected neural network with binary assignment to one of the SDGs as a target variable. SDG alignment has been constructed from Refinitiv ESG scores as described in the earlier

[10] The models are described in more detail in the LASSO and Random Regression Forest section of the Appendix.

section, SDG Dictionary and ESG Scores. The models are trained during the four years of the five-year rolling window and tested out-of-sample in the fifth year.

One particular problem with our dataset is that far fewer firms are aligned with a given SDG in a given year than not, which means our training sample is imbalanced in the target variable. Imbalanced training sets can lead to classification challenges. For example, a naïve classifier that always selects the majority class, regardless of the input, can artificially result in a greater than 50% accuracy rate.

One approach to addressing class imbalance is to change the distribution of the training data by oversampling from the minority class. We use the Synthetic Minority Oversampling Technique, SMOTE (Chawla et al. 2002). SMOTE augments the training data with randomly generated synthetic data points that are close to the nearest minority class generated from a common region of nearest neighbors. The approach has been shown to be effective as it creates new examples that are relatively close in feature space to the existing examples, and thus enables the classifier to learn better decision regions.

Doc2Vec. As an alternative to the Word2Vec models trained above, we also train a Doc2Vec model on the data for the SDG alignment prediction task. With Doc2Vec, in addition to mapping each word to a unique vector as in the Word2Vec, each document is also mapped to a unique vector (Le and Mikolov 2014). The document and word vectors are then concatenated into a context vector to predict the word in the context. The document vectors act as a memory of the context in which each word appears and is shared across all context vectors generated from the same document. The word vectors, on the other hand, can be shared across documents.

We describe the Doc2Vec model in more detail in Appendix A2.

Similar to the Word2Vec embedding model, we train the Doc2Vec embedding model for every five rolling years from 2010 to 2019 by feeding each processed and cleaned document to the model. We use a distributed bag of words (PV-DBOW) algorithm with a vector size of 100, and a negative sampling of five "noise" words. We set the minimum word count at five and train the model over 30 epochs. The document vector embeddings from the Doc2Vec model serve as features for our three classifiers (logistic regression, SVM, and dense shallow neural net) as described in more detail next. We illustrate the model's set-up in Exhibit 5.

Classification Models

Logistic Regression. Our baseline model is a standard linear logistic classifier, in which the alignment indicators are the target variables, and the word frequencies from the SDG dictionaries are the features. The model predicts the outcome variable using a sigmoid function that estimates the probability that a company is aligned with a particular SDG in a given year, given the frequencies of how often words aligned with that SDG appear in the CSR report. The decision boundary for prediction of alignment is set at a probability >0.5. The algorithm minimizes the following cost function using gradient descent:

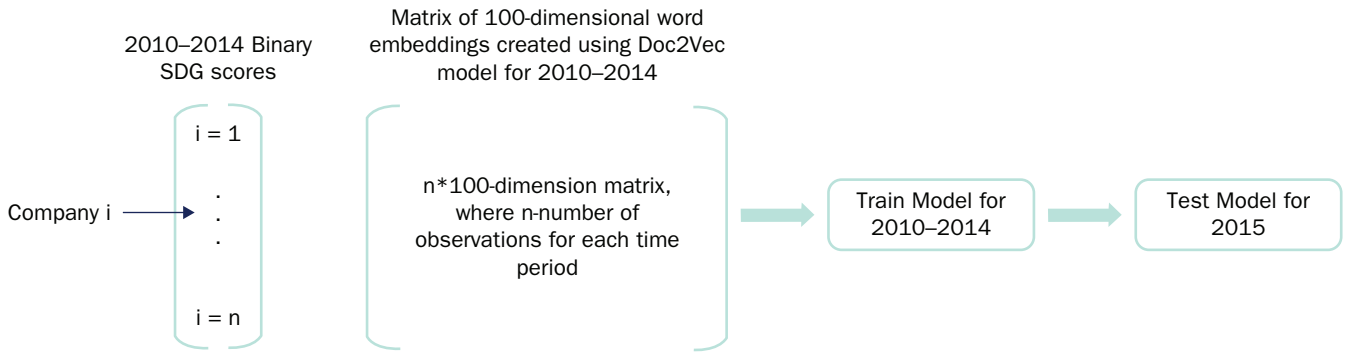
$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{N}(y^{(n)} log(h_\theta(x^{(n)})) + (1 - y^{(n)}) log(1 - h_\theta(x^{(n)}))), \text{ where} \quad (1)$$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^{Tx}}}. \quad (2)$$

Support Vector Machines. A second model we train for the prediction of SDG alignment is SVM. SVMs are supervised learning methods that are effective for

EXHIBIT 5

Illustration of the Doc2Vec Classification Set-Up



NOTE: This exhibit illustrates the use of Doc2Vec embeddings for the classification task.

classification tasks in high dimensional spaces. Through training on labelled data, the SVM builds a decision boundary that maps training examples to points in space by maximizing the gap between the different classes. When predicting the class labels for new observations, the model assigns a category based on which side of the decision boundary the observations fall. The SVM can be trained with linear and non-linear kernels. We calibrate the model using 10-fold cross-validation, choosing the best performing estimator among four different kernels: a linear kernel, a polynomial kernel, a radial basis function, and a sigmoid function. For a linear kernel, the cost function takes the form:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} max(1 - y^{(n)}(\theta^T x^{(n)} + \theta_0), 0) + \lambda \mid \theta \mid^2. \quad (3)$$

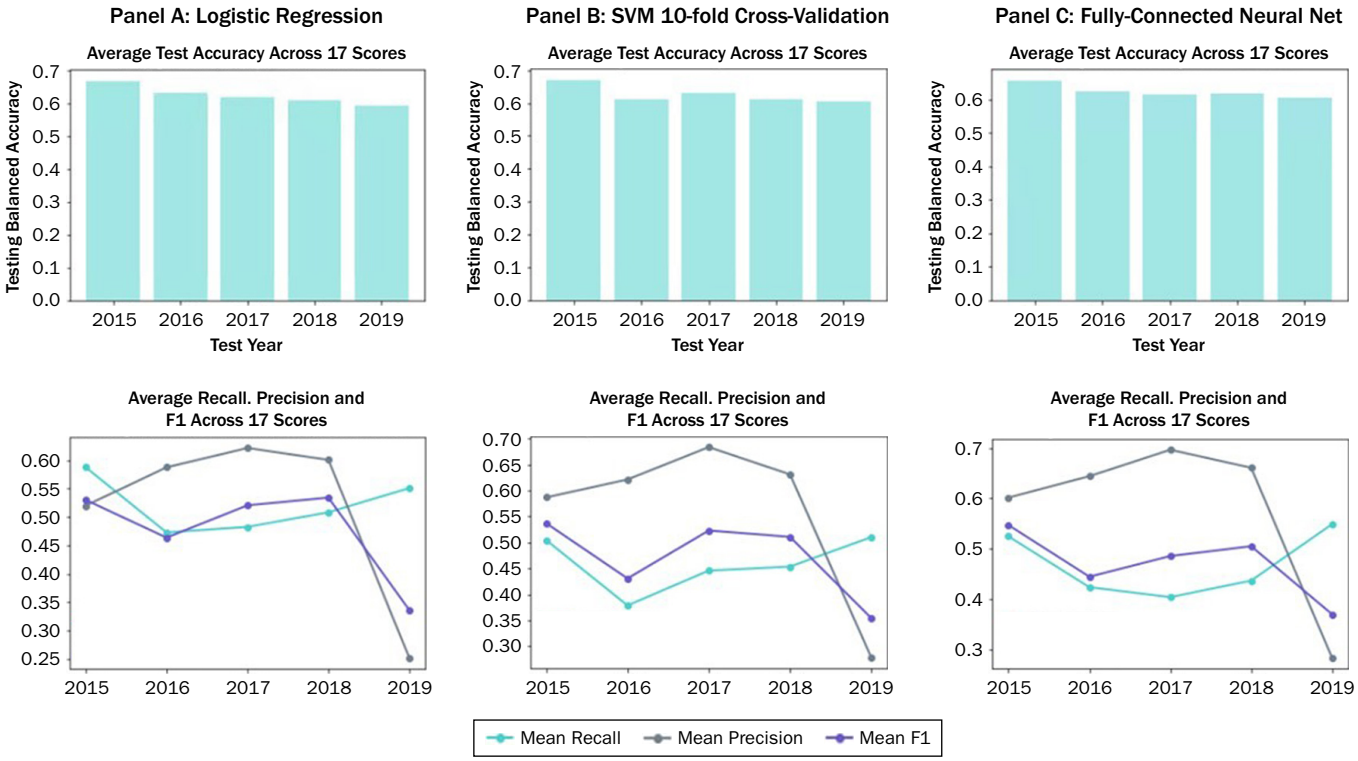
Neural Network. As a third classifier model, we use a fully-connected neural network. We use a shallow dense neural network with an input layer, a single hidden layer, and output layer. The hidden layer consists of 256 neurons. We use a rectified linear unit (ReLU) as an activation function for the hidden layer and a sigmoid function for the output layer. The ReLU function transforms each input x in a neuron to $max(0, x)$. To minimize the cost functions, the Adam optimizer is used. Adam, or Adaptive Moment Estimation, uses running averages of the first and second moments of the gradient to adapt the learning rate for each weight of the neural network (Kingma and Ba 2014). The learning rate is a hyperparameter of the training process that controls how much the weights are updated with respect to the loss gradient. The network is trained with a batch size of 128 samples over 100 epochs and a learning rate of 0.001.

Evaluation Metrics. We use several standard metrics to evaluate the prediction performance of the classification models on the test sample. As our sample is relatively unbalanced, a simple accuracy measure might result in misleading conclusions about the true accuracy of the models.[11] We therefore calculate the balanced out-of-sample accuracy, which individually represents the average, as well as the precision, the recall, and the F1 score, of the correctly predicted labels within each class.

[11] For example, consider a case with an extreme class imbalance where only one out of 100 observations is assigned the target variable, that is, equal to one, and the remainder is labeled zero. A naïve classifier that classifies the entire sample as not belonging to the target group, that is, classifies every observation as zero, attains a prediction accuracy of 99 percent. Clearly, such a classifier has, in fact, made no prediction at all and is not particularly useful.

EXHIBIT 6

Word2Vec Results



NOTES: This exhibit summarizes the prediction results on the test sample using Word2Vec generated input features of the Logistic Regression (Panel A), SVM classifier (Panel B) and Neural Network (Panel C). The top graph shows the average test accuracy across the 17 SDGs over the sample period and the bottom graph shows mean precision, recall and F1 score.

Precision measures the fraction of true positives to all samples that have been predicted positive (i.e., predicted as aligned with a particular SDG), and is the sum of true positives and false positives. Recall is equal to the fraction of true positives to all samples that are indeed positive, that is, the sum of true positives and false negatives. The F1-score is equal to the harmonic mean (i.e., the reciprocal of the arithmetic mean) of precision and recall.

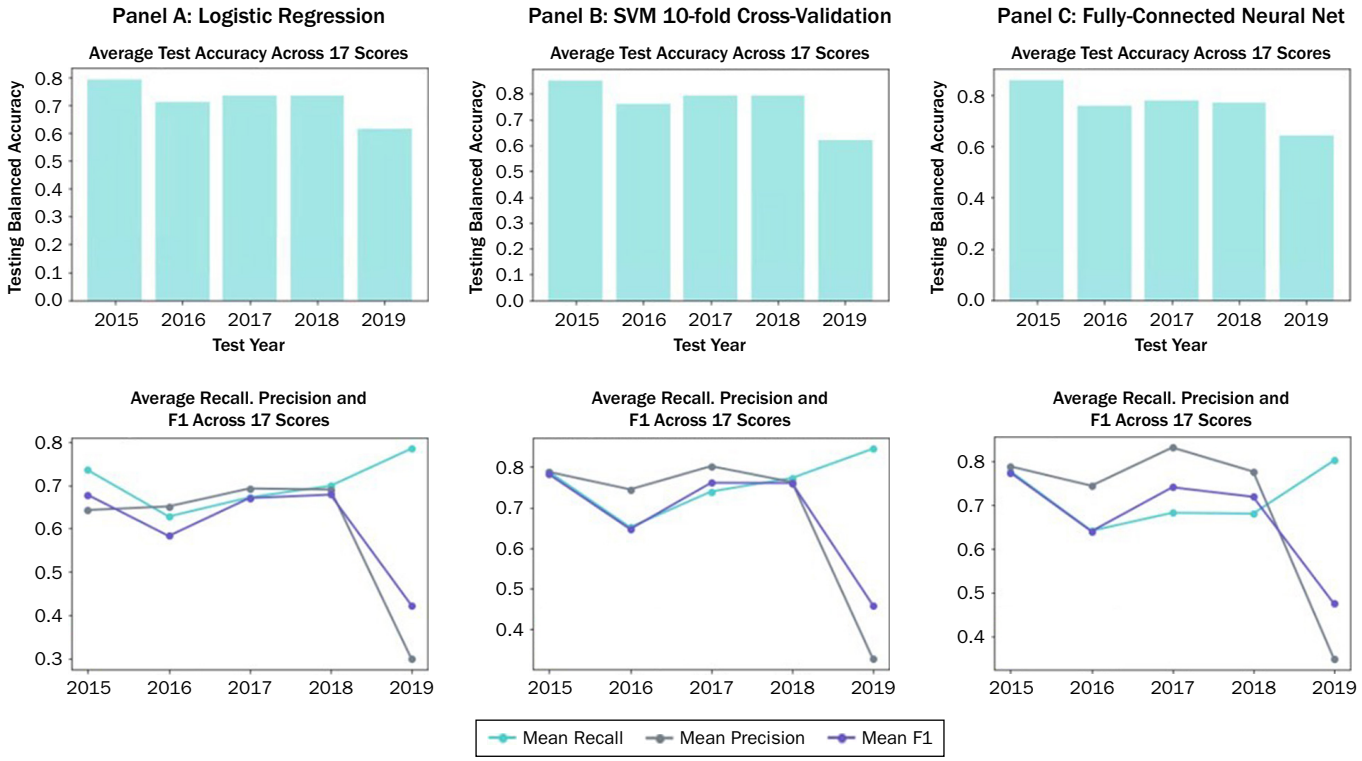
RESULTS

Word2Vec

Exhibit 6 summarizes the results for the test accuracy and recall, precision, and F1 scores for the three classification models using the Word2Vec-produced features over the test years 2015–2019. Panel (a) summarizes the results for the logistic classifier. The bar chart shows the average balanced accuracy score over all SDG classifications in each year. The average accuracy is 66.9% in 2015, reducing gradually to 59.5% in 2019. The line graph shows average precision, recall, and the F1 score per year. The average precision is above 50% for every year except 2019, when it drops off significantly. The mean recall is around 50%, while also starting off a bit higher in 2015. Similarly, the average F1 score remains around 53% in 2015–2018 (except for 2016), but drops to only 33.6% in 2019.

EXHIBIT 7

Doc2Vec Results



NOTES: This exhibit summarizes the prediction results on the test sample using Doc2Vec generated input features of the logistic regression (Panel A), SVM classifier (Panel B), and neural network (Panel C). The top graph shows the average test accuracy across the 17 SDGs over the sample period, and the bottom graph shows mean precision, recall, and F1 score.

Test scores for the SVM model shown in Panel (b) behave in a similar fashion, albeit being overall somewhat higher than those of the logistic regression. The average balanced test accuracy starts at 67.1% and drops to 60.5% over the five-year test period. The average precision of the model peaks at 68% in 2017, and, like the logistic regression, drops significantly in 2019. The mean F1 score is slightly higher than for the logistic classifier. Similar test results are obtained for the neural network model, shown in Panel (c), with slightly lower balanced accuracy over the test years than the SVM, but overall higher precision and recall, and thus also higher F1 score, than the SVM and the logistic classifier. Even though the test accuracy results for the Word2Vec-based models are reasonable, the precision and recall measures are relatively low, suggesting high type 1 and type 2 errors in the predictions. The gradually declining test scores over the test period further suggest that descriptions about the SDGs in companies' CSR reports change over time and are not fully captured with a rolling training window. We therefore extend our model using Doc2Vec embeddings that use context information about the entire documents for the classification as well. We present the results next.

Doc2Vec

Exhibit 7 summarizes the results for the Doc2Vec model. Compared to Word-2Vec, the Doc2Vec model results show an overall significant improvement in test scores for each model across the various metrics. The logistic classifier, shown

EXHIBIT 8

SVM Classifier Results for 2017 with Doc2Vec Features

	Balanced Test Accuracy	F1 Score	Recall	Precision
SDG1	0.87	0.70	0.81	0.62
SDG2	0.83	0.70	0.67	0.73
SDG3	0.79	0.80	0.74	0.86
SDG4	0.65	0.73	0.61	0.9
SDG5	0.67	0.76	0.66	0.89
SDG6	0.79	0.78	0.83	0.74
SDG7	0.78	0.79	0.76	0.82
SDG8	0.76	0.76	0.66	0.89
SDG9	0.81	0.81	0.83	0.79
SDG10	0.64	0.72	0.67	0.78
SDG11	0.88	0.73	0.82	0.66
SDG12	0.77	0.83	0.76	0.9
SDG13	0.79	0.77	0.67	0.92
SDG15	0.86	0.81	0.78	0.85
SDG16	0.76	0.79	0.71	0.89
SDG17	0.87	0.69	0.84	0.59

NOTE: This exhibit shows test sample accuracy, precision, recall, and F1 score for the SVM classification in 2017 for each SDG in the exhibit rows separately.

in Panel (a), achieves a mean balanced test accuracy of 76.1% in 2015, dropping to 67.4% in 2019—a marked improvement compared to the equivalent classifier using Word2Vec generated inputs. Average precision scores range from 64.3% to 69.3%, again with the exception of the last test year, for which the average precision falls dramatically. Mean recall scores range from 62.7% in 2016 to 78.5% in 2019. Both metrics, together with the F1 score, are shown in the lower line graph. The SVM achieves even higher balanced test accuracy as shown in the bar graph in Panel (b). The mean accuracy ranges from 69.9% to 83.5%, again tapering off over time. The model's precision and recall are also high, shown in the line graph at the lower half of Panel (b). The mean precision, excepting the precision for 2019, ranges from 74.4% to 80.2% and the mean recall from 65.1% to 84.4%. Overall, the SVM attains a fairly high F1 score, excepting the F1 score for 2019, at around 76% for 2017 and 2018.

The results for the neural net shown in Panel (c) are close to the SVM model, albeit a little lower. The mean balanced accuracy drops from 82.1% in 2015 to 69.2% in 2019. The mean precision and recall are also close to the SVM and exhibit a similar pattern over time.

The results so far showed the mean metrics across all predicted SDGs. The prediction performance naturally varies across SDG. Exhibit 8 individually shows the test metrics for each SDG prediction in the SVM model for 2017 as an example. Although we can observe some variation across SDGs, the predictive performance of the SVM remains relatively high across SDGs. F1 scores are all equal to or above 70%, and SDG 3, SDG 9, SDG 12, and SDG 15 reach F1 scores of at least 80%.

Case Examples

In this section, we provide case examples of companies that the algorithm predicts as aligned with certain SDGs. This serves as a "back-of-the-envelope" sense check to whether the SDG alignment prediction of our models have intuitive appeal. We randomly select four companies that have been consistently assigned to specific SDGs in every year of the sample period and therefore likely show strong alignment with the respective SDG.

The first example is Xcel Energy, a utilities and energy company headquartered in Minneapolis. This company is considered as one of the industry leaders in renewables and was one of the first power companies in the U.S. to announce that it would provide carbon-free energy by 2050. Its CSR report highlights large scale carbon emission reductions, innovation projects and partnerships with technology companies, reduced fresh water use and waste water production, and an emphasis on protecting wildlife during construction of wind energy projects. The company has also been recognized through regional and national awards for its environmental and climate leadership. Accordingly, our models consider the company aligned with SDG 6 (clean water), SDG 7 (clean energy), SDG 9 (industry and innovation), SDG 13 (climate action), SDG 15 (life on land), and SDG 17 (partnerships).

Another example is the Ford Motor Company. The models align Ford with SDG 3 (good health), SDG 6 (clean water), SDG 8 (decent work), SDG 9 (industry and innovation), SDG 12 (responsible consumption and production), SDG 13 (climate action), and SDG 16 (strong institutions). The company is the only U.S. car maker on the Carbon Disclosure Project's "A list for Climate Change" and "Water."[12] Inspection of Ford's CSR reports reveals emphasis on employment, employee well-being and training, support of health initiatives, and low incident rates at assembly facilities, as well as on strong supplier relationships (e.g., investmenting into minority- and women-owned companies). The reports further highlight the importance of respecting human rights, supply chain impact, and responsible materials sourcing.

Finally, our sample includes First Horizon Corp, a financial institution headquartered in Tennessee. The community bank emphasizes its commitments to a high minimum wage, as well as its investments into affordable housing, small businesses, and supplier diversity. The bank also manages a foundation that provides grants and engages in philanthropy to support education and financial literacy. The models consider the bank aligned with SDG 1 (no poverty), SDG 4 (quality education), SDG 11 (sustainable cities and communities), and SDG 17 (partnerships).

LIMITATIONS AND FUTURE RESEARCH

Overall, the "case check" in the previous section suggests that the model predictions result in classifications that are intuitive and consistent with companies' disclosures of their ESG activities. The case discussion also highlights, however, that our machine learning approach learns from the companies' own disclosures. That is, a major caveat of this approach is that the models rely on the companies' own descriptions to substantiate their SDG alignment. As our models do not rely on the sentiment of the disclosures, model predictions are less likely to be influenced by the tone of companies' descriptions. However, if companies are not entirely truthful about the *extent* of their ESG activities, our models might suggest SDG alignment, where in fact this may be overstated. For example, if companies that do not contribute to the SDGs mimic those that do in their descriptions in their CSR reports, the algorithms would possibly identify them as aligned. Moreover, we rely on Refinitiv Asset4 ESG scores, which we use as "ground truth" to train our models to be reliable indicators of companies' SDG-related activities.

An alternative to using companies' own disclosures is to rely on descriptions in news or NGO (non-governmental organizations) reports. These sources may be less likely to be positively biased; however, third parties likely possess less information about a company's SDG-related activities than company insiders. Our approach combines companies' own information with third-party ratings as a compromise.

A related second limitation is that this approach does not take into account negative contributions to the SDGs, and only considers how far particular activities of a company (as described in their CSR reports) are associated with certain SDGs by the nature of the activity. In other words, if a company is aligned with a subset of the SDGs (e.g., by using clean energy and recycling its waste materials), but negatively contributes on another SDG dimension (e.g., by paying low wages or exposing its workers to toxic materials), the models will only pick up the former—as they will likely be mentioned in company disclosures—but not the latter. Again, this is because the models measure alignment on a binary scale (with a zero lower bound) and not contribution on a continuous scale. Furthermore, to the extent that a company is negatively

[12] CDP recognizes companies on its "A List" that are leaders on the disclosure and management of climate change and water security risks.

contributing to the SDGs by the nature of its business model, or its products, our models may overestimate SDG alignment. One approach to address this limitation is to apply an exclusion screen to the company universe before applying the model.

More importantly, to gain a deeper understanding of how companies contribute to the SDGs, the models should measure a company's performance against targets set for each SDG and how a company improves on those targets over time. Although this is currently beyond the scope of this article, the algorithms could nevertheless help define and narrow the universe of eligible companies according to investors' preferences, making an initial cut of which companies seem to be at least qualitatively aligned with which SDG. Subsequently, human analysts could conduct a more detailed analysis on this narrower set of candidates to make further judgements about and quantify companies' contributions to the SDGs of interest.

Another, more advanced, algorithmic approach would be to apply a continuous scale (allowing for negative values) to measure the extent of a company's contribution, instead of simply using a binary scale for alignment. Several options are available in principle. In our current approach, we do not make use of the variation in ESG scores to measure the extent to which companies contribute to the respective SDG goal. As discussed in the earlier section, SDG Dictionary and ESG Scores, varying the ESG score threshold enables investors to vary the strictness with which companies are judged on their SDG alignment, resulting in a smaller investable universe. Another option would be to modify the prediction models from a classification problem (predicting SDG alignment) to a regression problem (predicting the extent of the contribution) to allow companies to be judged on a continuous scale (including negative values).

Perhaps the ideal alternative is to develop quantitative metrics to measure a company's contribution to the SDGs from the text, as well as quantitative information disclosed by the company and third parties. This is beyond the scope of this article, but could be explored in future research. Furthermore, from a methodological perspective, our results suggest that some SDGs are easier to predict than others, and we leave it to future research to examine the reasons for this variation in predictive performance across SDGs. We also find feature engineering (i.e., dictionary construction, choice of vocabulary, etc.) to play a more important role in determining predictive performance than the model choice. Nevertheless, future research might be able to improve on the models by using recent NLP breakthroughs of deep bidirectional transformer models using word representations that change based on the context, for example, BERT (Devlin et al. 2019). It is likely that with more accurate language modeling, we can improve on the F1 scores obtained in this article.

CONCLUSION

This article provides proof of concept for the use of machine learning and natural language processing to identify companies that are aligned with the UN SDGs. Although the SDGs are increasingly gaining the attention of the investment community, a paucity of disclosures by companies, whether and how they are contributing to the goals, and the lack of a clear mapping from ESG ratings to the SDGs has so far hampered investors' efforts to incorporate SDG alignment in asset allocation decisions.

The sustainability-minded investor needs to know how their portfolio and prospective investee companies are measuring against the SDGs to construct their portfolios closer to their sustainability preferences and to identify opportunities and risks to long-term investment returns. Currently few systematic and scalable methods exist, however, that could provide investors with the necessary information to understand public companies' contributions to the SDGs.

This study shows that advanced natural language processing methods can be applied to measure companies' alignment with the UN SDGs based on the text in their sustainability disclosures. We test these methods to CSR disclosures of Russell 1000 companies from 2010–2019. Specifically, we use word embeddings to augment dictionary-based input features, as well as features themselves, based on state-of-the art Word2Vec and Doc2Vec models to classify companies' alignment with the SDGs over time. Using a logistic classifier, SVM, and a fully-connected neural network, we find the SVM with Doc2Vec embeddings result in the highest average accuracy for predicting alignment.

Notwithstanding several limitations of this approach, the resulting SDG assignments could be used in portfolio construction to select companies that satisfy specific sustainability preferences of investors, for example, by selecting companies that are aligned with the SDGs important to the investor. With some modifications, they can also be used for measuring as to what extent existing portfolios and indexes are aligned with specific SDGs. We further provide several suggestions on how to build on this initial proof-of-concept to not only measure alignment more reliably, but also gauge the extent of companies' (positive and negative) contributions to the SDGs. Future refinements of this approach might allow investors to measure their portfolios' social and environmental impact—something that has proven challenging for public market investments.

APPENDIX

A1 WORD2VEC

Word2Vec is a so-called self-supervised machine learning algorithm developed by Google in 2013 (Mikolov et al. 2013). It is both unsupervised, because the input data is unlabeled, and supervised, because the input data itself provides enough context to infer the labels. Word2Vec is trained using a shallow fully-connected neural network to learn word embeddings. In practice, two model architectures are used, a continuous bag-of-words model (CBOW) and a skip-gram model with negative sampling. The former's objective is to predict a missing center word given surrounding words, and the latter learns to predict words surrounding a given center word. The intuition behind the model is that if words are frequently surrounded by a similar sets of words when used in sentences, then those words tend to be related in their semantic meaning. In this article, we train a CBOW model with negative sampling using the 'gensim' library in Python.

The inputs to the model are context words over sliding windows, where the window size is a hyperparameter of the model, which we set to five words. The word embeddings are extracted from the learned weight matrices that are the product of the prediction task. The size of the word embeddings is also a hyperparameter, which, in practice, ranges from a few hundred to a few thousand. Using higher dimensions captures more nuanced meanings, but is more computationally expensive to train. We set the embedding size to fifty.

To perform the learning task, the corpus of pre-processed CSR reports is transformed into word vector representations that are fed into the neural network. The inputs of the CBOW are one-hot representations of word vectors averaged over the sliding window and horizontally stacked into a matrix of size (vocabulary $\times$ sample size). In the CBOW, the center word is coded in one-hot representation also stacked horizontally, such that each column in the input matrix with the context words corresponds to the column in the output matrix with the center word.

The CBOW model is based on a shallow dense neural network with an input layer, a single hidden layer, and output layer. The number of neurons in the hidden layer corresponds to the dimensions of the word embeddings chosen. We use a rectified linear

EXHIBIT A1

Text Cleaning Process

Original Text

Our Credo has\nguided our actions in fulfilling our responsibilities to our \ncustomers,employees,communities,and stockholders since\n1943.In formulating these principles,General Robert Wood\nJohnson was ahead ofhis time.He recognized that our\nCompany's financial success depends on our ability to protect\nthe environment,respect our employees and be responsible to\nthe world community.In fact,that is sustainability:ensuring\nthat our customers,our employees,the communities in \nwhich we operate and the environment on which we depend,\nthrive with us.

Text after Cleaning and Lemmatization

credo guided action fulfilling responsibility customer employee community stockholder since. formulating principle general robert wood johnson ahead of his time recognized company financial success depends ability protect environment respect employee responsible world community. fact sustainability ensuring customer employee community operate environment depend thrive throughout history.

NOTE: This exhibit shows an example of the result of the text cleaning and lemmatization.

unit (ReLU) as the activation function for the hidden layer and a softmax function for the output layer. The loss function is the cross-entropy loss

$$J = -\sum_{k=1}^{V} y_k log\left(\hat{y}\right) \tag{4}$$

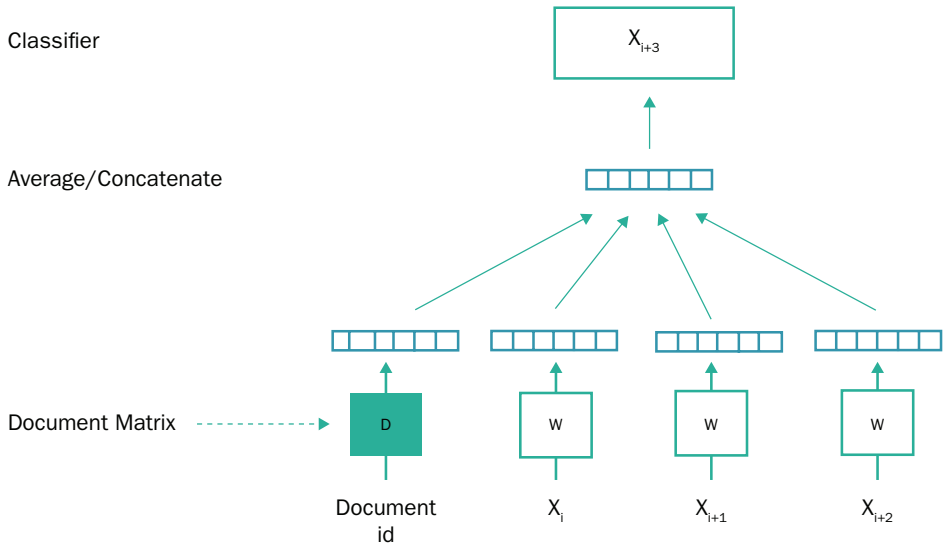
that is averaged for the entire set of examples/batches to derive the cost. The word embeddings are the outcome of the minimization of the cost in the form of the optimal input weights matrix.

A2 DOC2VEC

A Doc2Vec model, in principle, is set up and trained in the same manner as the Word2Vec model with the difference that, in addition to each word, each document is also mapped to a unique word vector (Le and Mikolov 2014). At each step in the training process, the Doc2Vec algorithm uses the document and word vectors as inputs to a shallow dense neural network as in the Word2Vec case. The document D and word vectors $[w_{T-k}, \ldots, w_{T-1}]$ are either concatenated or averaged into a context vector to predict the target word w_T in the context. The document vectors act as a memory of the context in which each word appears and are shared across all context vectors generated from the same document. The word vectors, on the other hand, can be shared across documents. Exhibit A2 shows a simplified framework for training process of the context word.

EXHIBIT A2

Doc2Vec



NOTES: This exhibit illustrates Doc2Vec. *D* represents the document or paragraph vector, *W* represents the word vectors, and x_{i+t} is a particular word in the sequence of words.

A3 LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO)

The least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996) is a linear regression and was introduced to avoid model overfitting by selecting only a subset of coefficients in the final prediction model. Similar to ridge regression, it adds a penalty (regularization) term to the cost function of regular OLS with the aim to reduce the size of the coefficients from the optimization. While ridge regularization adds the *L*2 or Euclidean norm, LASSO adds the *L*1-norm, that is, the sum of absolute values, which has a tendency to push the coefficients to zero. LASSO solves the following optimization problem:

$$\min_{\theta} J(\theta), \text{ where} \tag{5}$$

$$J(\theta) = \frac{1}{2n}\left(\sum_{i=1}^{N}(h_\theta(x^{(n)}) - y^{(n)})^2 + \lambda|\theta|\right) \tag{6}$$

with the regularization parameter $\lambda \geq 0$. The parameter λ controls the number of selected variables in the model. If $\lambda = 0$, the Lasso regression retains the same coefficients as OLS. As λ increases, the optimization penalizes the coefficient more heavily and fewer independent variables remain with a coefficient larger than zero.

A4 RANDOM REGRESSION FOREST

A random regression forest is a supervised ensemble learning approach that combines multiple regression trees by bootstrapping the training samples (Breiman 2001). A regression tree is a hierarchical structure where, at every "node" of the tree, the data is split into subsets based on a threshold value of a variable. During the construction of

the tree, starting from the root node, every split into branches of region R_j is chosen to minimize the mean squared error (MSE):

$$\sum_{j=1}^{J}\sum_{i\in R_j}\left(y_i - \widehat{y_{R_j}}\right)^2. \tag{7}$$

To avoid overfitting the training data, for example by splitting it into a very deep tree in which every resulting leaf only contains few sample observations (in the extreme, only one), the regression trees are pruned using cross-validation.[13]

To build a random forest, a random sample of all possible predictors is chosen as split criteria in each tree. This ensures low correlation among the trees that make up the forest, reducing the variance of the prediction. The two hyperparameters for the random forest are the number of regression trees that it consists of and their maximum depth. We choose 100 trees and allow the trees to expand until less than five samples remain in a branch. The random forest allows us to measure a predictor's importance by evaluating the reduction in MSE at every split that resulted from using that particular predictor as the splitting criterion averaged over all trees. This measure allows us to identify the most important tokens that improve the predictions of ESG scores. The mathematical construction of this variable importance measure is explained in Breiman et al. (1984).

ACKNOWLEDGMENT

We would like to thank Anastasia Demina for her excellent research work.

REFERENCES

Amel-Zadeh, A., R. Lustermans, and M. Pieterse-Bloem. 2020. "Sustainability and Private Wealth Investment Flows." Available at www.supremecourt.gov/opinions/19pdf/19-7_n6io.pdf.

Amel-Zadeh, A., and G. Serafeim. 2018. "Why and How Investors Use ESG Information: Evidence from a Global Survey." *Financial Analysts Journal* 74 (3): 87–103.

Berg, F., J. F. Kölbel, and R. Rigobon. 2020. "Aggregate Confusion: The Divergence of ESG Ratings." Available at https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3438533.

Blasco, J. L., A. King, and S. Jayaram. 2018. "How to Report on the SDGs." KPMG. Available at http://assets.kpmg/content/dam/kpmg/xx/pdf/2018/02/how-to-report-on-sdgs.pdf.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Taylor & Francis.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "Smote: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–357. http://dx.doi.org/10.1613/jair.953.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Confer ence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1: 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423.

[13] The cost function of the complexity pruning of regression trees includes a regularization parameter similar to LASSO.

Dimson, E., O. Karakaş, and X. Li. 2020. "Coordinated Engagements." Available at https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3209072.

Dimson, E., P. Marsh, and M. Staunton. 2020. "Divergent ESG Ratings." *The Journal of Portfolio Management* 47 (1): 75–87.

Dyck, A., K. V. Lins, L. Roth, and H. F. Wagner. 2019. "Do Institutional Investors Drive Corporate Social Responsibility? International Evidence." *Journal of Financial Economics* 131 (3): 693–714.

Gibson, R., and P. Krüger. 2018. "The Sustainability Footprint of Institutional Investors." *Swiss Finance Institute Research Paper* (17-05).

Global Sustainable Investment Alliance. 2020. "Global Sustainable Investment Review 2020." http://www.gsi-alliance.org/wp-content/uploads/2021/08/GSIR-20201.pdf.

He, Y., B. Kahraman, and M. Lowry. 2020. "ES Risks and Shareholder Voice." https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3284683.

Kingma, D. P., and J. Ba. 2014. "Adam: A method for Stochastic Optimization." *arXiv preprint arXiv:1412.6980*.

Krüger, P., Z. Sautner, and L. T. Starks. 2020. "The Importance of Climate Risks for Institutional Investors." *The Review of Financial Studies* 33 (3): 1067–1111.

Le, Q., and T. Mikolov. 2014. "Distributed Representations of Sentences and Documents." *Proceedings of the 31*st *International Conference on Machine Learning*, Beijing, China, 1188–1196.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vctor Space." *arXiv preprint arXiv:1301.3781*.

Schramade, W. 2017. "Investing in the Unsustainable Development Goals: Opportunities for Companies and Investors." *Journal of Applied Corporate Finance* 29 (2): 87–99.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.

World Economic Forum (WEF). 2019. *Global Financial Stability Report, Chapter 6: Sustainable Finance*, April 2019.

implementation of any specific investment program which cannot be fully accounted for in the preparation of model results and all of which can adversely affect actual trading results.

The information presented is based upon the hypothetical assumptions as a result of model generated scoring results discussed in this piece. Certain assumptions have been made for modeling purposes and may be unlikely to be realized. No representation or warranty is made as to the reasonableness of the assumptions made or that all assumptions used in achieving the returns have been stated or fully considered.

PanAgora is exempt from the requirement to hold an Australian financial services license under the Corporations Act 2001 in respect of the financial services.

PanAgora is regulated by the SEC under U.S. laws, which differ from Australian laws.