Technological Forecasting & Social Change 205 (2024) 123455

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

First impressions on sustainable innovation matter: Using NLP to replicate B-lab environmental index by analyzing companies' homepages

Pietro Cruciata *, Davide Pulizzotto, Catherine Beaudry

Department of Mathematics and Industrial Engineering, Polytechnique Montréal, 2500 Chem. de Polytechnique, Montréal, QC H3T 1J4, Canada

ARTICLE INFO

Keywords:
Zero-shot text classification
B-Corp data
Sustainability
Sustainable innovation
Natural language processing
Signal theory

ABSTRACT

This study explores the potential for developing web-based environmental culture indicators by analyzing signals extracted from the homepages of company websites. The primary aim is to assess the proposed method's ability to generate indicators that can serve as proxies for real environmental measures by leveraging the homepage content. We performed a Zero-Shot Text Classification (ZSTC) using a BERT-type Natural Language Processing (NLP) model, followed by a regression analysis to test the ability of these web-based indicators to replicate the B-Lab environmental index and comprehend the dynamics behind the results. This pilot study explains 57 % of the variance of the B-Lab environmental index using the results of the ZSTC score and companies' characteristics. This research makes two significant contributions. First, the text content of a company's homepage seems to provide insights into its environmental performance. Second, it introduces a generalizable methodology for studying the performance of companies through their websites without the need for heavy pre-processing, significantly reducing the time and cost of research. Furthermore, the method could provide policymakers with a real-time landscape to create and finetune policies about specific topics, partially addressing the problems associated with questionnaire-based surveys.

1. Introduction

Traditional innovation indicators built using public databases generally supplemented by questionnaire-based data are important sources of information for governments, academics, and the private sector. These sources of information are often incomplete (e.g., representative samples much smaller than the population of firms) or not specific, whereas questionnaire-based surveys (especially large-scale as the biennial European CIS or the annual MIP) lack regional granularity, coverage, timeliness, and more importantly, they are costly to run (Axenbeck and Breithaupt, 2021). Moreover, the number of low-cost web-based surveys sent to firms has sky-rocketed to the extent that obtaining a representative response rate has plummeted to lower than 5–10 % in most cases. For all these reasons, innovation indicators built using traditionally collected data hardly provide the full picture (Kinne and Lenz, 2021).

Alternative or complementary to these sources are web-based unstructured textual data. The increasing amount of data available in the form of digitalized text indeed offers new avenues for innovation studies. Among their noteworthy advantages, the rapidity of their evolution, their increasing quantity, variety, and availability opened new possibilities for policymakers and researchers (Gök et al., 2015). Although it seems difficult to measure and interpret "signals" of innovation dynamics in corporate websites or other web sources, researchers in innovation and technology management have obtained good results by building new indicators with large amounts of text. For example, Gök et al. (2015) created web indicators of R&D activities by extracting the keywords from companies' websites. Their study suggested that R&D activities captured through the web indicators were significantly more numerous, compared to the R&D activities documented in other sources. Libaers et al. (2016) harnessed the data from companies' websites to develop a taxonomy that identified strategies used by small firms to commercialize their innovations. The authors analyzed the content of firms' websites to extract the keywords related to possible strategies used by companies. Blazquez and Domenech (2018) used web-based variables built with keywords to predict firm export orientation. Héroux-Vaillancourt et al. (2020) built innovation indicators based on four core concepts (R&D, IP protection, collaboration, and external financing) from the complete texts of corporate websites of Canadian nanotechnology and advanced materials firms using keywords

* Corresponding author.
E-mail addresses: pietro.cruciata@polymtl.ca (P. Cruciata), davide.pulizzotto@polymtl.ca (D. Pulizzotto), catherine.beaudry@polymtl.ca (C. Beaudry).

https://doi.org/10.1016/j.techfore.2024.123455
Received 26 July 2023; Received in revised form 10 April 2024; Accepted 15 May 2024
Available online 8 June 2024

frequency analysis. Other researchers specifically studied different dimensions of sustainable development in companies through their environmental performances using their websites. For instance, Fernández-Vázquez and Sancho-Rodríguez (2020) analyzed texts and images from the websites of the Spanish IBEX 35 to investigate to which extent the companies address climate change in the construction of their reputational identity and to explore the types of narratives. Calabrese et al. (2021) examined the websites of 23 manufacturers from the fast-growing fitness equipment industry to study firms' strategies and their contribution to the SDGs.

All these pilot studies highlight the strong potential that these new sources of data bring to the field of innovation studies. Building on these encouraging findings, can we develop web-based environmental indices that mirror a real environmental index by analyzing the content of companies' homepages?

Companies' communications are divided into two main channels: external, towards clients and stakeholders; and internal, towards the workers who are part of the company. Through internal communication, a company expects to generate know-how necessary to fuel operational procedures, as well as the loyalty of employees, which motivates them to apply their expertise to the company's processes (Mazzei, 2010). External communication revolves around the company's relational network, serving to provide vital information to the business intelligence system, influence project specifications, facilitate industrial and financial package development, and foster trust with clients and partners (Goczol and Scoubeau, 2003). Therefore, an official website serves as a platform for conveying authentic, precise, and current information about companies, enabling visitors to make more informed decisions (Jiang et al., 2023). As a result, the information contained in a website provide a general understanding of the relevance of a particular element for the company (Héroux-Vaillancourt et al., 2020).

As policymakers shift their focus to adapt to climate change, mitigating its effects, and striving for a more positive socio-environmental impact through their policies, sustainable innovation (SI) emerges as a key solution. This paper examines the potential of developing web-based environmental culture indicators that analyze signals gleaned from the homepage of companies' websites. The primary objective is to explore the proposed method' capacity to create indicators as proxies for real environmental measures. This pilot study focuses on the environmental index of the B-Corp database, to evaluate the approach. It is one of 5 indices developed by B-Lab to assess various Environmental, Social, and Governance (ESG) dimensions of companies. The B-Corp certification has gained recognition in helping organizations stand out in the 'green revolution' (Kim and Schifeling, 2016, p. 32), establishing legitimacy (Blasi and Sedita, 2022; Cormier and Magnan, 2015), and projecting an authentic commitment to triple bottom line (TBL) practices (Cao et al., 2017; Kim and Schifeling, 2016). The database is therefore particularly well suited for our purposes: measuring the correlation between a company's website and the B-Lab indicator is the main goal for this pilot study.

The methodology comprises two steps: first, a Zero-Shot Text Classification (ZSTC) score is obtained using a BERT-type Natural Language Processing (NLP) model to extrapolate and study the environmental signal of each company's website; second a regression model is estimated to evaluate to what extent these web-based environmental culture indicators (the signal) explain the value of the environmental index attributed by B-Lab to the company. The results of the ZSTC score together with the companies' characteristics explain 57 % of the variance[1] of the B-Lab environmental index obtained by companies, thereby showing great promise for the proposed method.

The remainder of the paper is organized as follows. Section 2 presents the pertinent literature on sustainable innovation and signal theory. Section 3 describes the data collected and explains the methodology. Section 4 analyzes the results of the ZSTC, the correlation between the ZSTC scores and the B-Lab environmental index, and the Principal Component Analysis. Section 5 presents the Ordinary Least Squares (OLS) regression results while Section 6 discusses their implications. Finally, Section 7 concludes and highlights the limitations of the research and possible future works.

2. Literature review

The widely accepted viewpoint that innovation is driven solely by the combination of scientific research, technological advancements, their implementation by businesses, and distribution in the market, has evolved considerably. Innovation is no longer solely about enhancing market competitiveness and advancing technology in various industries. Instead, it is increasingly seen as a means to address social issues, improve quality of life, and enhance overall societal and environmental health. For instance, policymakers are now working to define and support the concept of SI, among other ideas linked to environmental, social, and governance (ESG) considerations. The origin of the SI concept can be dated back to the publication of the "Brundtland Report" (WCED, 1987), in which the World Commission on Environment and Development (WCED) coined the term *Sustainable Development*, which the report defined as "development that meets the requirements of the present without jeopardizing future generations' ability to meet their own needs" (Zhu and Hua, 2017, p. 893). Over time, governments have gradually placed greater emphasis on reducing the environmental footprint of economic activities. It was previously believed that economic objectives and environmental concerns were incompatible, but this notion was challenged by Weale's paper (1992). Moreover, the "triple bottom line" concept introduced by John Elkington in the 1990s has become the cornerstone of sustainable development. This concept seeks to harmonize environmental, economic, and social performance – a challenge that businesses must now address (Bossle et al., 2016).

Furthermore, the belief that companies can simultaneously pursue economic, environmental, and social goals has been reinforced by shifts in customer demands and stakeholder requirements. These changes are exerting increasing pressure on companies to implement sustainable initiatives and to measure, monitor, and report on sustainability performance. Customers and stakeholders are showing a growing interest in sustainable brands, with ethical and sustainable certification becoming a crucial factor that consumers consider when making purchasing decisions. Additionally, studies have shown the growth of B-Corp businesses after obtaining certification (Romi et al., 2018; Paelman et al., 2020). This underscores the importance of external communication in attracting customers.

In this context, several studies within the field of signal theory shed light on how companies strategically use their official websites to shape stakeholders' perceptions (e.g., Mavlanova et al., 2012; Yildiz et al., 2023). Signal theory defines a "signal" as an action initiated by a better-informed party in situations characterized by information asymmetry. The purpose of this signal is to effectively and credibly communicate the party's true characteristics to a less-informed counterpart (Connelly et al., 2011). Scholars in management have leveraged signal theory to elucidate the impact of information asymmetry across a variety of research domains. For instance, Mavlanova et al. (2012) conducted a study on the role of website signals as a means for online retailers to communicate their product quality, proposing and validating a three-dimensional framework. Jiang et al. (2023) argued that a corporate official website serves as a credible source of non-financial information for assessing the credit risk of Small and Medium Enterprises (SMEs). SMEs equipped with comprehensive official website information are less likely to default and are better positioned to secure financial support for further development. Yildiz et al. (2023) found that the presence of a "green label" on a hotel enhances the trustworthiness of the eco-conscious tourist brand. Lastly, Eccles et al. (2014) concluded in their research that companies with a strong emphasis on sustainability

[1] This is simply measured by the R-squared value of the regressions.

demonstrate increased levels of information transparency and accountability. Based on these findings, we put forth the following proposition:

Proposition 1. Environmental compliance indices are positively correlated with the environmental culture indices built using the text contained in companies' websites.

Several studies have highlighted the impact of both internal and external factors on companies' sustainability efforts (Hermundsdottir and Aspelund, 2021). In addition to the pressure from stakeholders and customers, national regulations, incentives, society's awareness, industrial norms, and regulations are some of the several factors that might directly impact companies in their pursuit of sustainable initiatives (Hermundsdottir and Aspelund, 2021). Doran and Ryan (2012) discovered that regulation and industrial agreements significantly influence a firm's decision to engage in eco-innovation. Aguilera-Caracuel and Ortiz-de-Mandojana (2013) proposed that policymakers play a crucial role in a firm's ability to transform SI into competitive advantages. Moreover, the authors suggested that countries with stricter environmental regulations tend to have a higher prevalence of green innovative firms. Finally, de Azevedo Rezende et al. (2019) demonstrated that there are differences in green innovation performances between Europe and North America due to their distinct approaches to regulations. Given the variations in performance and in the willingness to pursue SI highlighted in the literature, we suggest the following proposition:

Proposition 2. The country in which a company is located influences its environmental compliance index.

Additionally, Magnusson et al. (2011) suggested that the reputation of a brand's country of origin serves as a conspicuous and consistent signal that can shape consumer perceptions of corporate brand reputation. On the other hand, corporate brands originating from countries with more favorable sustainability reputations may not experience the same benefits from engaging in corporate social responsibility (CSR) or sustainability efforts, as these reputation-building strategies may be expected (Cowan and Guzman, 2020). Thus, we posit the following proposition:

Proposition 2m. The relationship between a company's web-based environmental culture indicators and its environmental compliance index is moderated by the country in which it is located.

It is reasonable to assume that a company's size may also affect its propensity to pursue sustainable initiatives. Aguilar-Fernández and Otegi-Olaso (2018) investigated the impact of size on the likelihood of firms to pursue SI, concluding that there is no consensus on this impact. On the one hand, large companies may have advantages in pursuing SI from both a supply chain and financial perspective. SMEs, on the other hand, may have more flexibility to adapt and change their business models. Additionally, large companies may face more pressure from stakeholders to achieve socio-environmental goals due to their greater exposure. Furthermore, the lack of resources and capacities may be a limit for SMEs (Aguilar-Fernández and Otegi-Olaso, 2018). For instance, Ketata et al. (2015) highlighted the positive impact of firm size in their study on SI in Germany. De Azevedo Rezende et al. (2019) also identified differences in green innovation performance according to company size in their analysis. The interplay between a company size and its digital communication strategy has been a focal point of various studies. In this regard, Kinne and Axenbeck (2020) found a correlation between the size of a company and the number of pages of its website. In a complementary vein, Callison (2003) posits that companies with larger market cap often possess greater financial and professional resources, which they can leverage to enhance their web presence. This perspective is further corroborated by the research of Jung Moon and Hyun (2014), who observed that large firms tend to have robust marketing teams dedicated to the upkeep of their websites. In light of the evidence presented, our proposition is as follows:

Proposition 3. The size of a company influences its environmental compliance index.

Furthermore, Hoehn-Weiss and Karim (2014) shed light on the advantages that young firms gain when they signal alliances with larger partners. This strategy can attract the general market and make an Initial Public Offering (IPO) a more appealing option than an acquisition. Given this intriguing finding on the signaling of small companies, we posit the following proposition:

Proposition 3m. The relationship between a company's web-based environmental culture indicators and its environmental compliance index is moderated by its size

As previously mentioned, it is commonly understood that economic benefits drive companies to adopt sustainable behavior, and the sector in which a company operates is also significant. De Azevedo Rezende et al. (2019) demonstrated differences in green innovation performance between manufacturing and non-manufacturing companies. The latter face more challenges in implementing green technologies in sectors such as services or information, while for manufacturing companies, green innovation can attract clients or increase efficiency, thereby generating a direct economic impact. Hermundsdottir and Aspelund (2021) highlighted how some industries adopt sustainable practices as standard, while others respond differently to environmental obstacles to SI. For these reasons, it is commonly asserted that the impact of various factors on the development of environmentally friendly products and processes varies depending on the industrial sector under examination. Numerous research endeavors have acknowledged that their conclusions are applicable solely within a specific industry, and have specified that their outcomes are limited to the context of that industry (Tariq et al., 2017). For the reasons mentioned above, the following proposition is suggested:

Proposition 4. The industrial sector in which a company operates influences its environmental compliance index

Moreover, as previously mentioned, Yildiz et al. (2023) provided empirical evidence of how environmental efforts signaled by hotels are advantageous for eco-tourists. The trust instilled by the "green label" significantly mediates the perceived green risk in online booking intentions within the hotel sector. This research exemplifies how a company's signaling approach is personalized and dependent on the specific needs of the sector in which the company operates (Fig. 1). In line with this finding, we formulate the following proposition:

Proposition 4m. The relationship between a company's web-based environmental culture indicators and its environmental compliance index is moderated by the industrial sector in which it operates

3. Data and methodology

3.1. Data

Two types of data are required to assess whether new web-based environmental culture indicators are good proxies for more traditionally built environmental compliance or certification indicators. For the latter, we selected the certification of the B-Corporation, henceforth

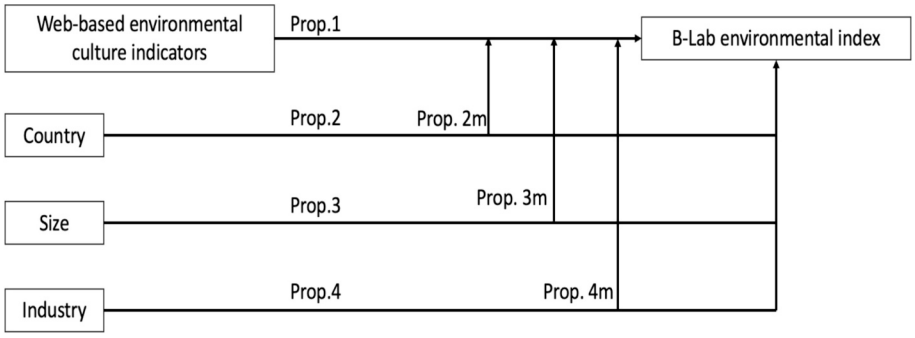

Fig. 1. Summary of the propositions to be tested.

referred to as B-Corp, which is a type of for-profit corporation that has been certified by the non-profit organization B-Lab to meet certain standards of social and environmental performance, accountability, and transparency. These standards are set by the B-Lab and are verified through a rigorous assessment process. The B-Corp Certification is comprehensive and adopts a holistic approach to environmental, social, and governance (ESG) issues. Furthermore, obtaining and maintaining accreditation is a rigorous procedure that involves teams and departments from across the organization. B-Corp firms are committed to making a positive impact on society and the environment, and to conducting in a way that is transparent, accountable, and sustainable. Unlike traditional corporations, B-Corps are required to consider the impact of their decisions on their employees, customers, suppliers, community, and the environment.

B-Corp data includes one main index, "overall score", which is an aggregation of five other indices evaluating specific dimensions: governance, customers, workers, community, and environment. These dimensions are further divided into several items. In this paper, we focus on the B-Corp indicator concerning the "impact area environment". We refer to this environmental compliance index variable "B-Lab environmental Index" (*BLabEnvIndex*). The B-Lab Environmental Index is derived from the B Impact Assessment, a tool that assesses a company's social and environmental performance. Specifically, the B-Lab environmental Index evaluates a company's overall environmental stewardship. This includes how the company manages general environmental impacts, air and climate issues, water sustainability, and impacts on land and life. The scoring system used in the assessment allows for comparability across companies and identifies areas for improvement over time. The scoring criteria are customized and evolve with each version of the assessment, based on the specific track of the company being evaluated. We incorporated control variables related to two other specific ESG areas, community and governance, because the minimum score required to pass the assessment is determined by the sum of the scores for the aforementioned five ESG areas. This means that a company could potentially invest more in other areas than the environmental one and still qualify as a B-Corp (Liute and De Giacomo, 2022). Additionally, we introduced a dummy variable representing the assessment year—the year in which the company completed the B-Lab test designed to measure, manage, and enhance positive impact performance for the environment, communities, customers, suppliers, employees, and shareholders. The control variables related to the assessment year account for changes that can affect the assessment test[2].

This pilot study uses only a subset of the B-Corp data limited to Canadian and US companies. The reason for this choice is straightforward. We aimed to ensure a certain degree of homogeneity in the sample considered, i.e., that the results are not affected by widely different national systems, or languages, or by considering several dimensions simultaneously. As of March 2022, B-Corp had 8799 certifications from 5631 companies (the certification lasts 3 years) across 86 countries.

Fig. 2[3] illustrates the primary steps of our data collection process. As mentioned earlier, we began with 8799 certifications of companies[4], narrowing down our selection to 1741 certified companies in Canada and the USA. For textual data collection, we utilized the website URLs provided by B-Corp, leveraging the Wayback Machine, a web archive tool. This allowed us to obtain snapshots of company homepages corresponding to the certification years between 2007 and 2022 for each Canadian or US company in the B-Corp data. The objective was to collect data from the company's website for the year of the assessment recorded in the B-Corp database. In instances where a company's snapshot for the specific year was unavailable, we gathered data within a three-year range, encompassing the target year as well as the preceding and following years. We extracted the text from the homepages using specific HTML tags (i.e., <p>, <li>, <h1>, <h2>, etc.), effectively matching the companies with their respective snapshots and reducing the sample to 1256. Subsequently, we filtered out snapshots deemed irrelevant based on the following criteria: non-English websites[5], instances with less than one sentence of text, and manual removal of snapshots resulting from errors in the Wayback Machine. Our final sample comprises 1110 companies, with 195 of them being Canadian firms.

3.2. Methodology

Once the data is prepared, the first step of the analysis involves "understanding" the text of the corporate websites. Instead of counting specific keywords related to predetermined topics, as most of the literature mentioned in the introduction does, we employ the Zero-Shot Text Classification (ZSTC) method. This method is a Natural Language Processing (NLP) task designed to answer the question: "Is this text about label X?". The response to this question serves as an indicator of the confidence that the given text pertains to the label X. The 31 labels used for this purpose correspond to the names of the 31 items that compose the B-Corp environmental certification.

Within the realm of the Natural Language Understanding (NLU),

[2] See https://kb.bimpactassessment.net/support/solutions/articles/43000547789-overview-of-changes-in-version-6-of-the-b-impact-assessment, accessed on 26th Nov. 2024.

[3] To assess the comparability between the 1256 companies with a snapshot in the online archive and the 485 excluded companies, we conducted a non-parametric Mann-Whitney Anova test. The results revealed no significant differences in averages between the two groups. Furthermore, we performed the same test between the final sample of 1110 companies and the 631 eliminated companies (due to no snapshot found and specific criteria such as no more than a sentence, non-English language, and manual elimination for lack of meaningful content).

[4] Some companies have been certified multiple times.

[5] Our methodology will use a NLP model that is only trained in English document.

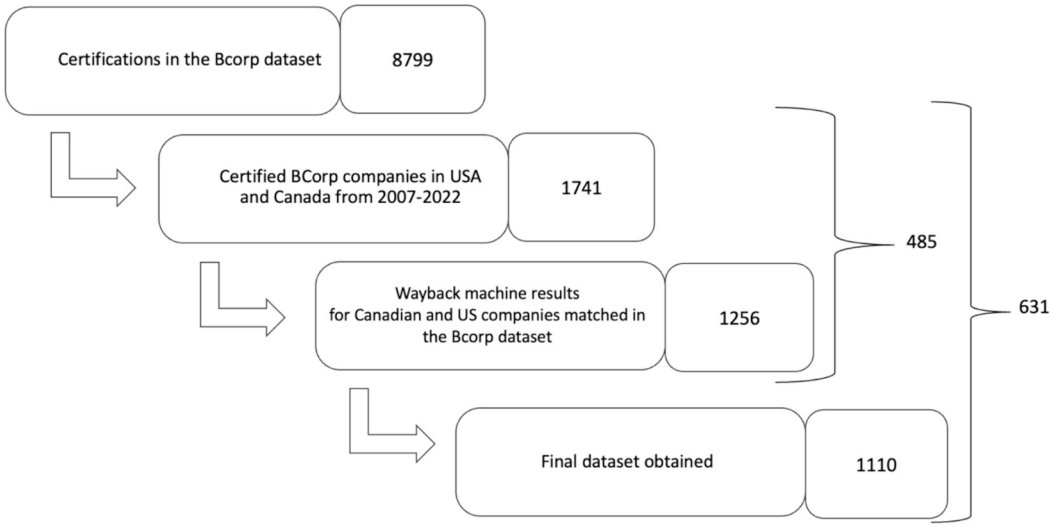

Fig. 2. Pre-processing steps.

ZSTC is a challenging task that necessitates the use of syntactic and semantic analysis to comprehend the actual meaning and sentiment of human language. More specifically, ZSTC refers to a task where the model classifies text into classes that were not present in the training corpus. In other words, ZSTC aims to associate an appropriate label with a piece of text, regardless of the text's domain or predefined label categories. ZSTC was initially applied in a Dataless Classification scenario, similar to the problem we are currently addressing, where it was used to select the appropriate label for a text through Explicit Semantic Analysis. With the rise of word embeddings, various approaches have been proposed for this purpose. For instance, generative Long Short-Term Memory has been used to generate text given the vector labels, and the vector representation of the label has been used to represent the text in multilabel classifiers (Yin et al., 2019).

The core of ZSTC is the NLP model "Bidirectional and Auto-Regressive Transformers" or BART (Lewis et al., 2019), a transformer-based deep learning model for NLP developed by Facebook AI. This model combines the most significant characteristics of BERT[6] and GPT[7]. BART was pre-trained on the English Wikipedia and BooksCorpus, using a two-step processes: first, the text is altered by adding a noise factor (e.g., changing the words randomly); then, the model learns to reconstruct the original text. This innovative approach allows BART to reach state-of-the-art performances in several NLP challenges. Indeed, BART excels in text generation, but it has also been tested in a wide range of tasks, including discriminative tasks such as General Language Understanding and the Stanford Question Answering Dataset (Lewis et al., 2019).

Performing the ZSTC requires the selection of both the labels and the corpus. Since our aim is to create an environmental culture indicator, we utilized the labels of the items that constitute the "impact area environment" index of the B-Corp data (Table 1). After experimenting with several settings, we decided to divide each website into groups of three sentences to create the corpus[8]. This decision was based on the observation that the ZSTC performed better when the input was a text longer than a single sentence but shorter than the full homepage. Consequently, for each website, we applied the ZSTC on each group of sentences. Then, to prepare the results for the subsequent Pearson correlation test, we calculated the average ZSTC scores for each label listed in Table 1, for each website.

Table 1
Labels used in the Zero-Shot Text Classification.

- Air climate
- Certification
- Community
- Construction practices
- Designed to conserve agriculture process
- Designed to conserve manufacturing process
- Designed to conserve wholesale process
- Energy water efficiency
- Environment products services introduction
- Environmental education information
- Environmental management
- Environmentally innovative agricultural process
- Environmentally innovative manufacturing process
- Environmentally innovative wholesale process
- Green investing
- Green lending
- Inputs
- Land life
- Land office plant
- Land wildlife conservation
- Material energy use
- Materials codes
- Outputs
- Renewable energy
- Cleaner burning energy
- Resource conservation
- Safety
- Toxin reduction remediation
- Training collaboration
- Transportation distribution suppliers
- Water

Source: https://data.world/blab/b-corp-impact-data/workspace/data-dictionary.

It is cruciate to note that we configured the multilabel output for the ZSTC. When the output is multilabel, the ZSTC generates a score among the class using cosine similarity metrics between the word-embedding vectors created by BART, representing the label and the target corpus. The cosine similarity is a metric in a range from −1 to 1, where a value of 1 signifies identical vectors, 0 indicates they are orthogonal (i.e., completely dissimilar), and − 1 implies they are diametrically opposed. Generally, the closer the cosine similarity is to 1, the more similar the vectors are to each other. In this task, the corpus and the label are transformed into vectors using BART's word embedding representation, and then the cosine similarity is calculated. Consequently, when we ask the model "Is this text about label X?" The score within the 0 to 1 range can be identical for multiple labels.

To examine the correlation between the web-based environmental culture indicators and B-Lab environmental index (the environmental compliance index of the propositions), we calculate Pearson correlations. This calculation measures which different items of the B-Corp environmental certification, detected with the ZSTC in the text of the corporate websites, serves as good proxies for the environmental score obtained by these firms. From the list of labels in Table 1, we expect the ZSTC to generate a series of indicators that are likely to correlated to various extents. From this point, we want to comprehend the different

[6] Bidirectional Encoder Representations from Transformers (Devlin et al., 2018).
[7] Generative Pre-Training (Radford et al., 2018).
[8] We use the python package spacy for this purpose.

Table 2
Zero-Shot Text Classification (ZSTC) results.

Labels	Mean	Std. dev.	Min	25 %	50 %	75 %	Max
Inputs	0.435	0.107	0.020	0.375	0.445	0.505	0.793
Outputs	0.103	0.104	0.000	0.046	0.076	0.127	0.969
Community	0.104	0.082	0.000	0.043	0.087	0.145	0.668
Designed to conserve wholesale process	0.698	0.169	0.042	0.602	0.710	0.824	0.996
Land office plant	0.138	0.168	0.000	0.043	0.090	0.169	0.991
Designed to conserve manufacturing process	0.142	0.079	0.001	0.088	0.138	0.188	0.543
Green investing	0.073	0.065	0.000	0.033	0.060	0.095	0.945
Water	0.223	0.115	0.001	0.143	0.211	0.289	0.861
Training collaboration	0.153	0.106	0.001	0.079	0.135	0.207	0.754
Energy water efficiency	0.217	0.214	0.000	0.059	0.135	0.315	0.973
Green lending	0.080	0.076	0.000	0.020	0.058	0.117	0.512
Air climate	0.272	0.213	0.001	0.111	0.216	0.377	0.986
Designed to conserve agriculture process	0.212	0.101	0.000	0.145	0.203	0.270	0.777
Renewable energy	0.125	0.113	0.000	0.064	0.094	0.143	0.993
Construction practices	0.208	0.137	0.001	0.114	0.179	0.264	0.995
Land life	0.147	0.109	0.000	0.079	0.125	0.182	0.986
Environment products services introduction	0.100	0.104	0.000	0.033	0.074	0.133	0.848
Environmentally innovative wholesale process	0.243	0.130	0.000	0.158	0.224	0.306	0.873
Environmentally innovative manufacturing process	0.059	0.068	0.000	0.020	0.038	0.074	0.783
Material energy use	0.175	0.145	0.002	0.077	0.125	0.223	0.951
Certification	0.104	0.115	0.000	0.038	0.074	0.119	0.991
Cleaner burning energy	0.207	0.088	0.010	0.143	0.197	0.255	0.624
Environmental management	0.252	0.200	0.001	0.098	0.188	0.360	0.939
Resource conservation	0.259	0.147	0.004	0.149	0.237	0.351	0.898
Materials codes	0.259	0.146	0.002	0.155	0.238	0.342	0.952
Land wildlife conservation	0.059	0.046	0.000	0.027	0.049	0.078	0.404
Environmentally innovative agricultural process	0.291	0.131	0.010	0.202	0.273	0.366	0.905
Transportation distribution suppliers	0.208	0.122	0.004	0.127	0.182	0.265	0.924
Safety	0.316	0.152	0.002	0.205	0.291	0.398	0.979
Environmental education information	0.225	0.100	0.003	0.157	0.212	0.277	0.918
Toxin reduction remediation	0.094	0.086	0.000	0.037	0.069	0.122	0.761

dynamics behind the indicators produced by exploiting the B-Corp subset. Therefore, we leverage these correlations by performing a Principal Component Analysis (PCA). This analysis serves two purposes: firstly, it reduces the 31 indicators (for each of the 31 labels) to a manageable number; and secondly, it constructs a series of aggregated web-based environmental culture indicators that are orthogonal to one another.

Finally, the concluding step of our analysis, which aims to verify our propositions, involves estimating a series of Ordinary Least Squares (OLS) regressions. The goal is to estimate the proportion of the variance of the B-Lab environmental index that can be explained using our web-based environmental culture indicators. The structure of the regression model to be estimated is as follows:

$$Y_i = \alpha + \rho CommInd + \tau GovernInd + \sum_k^K \eta_k dyear_k + \sum_j^J \beta_j pca_{ij} + \varepsilon_i \quad (1)$$

$$Y_i = \alpha + \rho CommInd + \tau GovernInd + \sum_k^K \eta_k dyear_k + \sum_j^J \beta_j pca_{ij} + \gamma dCanada_i + \sum_l^L \delta_{1l} dsize_{il} + \sum_m^M \theta_m dIndustry_m + \varepsilon_i \quad (2)$$

where Y_i is the dependent variable B-Lab environmental index of firm i that we are trying to predict, with $i \in N,\ 1 \le i \le 1110$, *CommInd* and *GovernInd* are the two control variables representing respectively the log of the B-Lab impact on the community indicator and the B-Lab impact on the governance indicator, the variable *dyear* with $k \in \{2015, 2016, 2017, 2018, 2019, 2020, 2021\}$ represents the dummy variables related to the year of the assessment test filed by the company[9]. The variables pca_{ij} represent the results of the factor analysis, with $j \in N, 1 \le j \le 6$, where 6 is the number of the factors, $\alpha, \rho, \tau, \beta, \delta, \theta \in R$, are the coefficients in the regression model, and $\varepsilon \in R$ indicates the error term of the regression. The dummy variable $dCanada_i$ takes the value 1 if the firm is located in Canada, and 0 otherwise (i.e., it is an American company). The dummy variables $dsize_{il}$ represent the company sizes $l \in N, 1 \le l \le 3$[10]. The dummy variables $dIndustry_m$ represent each industry category $m \in N, 1 \le m \le 9$[11].

4. Web-based variable construction

4.1. Zero-Shot Text Classification

Table 2 presents the results of the initial step of our analysis. For all the 1110 websites, the ZSTC provides the average score, which ranges from 0 to 1, for each of the 31 web-based environmental labels. Considering their mean score, "designed to conserve wholesale process" and "inputs" are the labels with the highest average. In other words, in the full text of the websites, there are, on average, more groups of sentences that, according to the model, refer to these labels. While all

[9] Upon realizing that there was only one observation in *d2015* and noting a correlation of −0.39 between d2019 and d2018 in the correlation matrix, we decided to omit both the two assessment years, d2015 and d2019.

[10] We verified that there were no significant differences between the *size_0* (no employees) and the *size_1_9* (1 to 9 employees) on the dependent variable using the Mann-Whitney test before merging the two B-Lab classifications into *dmicro* (0 to 9 employees). Likewise, we created the dummy variable *dlarge* by merging the *size_250_999* (250 to 999 employees) and *size_1000+* (>1000 employees). Finally, we obtained 4 size categories and *dmicro* is the omitted size dummy in the regression analysis.

[11] We repeated the same Mann-Whitney tests for the industrial classifications provided by B-Lab. The results allowed to merge 4 of the B-Lab industry categories: Media with Restaurant, Hospitality & Travel; Legal Services with Finance Services; and Retail with Transportation & Logistics. Finally, we obtained 10 industry categories and *dConsPdct* is the omitted industry category in the regression analysis.

Table 3
Pearson correlation results.

Labels	Correlation	p-value
Green investing**	0.497	0.000
Resource conservation**	0.472	0.000
Environmentally innovative wholesale process**	0.467	0.000
Green lending**	0.430	0.000
Environmental management**	0.390	0.000
Designed to conserve wholesale process**	0.356	0.000
Designed to conserve agriculture process**	0.349	0.000
Environmental education information**	0.320	0.000
Environmentally innovative manufacturing process***	0.297	0.000
Materials codes**	0.284	0.000
Designed to conserve manufacturing process**	0.283	0.000
Environment products services introduction**	0.266	0.000
Certification**	0.248	0.000
Environmentally innovative agricultural process***	0.215	0.000
Material energy use**	0.214	0.000
Outputs**	0.161	0.000
Land life**	0.108	0.000
Community*	0.081	0.007
Cleaner burning energy***	0.052	0.086
Renewable energy***	0.039	0.199
Inputs*	0.007	0.812
Air climate**	−0.003	0.922
Water***	−0.022	0.460
Safety**	−0.024	0.433
Land office plant**	−0.036	0.233
Toxin reduction remediation**	−0.067	0.025
Construction practices**	−0.079	0.008
Transportation distribution suppliers***	−0.080	0.008
Energy water efficiency**	−0.098	0.001
Training collaboration**	−0.161	0.000
Land wildlife conservation**	−0.220	0.000

Notes: The labels with * are transformed with the formula ln((label) + 1).
The labels with ** are transformed with the formula ln ((label *10) +1).
The labels with *** are transformed with the formula ln((label*100) +1).

variables have a minimum score close to 0, the maximum value varies across the labels. Out of the 31 labels, 17 have a maximum score above 0.90. This indicates that, according to the model, each of these labels was the dominant one on at least one website. Conversely, the labels "clear burning energy", "green lending", "community", "designed to conserve manufacturing process" and "land wildlife conservation" display the lowest maximum values in the table, with the latter showing a maximum score of <0.5. This suggests that generally, our sample does not include websites where the "land wildlife conservation" label has a ZSTC score higher than the other labels. Moreover, "land wildlife conservation", "environmentally innovative manufacturing process" and "green investing" all exhibit a low average score. This implies that the model infrequently identifies groups of sentences that correspond to these labels.

Generally, the minimum score value is closer to the mean than to the maximum value, except for the labels "designed to conserve wholesale process" and "inputs". In other words, while the other labels are found on several websites with low scores or are not found at all, the model considers the labels "designed to conserve wholesale process" and "inputs" only when their value is significantly higher than the other scores. Additionally, the scores displayed in Table 2 do not follow a normal distribution: the mean and median are not equal and for most of the scores, the third quartile is closer to the minimum value, suggesting possible outliers. Consequently, we perform several transformations of the web-based indices before proceeding with the Pearson Correlation test.

4.2. Correlation tests

Table 3 shows the Pearson correlations[12] between each web-based environmental culture indicator and the B-Lab environmental index, the dependent variable of our regression model. Normality of all the variables is ensured by using a natural logarithm transformation (see the notes of Table 3).

To ease the interpretation of the results, we divide Table 3 into three sections. The lower section of the table contains the labels that have either a negative or null correlation with the B-Lab environmental index. Only the last 5 have p-values <0.005, with the last two presenting p-values <0.001. The last two labels have a score that is weakly inversely related to the B-Lab environmental index. As we move up to the middle of the table, in the positive but lesser than 0.3 correlation portion, only the bottom two labels exhibit low and non-significant correlations. The most interesting section is located in the top part of Table 3. There, we find the labels that have the strongest and most significant correlation with the B-Lab environmental index. Among these labels, the top four have a correlation higher than 0.4 with green investing reaching nearly 0.5 (0.497).

4.3. Principal component analysis

To further investigate the relationship between the indices produced and the B-Lab environment index, we first conducted a Principal Component Analysis (PCA). This analysis groups the web-based indicators created with the ZSTC into latent variables. This step has a twofold effect: 1) reduce the correlation among the labels; and 2) it groups them into a smaller set of components maintaining trends and characteristics. The PCA analysis[13] groups the 31 items into 6 components or dimensions. The PCA, presented in Table 4, yields a high Kaiser-Meyer-Olkin (KMO) of 0,865 and a cumulative variance of 68,828 %. The resulting six factors, interpreted using their B-Lab description[14], are as follows: 1) The first component relates to management and finance (*ManFin*); 2) The second focuses on the environmental impact of companies on water and land (*WatLand*); 3) The third covers the energy efficiency of companies (*EnergyEff*); 4) The fourth concerns the impact of companies on the community, including air pollution, economic impact on the area, diversity, civic engagement, and public collaboration (*ComImp*); 5) The fifth is associated with the impact of the companies in agriculture processes and practices (*Agri*); and 6) Lastly, the sixth deals with the processes put in place by the companies to reduce the impact on the environment of manufacturing and transportation process, as well as the safety measure applied by the companies (*ManTransSaf*).

Five of the six components present a high level of reliability with a Cronbach's alpha score >0.70. Although the general rule of thumb suggests that a Cronbach's alpha >0.60 is acceptable, we retain this last component (*ManTransSaf*), despite its Cronbach's alpha on the low side (0,577), because this study is an exploratory analysis (Hair et al., 1998). Before proceeding to the regression analysis, we calculate the score of the PCA for each factor using the SPSS option regression factor to obtain orthogonal components. The six resulting factors will be used as web-based environmental culture indicators in the regression analysis.

[12] The Pearson Correlation matrix and the descriptive statistics were performed by STATA software v. 16.1
[13] The Principal Component Analysis (PCA) was performed using the IBM SPSS software v.29.
[14] See https://data.world/blab/b-corp-impact-data/workspace/data-dictionary.

Table 4
PCA solution and factor loadings.

Labels	FinMan	WatLand	EnergyEff	ComImp	Agri	ManTransSaf
Environment products services introduction	**0.696**	0.244	0.202	0.256	−0.035	0.161
Environmental management	**0.822**	0.251	0.162	−0.048	0.095	−0.062
Resource conservation	**0.882**	0.110	0.052	0.015	0.228	0.105
Environmentally innovative manufacturing process	**0.629**	−0.052	0.194	−0.111	0.157	0.425
Environmentally innovative wholesale process	**0.811**	−0.081	0.295	−0.019	0.123	0.107
Green investing	**0.829**	−0.037	0.317	0.091	0.220	0.000
Green lending	**0.701**	0.057	0.177	0.137	0.270	0.167
Environmental education information	**0.718**	0.250	−0.047	0.147	0.025	−0.084
Water	0.400	**0.628**	−0.167	0.066	−0.037	−0.098
Energy water efficiency	0.181	**0.771**	0.330	0.121	0.042	0.057
Land office plant	0.038	**0.628**	0.303	0.305	0.228	0.152
Land wildlife conservation	0.007	**0.782**	0.113	0.065	0.130	0.075
Toxin reduction remediation	0.142	**0.708**	0.130	0.068	0.067	0.339
Renewable energy	0.313	0.075	**0.719**	0.107	−0.079	−0.156
Material energy use	0.398	0.208	**0.674**	0.049	−0.094	0.296
Construction practices	0.006	0.434	**0.509**	0.211	0.020	0.133
Cleaner burning energy	0.270	0.218	**0.647**	−0.042	0.162	0.097
Community	0.195	0.021	−0.067	**0.786**	0.171	0.062
Air climate	0.077	0.200	0.380	**0.630**	0.018	0.217
Training collaboration	−0.037	0.139	0.059	**0.766**	−0.009	−0.024
Designed to conserve agriculture process	0.491	0.134	−0.002	0.135	**0.688**	0.257
Land life	0.134	0.277	0.075	0.469	**0.641**	−0.063
Environmentally innovative agricultural process	0.276	0.077	−0.027	−0.035	**0.844**	0.020
Transportation distribution suppliers	−0.129	0.356	0.161	−0.103	0.100	**0.580**
Designed to conserve manufacturing process	0.412	−0.021	0.160	0.041	0.199	**0.695**
Safety	0.093	0.194	−0.120	0.307	−0.125	**0.720**
KMO						0.865
Eigen Values	5.805	3.317	2.460	2.258	2.033	2.023
% var	22.327	12.756	9.463	8.684	7.819	7.780
% var. cum	22.327	35.082	44.546	53.230	61.048	68.828
Cronbach's alpha	0.917	0.797	0.698	0.676	0.787	0.547
Cronbach's alpha based on standardized items	0.926	0.826	0.747	0.689	0.790	0.577
Number of items	8.000	5.000	4.000	3.000	3.000	3.000

In the table, the items that contribute to the factor identified by the Principal Component Analysis are highlighted in bold and listed under the corresponding factor name.

5. Regression results

Tables 7, 8, and 9 present the results of the OLS regressions[15,16,17,18]. In Table 7, Reg0 presents the basic regression including only control variables and six factors from the PCA analysis. Reg1 shows the results of the regression that includes variables related to industry, size, and country, while the others present the results of the regression with both the direct and moderating effects of industry (Reg1 to Reg8 in Table 7), firm size (Reg 9 to Reg14 in Table 8), and country (Reg15 to Reg20 in Table 9). The regression exhibits a very high R^2 of approximately 0.58, i.e., we can predict 58 % of the variance of the dependent variable.

5.1. Direct effects

Table 7 shows 9 different regressions: the basic regression (Reg1) that includes the control variables and the six orthogonal components, the complete regression (Reg2) used to test the propositions for the direct effect, and Reg 3–8, which are the regressions with the moderating effect for the industry category. Reg0 demonstrates the significant and negative impact of the two ESG variables, *CommInd* (impact area community) and *GovernInd* (impact area governance). This result aligns with the idea that companies do not exert the same efforts in all the ESG areas (Liute and De Giacomo, 2022). The factors associated with Management and Finance (*FinMan*), Agriculture (*Agri*) and Energy efficiency (*EnergyEff*) have a positive and significant impact on the B-Lab environmental index. More specifically, companies that mention specific topics related to green investing or their good management of the company resources or renewable energy on their website are also those that exhibit a higher environmental index as measured by B-Lab. Likewise, highlighting good stewardship of the land through environmentally friendly processes, such as conserving natural resources or developing innovative agricultural processes, yields a higher score on this environmental index. However, the association is negative and significant for Water and Land (*WatLand*), which is built from the labels "water", "energy water efficiency", "land office plant", "land wildlife conservation" and "toxin reduction remediation", as well as Community impact (*ComImp*) built from the labels "air climate", "training collaboration" and "community". This suggests that some key topics for the environment, such as "energy water efficiency", "land wildlife conservation", and "toxin reduction remediation" do not positively impact the B-lab environment index as expected. While this result requires further analysis, one possible interpretation is that these topics refer to long-term projects. Specifically, companies may only be discussing future

[15] The Multivariate Ordinary Least Square regression was performed by STATA software v. 16.1.

[16] Three of the PCA web-based environmental indicators had to be normalised prior to the regression analysis: *EnergyEff*, *Agri*, and *ManTransSaf*. Details of the transformations on Table 10.

[17] We conducted normality tests, confirming that residuals fall within acceptable bounds for skewness and kurtosis. We also examined autocorrelation between residuals using the Durbin-Watson statistic, observing no autocorrelation within the limits. However, the Breusch-Pagan test rejected the hypothesis of constant variance (homoskedasticity). Due to the heteroskedastic residuals, we employed the "vce robust" option in Stata to mitigate this effect (https://www.stata.com/manuals/semintro8.pdf, Nov. 24th 2023).

[18] We performed a Tobit regression to ensure robustness, yielding results highly similar to the linear regression, which are presented in Annex 3. This might be caused by the fact that we encountered only 7 observations in the left-censored category (0) and one observation in the right-censored category (66.10).

projects that do not reflect the current state of the B-Lab environmental index, which evaluates projects already implemented by the company.

The factors related to Management, Transportation, and Safety (*ManTransSaf*) show no impact on the environmental culture. In contrast, and quite surprisingly, companies that emphasize their sense of community and collaboration have a negative impact on the B-Lab environmental index. To explore the dynamics identified through the Pearson correlation in Proposition 1, we examined Reg1, which incorporates the variables from the basic regression and those related to the country, size, and industry category of the companies. In Reg1, it is observed that among the ESG control variables, only *CommInd* (impact area community) is significant, exhibiting a negative association. This suggests that a company with a higher impact community score typically has a lower B-Lab environmental index. Concerning the six PCA factors, the *EnergyEff* factor becomes insignificant when introducing variables related to industry, size, and country. As suspected in the theoretical framework that led to our propositions, industrial differences are probably at play here. Before exploring the moderating effects, let us first examine the direct effects. Depending on the industry category, companies may have more interest in applying good environmental practices when a specific industry category is known for its pollution, or depending on the category, companies could have a direct economic advantage in applying good environmental practices (e.g., de Azevedo Rezende et al., 2019; Hermundsdottir and Aspelund, 2021). In general, compared to the Consumer product and service industry category (the omitted category or baseline), all industry categories but Retail, Transportation and Logistics (*dRetTransLog*) yield significant coefficients. The only industries that have a negative association with the dependent variable compared to the baseline are Media, Restaurant, and Hospitality (*dMediaRestHosp*), and Business Product and Service (*dBusinessPro*) industry categories. As mentioned in Section 3, the impact of these industry categories on the environmental index of B-lab was expected.

Surprisingly, the country does not seem to matter in our model (Table 9), i.e., once we have controlled for industry categories and size: the coefficient of *dCanada* compared to the US as the baseline is not significant. This result is not aligned with the literature (e.g., Doran and Ryan, 2012; Aguilera-Caracuel and Ortiz-de-Mandojana, 2013), as researchers studying SI and evidence of green innovation performances have generally shown how several factors contribute to differentiated impacts of the country on its environmental performances. For example, policymakers can incentivize and reinforce positive environmental behavior using subsidies and increasing the regulatory constraints on pollution.

Compared to the micro companies (0 to 9 employees), our baseline, Table 8 indicates that medium and large firms have a positive and significant association with the B-Lab environmental index while small firms do not show a significant difference compared to micro firms. Increasing the size of a company from a range of 0–9 to a number of employees higher than 49 results in a higher B-Lab environmental index. The results align perfectly with the literature (e.g., Aguilar-Fernández and Otegi-Olaso, 2018) that suggests firm size impacts a companies' pursuit of SI. While the impact of medium and large enterprises is not statistically different from one another, their differentiated influence contrasts with both small and very small firms.

5.2. Moderating effects

This section focuses on the coefficients of the interaction terms presented in the lower part of each regression results table. Presents the moderating effects of the different industry categories. In Reg2, the negative coefficient associated with Business Product and Service industry category is partially mitigated by a strong signal regarding environmental management and manufacturing processes, green investing, and lending (*FinMan*). In Reg3, the industry categories where water conservation and good stewardship of the land matter most, i.e., Agriculture (*dAgricul*) and Building (*dBuild*), are the main contributors to the positive association of this web-based environmental culture indicator with the B-Lab environmental index. However, *WatLand* does not have a significant impact when moderating the industry. Reg5 shows that most of the industry categories exhibit a negative moderating effect on the importance of community and collaboration. With the exception of Agriculture, Building, and Energy Environment industry categories where we observe a positive effect of this web-based environmental culture indicator on the B-Lab environmental index. Additionally, Reg5 displays no significant moderating effect of *ComImp* for the industry categories. In Reg6, all the industry categories are not significant and only the moderating effect of Finance and Legal service category, when moderating the relation between the agriculture factor (*Agri*) and the B-Lab environmental index, becomes slightly significant and negative. Thus, the items associated with the factor Agriculture, when addressed by companies in Finance and Legal Services have a negative impact on their environmental score.

Given the non-significant impact of the *EnergyEff* and *ManTransSaf* factors, further analysis was warranted. Specifically, we compared the moderating effect of industry categories on the impact of these factors on the B-Lab environmental index. The web-based environmental culture indicator linked to Energy Efficiency (*EnergyEff*), becomes negative and significant when moderated by the Agriculture industry category (*dAgricul*) and positive and significant when moderated by the Health and Human industry (*dHealthHuman*) (see Reg5). From an industrial perspective, Agriculture firms show a reduced correlation with the B-Lab environmental index when influenced by a robust web-based environmental culture indicator related to *EnergyEff*, although the overall impact remains positive. Conversely, in the case of *dHealthHuman*, the *EnergyEff* moderated by *dHealthHuman* is positive, but the general effect remains negative. Table 5 shows that Agriculture has a negative and significant moderating effect on the relationship between *EnergyEff* and environmental impact, compared to other industry categories. The results suggest that Agriculture companies that include concepts related to "renewable energy" (driven items of the *EnergyEff* factor according to the loading of the PCA analysis) in their websites tend to have a lower environmental index. Most other sectors show a positive effect, except for Media, Restaurant and Hospitality (*dMediaRestHosp*), Business Products and Services (*dBusinessProd*), and Education and Training (*dEducationTr*).

Although the Agriculture industry category is broad, encompassing companies producing tractors to shops selling fruits, we observe that companies tend to focus more on the products they are selling or producing, and less on the sustainability of their company (e.g., energy usage to run their activities). Thus, for this industry, the average score of the ZSTC for the items composing *EnergyEff* is less than the average score of the items for companies in other industry categories Additionally, the distinctly negative coefficient for *EnergyEff* moderated by Agriculture, as seen in the table, consistently reflects a positive delta compared to *dConsumPrd*, *dEducationTr*, *dMedRestHosp,* and *dHealthHuman*. For the latter (see Reg4), we observe a positive moderating effect for an industry that includes companies related to health and care. This sector encompasses companies such as veterinary, mental health, and homecare, and the average score of the ZSTC for the items composing *EnergyEff* is generally low due to the lesser importance given to this topic compared to other topics such as community and safety.

The regression results in Reg 7 reveal a substantial negative impact of the cluster comprising the Media, Restaurant, and Hospitality industry categories on the B-Lab environmental index. Similarly, Finance and Legal Services demonstrate a significant negative effect on the B-Lab environmental index. In terms of moderating effects, Reg 7 underscores Agriculture and, Health and Human industry categories as notable for their significant moderating influence on this factor. Companies in the Agriculture industry category, emphasizing concepts related to the items comprising *ManTransSaf*, exhibit a lower B-Lab environmental index. Conversely, a positive effect is observed when the moderating industry is Health and Human. Table 6 illustrates that, when moderated by the

Table 5
Moderating effect of the industry category for the EnergEff factor.

Industry		Coeff.	1	2	3	4	5	6	7	8	9	10
			−46.880	20.419	2.038	−7.394	10.960	2.366	36.997	−8.814	8.429	−2.454
dAgricul	1	−46.880										
dBuild	2	20.419	++									
dConsumPrd	3	2.038										
dEducationTr	4	−7.394										
dEnergyEnvir	5	10.960	++									
dFinLegservic	6	2.366	++									
dHealthHuman	7	36.997	+++		+	+						
dMedRestHosp	8	−8.814							-			
dRetTransLog	9	8.429	++									
dBusinessPro	10	−2.454	+									

+ if the difference between two cells is positive and the p-value ≤0.1.
++ if the difference between two cells is positive and the p-value ≤0.05.
+++ if the difference between two cells is positive and the p-value ≤0.01.
- if the difference between two cells is negative and the p-value l ≤ 0.1.
– if the difference between two cells is negative and the p-value ≤0.05.
— if the difference between two cells is negative and the p-value ≤0.01.

Table 6
Moderating effect of the industry categories for the ManTransSaf.

Industry		Coeff.	1	2	3	4	5	6	7	8	9	10
			−4.376	−5.454	1.466	−9.736	1.581	6.315	7.032	16.260	−9.372	−2.233
dAgricul	1	−4.376										
dBuild	2	−5.454										
dConsumPrd	3	1.466	+++									
dEducationTr	4	−9.736										
dEnergyEnvir	5	1.581										
dFinLegservic	6	6.315				++						
dHealthHuman	7	7.032				+						
dMedRestHosp	8	16.260	+	++	+++	+++						
dRetTransLog	9	−9.372			—							
dBusinessPro	10	−2.233										

+ if the difference between two cells is positive and the p-value ≤0.1.
++ if the difference between two cells is positive and the p-value ≤0.05.
+++ if the difference between two cells is positive and the p-value ≤0.01.
- if the difference between two cells is negative and the p-value l ≤ 0.1.
– if the difference between two cells is negative and the p-value ≤0.05.
— if the difference between two cells is negative and the p-value ≤0.01.

Agriculture industry category, the difference in effects is positive compared to Media, Restaurant and Hospitality, Consumer Products, and Services. Notably, Media, Restaurant and Hospitality also exhibit a positive and significant difference compared to Building Consumer Products and Services and Education and Training.

The PCA reveals that "safety" is the driving item for the *ManTransSaf* factor. Although the Media, Restaurant, and Hospitality industry category includes companies ranging from lobster sellers to book publishers, we observed that companies performing well in this category tend to emphasize the concept of "safety". As B-Lab does not provide a precise definition of "safety", we used the knowledge of the NLP model BART to understand the meaning of the word. Our exploration revealed that the context of 'safety' is remarkably broad, encompassing social dimensions like protection from bullies in the Education and Training industry category to physical security concerns, such as ensuring a safe safari experience in the Media, Restaurant, and Hospitality industry category. This broad interpretation of "safety" is reflected in the significant result observed for *ManTransSaf* moderated by *dHealthHuman*, aligning with the nature of companies in this category that prioritize safety measures for the well-being of individuals. However, it's essential to acknowledge that the lack of a precise definition for the label "safety" may contribute to explaining the contrasting results obtained in Reg6.

Regressions 9 to 14 (in Table 8) show the moderating effect of the size categories. Small firms that signal a strong web-based environmental culture index regarding green finance and environmental management (*FinMan*) are associated with a higher B-Lab environmental index, hence partially compensating for their small size status compared to their larger counterparts. In contrast, web-based indicators related to water energy conservation and better land management (*WatLand*) have a negative moderating impact for medium-sized companies, implying that the negative association between the web-based measures and the B-Lab index is due to the signals sent by medium-sized companies on their websites. The web-based indicator related to the projects put in place by the companies to reduce the impact on the environment of the manufacturing and transportation process, as well as the safety measures applied (*ManTransSaf*) has a weakly significant and positive impact on the large firms. Overall, the moderating effects due to the size of firms are very small. We suspect that with a larger sample of firms, most of these effects may disappear.

Finally, the last six regressions (Table 9) explore the moderating effects of the country (Canada compared with the US) on the relationship between the web-based environmental culture indices and the B-Lab environmental index. Table 9 shows that there is no moderating effect of the country on the web-based environmental culture indices. The introduction of these moderating effects in the regression models does not significantly influence the sign or level of significance of the coefficients of the variables *EnergyEff*, and *ManTransSaf,* which remain non-significant. In other words, Canadian firms do not have a different effect on the web-based environmental culture indices compared to the American.

Table 7
Basic regression results and moderating effect of industry.

Variables	Reg1		Reg2		Reg3		Reg4		Reg5		Reg6		Reg7		Reg8	
CommInd	−1.658	**	−1.613	**	−1.590	**	−1.623	**	−1.559	**	−1.661	**	−1.551	**	−1.662	**
	(0.805)		(0.737)		(0.701)		(0.730)		(0.737)		(0.739)		(0.745)		(0.744)	
GovernInd	−0.306	***	−0.050		−0.053		−0.045		−0.038		−0.047		−0.046		−0.051	
	(0.077)		(0.065)		(0.063)		(0.066)		(0.064)		(0.064)		(0.063)		(0.066)	
dsmall			0.885		0.762		0.823		0.717		0.818		0.923		0.989	*
			(0.583)		(0.565)		(0.580)		(0.585)		(0.573)		(0.587)		(0.585)	
dmedium			3.276	***	3.139	***	3.334	***	3.203	***	3.258	***	3.312	***	3.306	***
			(0.760)		(0.747)		(0.752)		(0.756)		(0.756)		(0.767)		(0.760)	
dlarge			3.088	***	2.645	**	3.062	***	3.092	***	3.110	***	2.987	**	3.089	***
			(1.140)		(1.125)		(1.140)		(1.120)		(1.161)		(1.167)		(1.136)	
dCanada			0.794		0.758		0.742		0.809		0.757		0.745		0.828	
			(0.663)		(0.638)		(0.672)		(0.669)		(0.673)		(0.666)		(0.664)	
dAgricul			7.407	***	6.846	***	10.416	***	117.483	**	8.104	***	−6.024		24.427	
			(2.589)		(2.571)		(2.749)		(52.064)		(2.563)		(47.853)		(31.968)	
dBuild			7.430	***	6.881	***	6.685	***	−43.125		9.295	***	−81.099		28.268	
			(2.097)		(2.077)		(1.930)		(58.975)		(2.328)		(61.393)		(27.241)	
dEducationTr			−12.385	***	−15.068	***	−12.412	***	5.265		−13.160	***	−42.702		24.950	
			(1.171)		(1.790)		(1.164)		(38.459)		(1.577)		(53.580)		(25.625)	
dEnergyEnvir			9.457	***	12.068	***	8.275	***	−17.907		9.073	***	−16.541		3.276	
			(1.818)		(2.755)		(2.016)		(32.702)		(1.939)		(41.319)		(22.737)	
dFinLegservic			−13.002	***	−14.223	***	−12.725	***	−18.692		−12.897	***	26.542		−37.783	**
			(0.772)		(0.873)		(0.785)		(31.415)		(0.766)		(20.662)		(16.363)	
dHealthHuman			−10.176	***	−11.812	***	−10.053	***	−97.652	*	−9.709	***	51.703		−37.587	
			(1.177)		(1.417)		(1.166)		(52.633)		(1.208)		(45.102)		(25.465)	
dMedRestHosp			−7.773	***	−8.330	***	−7.481	***	12.919		−7.776	***	10.741		−69.786	**
			(1.895)		(1.864)		(1.872)		(29.606)		(1.843)		(31.144)		(29.882)	
dRetTransLog			−1.736		−2.452		−2.039		−21.697		−1.563		−42.772		35.406	
			(1.565)		(1.505)		(1.564)		(40.807)		(1.552)		(36.767)		(26.178)	
dBusinessPro			−9.056	***	−8.717	***	−9.129	***	−3.138		−8.857	***	25.453		−0.402	
			(0.802)		(0.848)		(0.801)		(33.765)		(0.801)		(27.629)		(16.533)	
FinMan	6.667	***	4.144	***	2.780	***	4.181	***	4.105	***	4.111	***	4.165	***	4.170	***
	(0.367)		(0.392)		(0.502)		(0.394)		(0.387)		(0.387)		(0.397)		(0.394)	
WatLand	−2.430	***	−0.664	**	−0.698	**	−0.679		−0.649	*	−0.680	**	−0.560		−0.631	*
	(0.362)		(0.324)		(0.338)		(0.548)		(0.348)		(0.323)		(0.345)		(0.329)	
EnergEff	10.514	**	2.515		2.385		0.080		2.038		1.319		2.982		3.522	
	(4.575)		(4.484)		(4.100)		(4.623)		(7.005)		(4.324)		(4.541)		(4.616)	
ComImp	−0.750	**	−0.588	**	−0.576	**	−0.650	**	−0.601	**	−1.293	**	−0.603	**	−0.611	**
	(0.315)		(0.283)		(0.274)		(0.280)		(0.278)		(0.583)		(0.283)		(0.284)	
Agri	25.714	***	15.548	***	14.226	***	15.574	***	16.104	***	15.830	***	16.932	***	14.642	***
	(4.302)		(3.862)		(3.808)		(4.021)		(3.973)		(3.864)		(5.573)		(3.918)	
ManTransSaf	2.346		1.249		1.745		1.593		1.691		1.080		1.340		1.466	
	(1.787)		(1.603)		(1.675)		(1.633)		(1.659)		(1.607)		(1.647)		(2.493)	
DummyYears	Yes		Yes		Yes		Yes		Yes		Yes		Yes		Yes	
Constant	−69.551	***	−21.839		−19.636		−17.646		−24.154		−18.959		−26.931		−22.809	
	(14.958)		(14.758)		(14.402)		(15.148)		(19.502)		(14.449)		(17.602)		(16.835)	

Interaction x ...			…FinMan		…WatLand	…EnergyEff		…ComImp	…Agri		…ManTransSaf	
dAgricul			2.487		4.852	−46.880	**	3.922	5.105		−4.376	**
			(2.890)		(3.296)	(21.969)		(2.920)	(18.557)		(8.272)	
dBuild			2.502		3.574	20.419		−3.242	36.904		−5.454	
			(2.269)		(1.989)	(23.972)		(2.386)	(25.555)		(7.139)	
dConsumProdu												
dEducationTr			−2.733		−1.905	−7.394		1.638	12.907		−9.736	
			(2.400)		(1.069)	(16.340)		(1.253)	(22.676)		(6.655)	
dEnergyEnvir			−1.032		−2.083	10.960		2.384	11.238		1.581	
			(1.676)		(1.327)	(13.074)		(1.473)	(17.566)		(5.974)	
dFinLegservic			−0.716		−1.037	2.366		0.959	−16.519	*	6.315	
			(0.981)		(0.980)	(13.176)		(0.758)	(8.624)		(4.118)	
dHealthHuman			−0.133		0.115	36.997	*	−0.103	−26.036		7.032	*
			(1.849)		(1.200)	(22.278)		(0.939)	(18.899)		(6.453)	
dMedRestHosp			2.220		−1.892	−8.814		1.452	−7.863		16.260	
			(1.812)		(1.912)	(12.384)		(2.122)	(13.156)		(7.894)	
dRetTransLog			1.781		−1.373	8.429		1.143	17.159		−9.372	
			(1.323)		(1.733)	(17.105)		(1.386)	(15.348)		(6.634)	
dBusinessPro			4.313	***	0.574	−2.454		1.091	−14.496		−2.233	
			(0.988)		(0.806)	(14.123)		(0.736)	(11.519)		(4.195)	
Nb obs.	1110	1110	1110		1110	1110		1110	1110		1110	
F	37.620	61.566	51.944		49.877	52.904		48.604	46.821		47.780	
R^2	0.370	0.587	0.605		0.595	0.592		0.593	0.591		0.591	
Adjusted R^2	0.363	0.577	0.593		0.581	0.579		0.580	0.578		0.578	
Kurtosis	4.148											

(*continued on next page*)

Table 7 (*continued*)

Interaction x ...			…FinMan	…WatLand	…EnergyEff	…ComImp	…Agri	…ManTransSaf
Durbin-Watson	2.004							
d_l	1.855							
d_u	1.937							
4-du	2.063							
4-dl	2.145							
Breusch-Pagan	72.690	***						

Notes:***$p \leq 0.001$, **$p \leq 0.05$, *$p \leq 0.1$. The Breusch-Pagan test is a χ^2 with 1 degree of freedom. "dl" and "du" are the lower and upper critical values of the Durbin-Watson test. Since '2.004' falls between "du" and "4-du" there is no autocorrelation.
DummyYears refers to the control variables for the assessment years. Compared to the omitted variables *d2015&d2019* only *d2016* is significant for all the regressions but Reg1.
dMicro, very small firms, is the omitted firm size category, *dConsPdct*, Consumer products, is the omitted industry category.
EnergyEff represents the transformation ln (EnergyEff+11).
Agri represents the transformation ln (Agri+11).
ManTransSaf represents the transformation ln((ManTransSaf+5) *10 + 1).
CommInd represents the transformation ln (CommInd+1).

6. Discussion and conclusion

The main objective of this research was to determine whether web-based sustainable innovation indicators can serve as a proxy for performance measures and indices built using traditional survey-based and administrative data. This pilot study compares web-based environmental culture indicators with the environmental index created by B-Lab.

The first proposition was based on three main assumptions: first, nowadays companies feel greater pressure to implement environmental initiatives; second, environmentally friendly companies pursuing green goals have competitive advantages (Paelman et al., 2020); and third an official website serves as a platform for delivering authentic, precise, and up-to-date information about companies (Jiang et al., 2023). This proposition was accepted based on the moderate correlation of certain topics with the environmental index (Table 3) and the results of the regression analysis. The latter shows that the web-based indicators created from ZSTC, together with control variables, industry, size, and country attributes, contribute to explaining over 57 % of the variance of the B-Lab environmental index.

The second proposition concerns the direct impact of the country in which the company operates on the evidence of its appropriate environmental approach. This proposition is rejected because our study shows no significant differences between the firms located in Canada and the United States on the B-Lab environmental index. Arguably, the well-aligned green policies and long-standing agreements in protecting the environment and decarbonizing the industry between Canada and the United States contribute to explaining this result[19]. The results by no means imply that the country does not influence the way companies address environmental issues, take actions, and write about them on their external communication channels such as their corporate websites. Also, proposition 2 m is rejected. Although Magnusson et al. (2011) suggested that the country shapes the perception of corporate sustainability of the customers, we found no differences in the signal captured on the website of Canadian and United States companies.

The third proposition examined the impact of a company's size on its approach to sustainable environmental practices and is accepted. Researchers (e.g., Ketata et al., 2015) suggest that the size of a company may impact its approach to environmental practices, but there is no consensus in the literature on whether the impact is positive or negative. Our study clearly shows that larger companies have a higher environmental index, indicating a positive impact of size. There are likely both internal and external reasons for this finding. As highlighted by previous research (e.g., Aguilar-Fernández and Otegi-Olaso, 2018), larger companies are more exposed to criticism from customers and stakeholders, which can pressure them to pursue eco-conscious practices. Larger companies have more control over the market, as they can more easily choose suppliers that adopt green practices. Furthermore, they generally have more resources to invest compared to smaller companies (Aguilar-Fernández and Otegi-Olaso, 2018). However, we found that the moderating effect of size is weak or infrequent and the evidence generated only partially supports proposition 3 m. Some industry categories are likely to be dominated by firms of specific sizes. We suspect that a triple interaction between web-based environmental indicators, industry category, and size is needed to disentangle this.

Lastly, the fourth proposition regarding the direct effect of industry categories on the B-Lab environmental index is accepted. With the exception of Retail, Transportation, and Logistics, all industry categories showed a significant impact compared to the baseline. There are specific industrial characteristics that influence firms' approach to sustainable innovation, with some industries having more interest in pursuing environmental goals due to economic or social pressure (Hermundsdottir and Aspelund, 2021). Proposition 4 m, which aimed to explore the moderating effect of industry category on the relationship between our web-based environmental indicators and the B-Lab index, is also accepted. All six web-based indicators have a significant impact on the B-Lab environmental index when the industry category is changed. Both Agriculture and Health and Human industry categories moderate the relationship between the B-Lab index and the web-based indicators regarding water and land stewardship, energy efficiency, and community implication. The web-based indicator for Manufacturing, Transportation, and Safety moderates that relationship for the industry categories of Agriculture and Health and Human.

This study makes significant advancements both in terms of theoretical frameworks and methodological approaches, enriching the discourse in Sustainability Research and Signal Theory. On the theoretical front, our paper makes a dual contribution to sustainability studies. Primarily, by establishing a correlation between our web-based indices and the B-Lab environmental index, our research underscores the potential of website communication to reflect a company's environmental "performance". Furthermore, through the regression analysis, we disentangle the effect of the web indices on the B-Lab environment index shedding light on the subjects correlated to a higher B-Lab environment index. Concerning Signal Theory, our work responds to the critical inquiry posed by Connelly et al. (2011): "Does the signal represent a valid and reliable measure of the underlying quality that the signaler is attempting to communicate?" In answering this, our investigation confirms the feasibility of extracting meaningful signals from company websites, validating these within the sustainability domain via the B-Lab environmental index, thus bridging a vital research gap. From

[19] See for instance the joint statement: https://www.canada.ca/en/environment-climate-change/news/2021/04/joint-statement-by-the-us-environmental-protection-agency-and-environment-and-climate-change-canada-on-environment-and-climate-change.html.

Table 8
Regression results exploring the moderating effect of size.

Variable	Reg9		Reg10		Reg11		Reg12		Reg13		Reg14	
CommInd	−1.574	**	−1.661	**	−1.614	**	−1.612	**	−1.684	**	−1.607	**
	(0.732)		(0.739)		(0.735)		(0.737)		(0.739)		(0.739)	
GovernInd	−0.046		−0.045		−0.058		−0.050		−0.051		−0.045	
	(0.065)		(0.066)		(0.065)		(0.065)		(0.065)		(0.066)	
dAgricul	7.280	***	7.357	***	7.447	***	7.402	***	7.382	***	7.463	***
	(2.579)		(2.594)		(2.582)		(2.581)		(2.550)		(2.591)	
dBuild	7.545	***	7.355	***	7.401	***	7.465	***	7.486	***	7.406	***
	(2.082)		(2.091)		(2.090)		(2.108)		(2.116)		(2.106)	
dEducationTr	−12.374	***	−12.373	***	−12.242	***	−12.451	***	−12.307	***	−12.326	***
	(1.155)		(1.185)		(1.183)		(1.177)		(1.172)		(1.179)	
dEnergyEnvir	9.128	***	9.483	***	9.245	***	9.506	***	9.771	***	9.361	***
	(1.833)		(1.800)		(1.845)		(1.815)		(1.828)		(1.822)	
dFinLegservic	−12.967	***	−13.065	***	−12.967	***	−12.997	***	−12.924	***	−13.053	***
	(0.773)		(0.777)		(0.778)		(0.778)		(0.775)		(0.770)	
dHealthHuman	−10.197	***	−10.300	***	−10.204	***	−10.273	***	−10.183	***	−9.917	***
	(1.190)		(1.191)		(1.186)		(1.200)		(1.181)		(1.153)	
dMedRestHosp	−7.856	***	−7.861	***	−7.780	***	−7.814	***	−7.782	***	−7.802	***
	(1.910)		(1.865)		(1.911)		(1.899)		(1.901)		(1.881)	
dRetTransLog	−1.654		−1.631		−1.608		−1.694		−1.546		−1.687	
	(1.584)		(1.546)		(1.573)		(1.569)		(1.567)		(1.598)	
dBusinessPro	−9.077	***	−9.091	***	−9.058	***	−9.051	***	−9.016	***	−9.031	***
	(0.797)		(0.802)		(0.801)		(0.803)		(0.802)		(0.799)	
dCanada	0.785		0.750		0.809		0.796		0.813		0.751	
	(0.662)		(0.659)		(0.664)		(0.668)		(0.661)		(0.666)	
dsmall	0.845		0.908		−3.457		0.881		−17.652		8.952	
	(0.573)		(0.584)		(21.111)		(0.583)		(19.475)		(12.906)	
dmedium	3.286	***	2.946	***	−25.990		3.239	***	−29.394		−10.630	
	(0.773)		(0.776)		(24.519)		(0.763)		(23.957)		(16.511)	
dlarge	2.971		3.096	***	42.135		3.104	***	−2.059		−40.373	
	(1.893)		(1.136)		(55.087)		(1.154)		(34.307)		(24.864)	
FinMan	3.524	***	4.154	***	4.169	***	4.143	***	4.146	***	4.160	***
	(0.483)		(0.390)		(0.392)		(0.393)		(0.391)		(0.394)	
WatLand	−0.664	**	−0.387		−0.610	*	−0.657	**	−0.646	**	−0.739	**
	(0.328)		(0.474)		(0.329)		(0.325)		(0.323)		(0.329)	
EnergEff	2.712		1.946		0.554		2.342		2.163		2.579	
	(4.557)		(4.468)		(6.361)		(4.480)		(4.460)		(4.511)	
ComImp	−0.611	**	−0.617	**	−0.621	**	−0.564		−0.587	**	−0.592	**
	(0.281)		(0.284)		(0.282)		(0.426)		(0.282)		(0.283)	
Agri	15.606	***	15.493	***	15.357	***	15.502	***	10.774	**	15.515	***
	(3.845)		(3.828)		(3.877)		(3.857)		(5.407)		(3.906)	
ManTransSaf	1.404		1.453		1.228		1.230		1.211		0.895	
	(1.622)		(1.613)		(1.619)		(1.605)		(1.601)		(2.389)	
DummyYears	Yes		Yes		Yes		Yes		Yes		Yes	
Constant	−23.228		−21.013		−16.535		−21.231		−9.175		−20.613	
	(15.052)		(14.596)		(19.411)		(14.676)		(18.502)		(17.559)	

Interaction x …	…FinMan		…WatLand		…EnergyEff	…ComImp	…Agri	…ManTransSaf	
dsmall	1.470	**	−0.128		1.823	−0.251	7.727	−2.058	
	(0.697)		(0.689)		(8.817)	(0.618)	(8.128)	(3.282)	
dmedium	0.940		−1.966	*	12.287	0.413	13.681	3.550	
	(1.015)		(1.048)		(10.264)	(0.789)	(10.047)	(4.210)	
dlarge	0.459		−0.252		−16.330	−0.093	2.156	10.896	*
	(2.708)		(0.897)		(23.088)	(0.995)	(14.367)	(6.299)	
Nb obs.	1110		1110		1110	1110	1110	1110	
F	55.775		56.444		55.938	55.299	56.206	55.737	
R^2	0.589		0.589		0.588	0.587	0.588	0.589	
Adjusted R^2	0.578		0.578		0.577	0.576	0.577	0.578	

Notes: ***p ≤ 0.001, **p ≤ 0.05, *p ≤ 0.1. *dMicro*, very small firms, is the omitted firm size category, *dConsPdct*, Consumer products, is the omitted industry category. *DummyYears* refers to the control variables for the assessment years. Compared to the omitted variables *d2015&d2019* only *d2016* is significant for all the regressions.
EnergyEff represents the transformation ln (EnergyEff+11).
Agri represents the transformation ln (Agri+11).
ManTransSaf represents the transformation ln((ManTransSaf+5) *10 + 1).
CommInd represents the transformation ln (CommInd+1).

a methodological perspective, our study brings to light the potential to accurately replicate the B-Lab index, which can have a game-changing impact in several aspects. Leveraging Zero-Shot Text Classification (ZSTC) alongside a pre-trained language model, our approach yields impressive results, explaining upwards of 57 % of the variance in the B-Lab index. This efficiency is achieved without extensive textual preprocessing or reliance on an annotated, which streamlines research efforts in terms of time and cost.

Moreover, with further validation and considering the limitations presented in the following section, our methodology holds promise for forecasting the B-Lab scores assigned to corporations. Furthermore, by relying on a pre-trained language model and employing a semantic

Table 9
Regression results exploring the moderating effect of country.

Variables	Reg15		Reg16		Reg17		Reg18		Reg19		Reg20	
CommInd	−1.593	**	−1.639	**	−1.613	**	−1.613	**	−1.616	**	−1.610	**
	(0.737)		(0.739)		(0.736)		(0.738)		(0.737)		(0.737)	
GovernInd	−0.050		−0.053		−0.045		−0.050		−0.051		−0.049	
	(0.065)		(0.065)		(0.065)		(0.065)		(0.065)		(0.065)	
dsmall	0.888		0.864		0.835		0.885		0.891		0.880	
	(0.583)		(0.584)		(0.585)		(0.585)		(0.584)		(0.583)	
dmedium	3.289	***	3.273	***	3.263	***	3.276	***	3.277	***	3.311	***
	(0.758)		(0.760)		(0.759)		(0.762)		(0.760)		(0.761)	
dlarge	3.100	***	3.046	***	3.106	***	3.087	***	3.096	***	3.173	***
	(1.140)		(1.143)		(1.137)		(1.140)		(1.141)		(1.132)	
dCanada	0.766		0.707		27.868		0.794		−5.851		24.223	
	(0.638)		(0.644)		(23.496)		(0.663)		(25.589)		(17.495)	
dAgricul	7.454	***	7.346	***	7.477	***	7.407	***	7.474	***	7.486	***
	(2.595)		(2.592)		(2.593)		(2.591)		(2.591)		(2.591)	
dBuild	7.490	***	7.460	***	7.377	***	7.431	***	7.435	***	7.524	***
	(2.096)		(2.101)		(2.093)		(2.098)		(2.098)		(2.096)	
dEducationTr	−12.345	***	−12.373	***	−12.384	***	−12.386	***	−12.389	***	−12.395	***
	(1.168)		(1.165)		(1.179)		(1.172)		(1.169)		(1.169)	
dEnergyEnvir	9.466	***	9.436	***	9.460	***	9.457	***	9.461	***	9.559	***
	(1.821)		(1.828)		(1.806)		(1.819)		(1.820)		(1.818)	
dFinLegservic	−12.973	***	−12.956	***	−13.064	***	−13.002	***	−12.992	***	−12.964	***
	(0.770)		(0.773)		(0.772)		(0.774)		(0.775)		(0.771)	
dHealthHuman	−10.140	***	−10.150	***	−10.182	***	−10.177	***	−10.174	***	−10.214	***
	(1.174)		(1.178)		(1.181)		(1.179)		(1.176)		(1.183)	
dMedRestHosp	−7.714	***	−7.725	***	−7.721	***	−7.773	***	−7.771	***	−7.729	***
	(1.904)		(1.892)		(1.886)		(1.897)		(1.897)		(1.880)	
dRetTransLog	−1.757		−1.699		−1.800		−1.736		−1.729		−1.644	
	(1.565)		(1.563)		(1.560)		(1.566)		(1.570)		(1.555)	
dBusinessPro	−9.029	***	−9.048	***	−9.054	***	−9.056	***	−9.050	***	−9.003	***
	(0.801)		(0.805)		(0.803)		(0.805)		(0.803)		(0.805)	
FinMan	4.081	***	4.140	***	4.134	***	4.144	***	4.145	***	4.149	***
	(0.421)		(0.392)		(0.392)		(0.392)		(0.392)		(0.391)	
WatLand	−0.672	**	−0.791	**	−0.640	**	−0.664	**	−0.661	**	−0.689	**
	(0.323)		(0.341)		(0.323)		(0.328)		(0.324)		(0.322)	
EnergEff	2.553		2.309		4.471		2.515		2.486		2.535	
	(4.507)		(4.493)		(4.873)		(4.486)		(4.485)		(4.472)	
ComImp	−0.589	**	−0.572	**	−0.579	**	−0.586	*	−0.585	**	−0.568	**
	(0.283)		(0.286)		(0.282)		(0.311)		(0.283)		(0.284)	
Agri	15.544	***	15.684	***	15.632	***	15.548	***	15.141	***	15.489	***
	(3.853)		(3.849)		(3.825)		(3.858)		(4.073)		(3.845)	
ManTransSaf	1.251		1.323		1.305		1.249		1.262		2.124	
	(1.603)		(1.600)		(1.606)		(1.607)		(1.607)		(1.684)	
DummyYears	Yes		Yes		Yes		Yes		Yes		Yes	
Constant	−22.026		−21.846		−26.983		−21.840		−20.835		−25.240	
	(14.817)		(14.714)		(15.557)		(14.773)		(15.082)		(14.783)	

Interaction x …	…FinMan	…WatLand	…EnergyEff	…ComImp	…Agri	…ManTransSaf
dCanada	0.391	0.662	−11.319	−0.008	2.780	−5.976
	(0.781)	(0.871)	(9.809)	(0.739)	(10.682)	(4.440)
Nb obs.	1110	1110	1110	1110	1110	1110
F	59.254	59.167	60.045	59.276	59.272	59.494
R^2	0.587	0.587	0.587	0.587	0.587	0.588
Adjusted R^2	0.576	0.577	0.577	0.576	0.576	0.577

Notes: ***p ≤ 0.001, **p ≤ 0.05, *p ≤ 0.1. *dMicro*, very small firms, is the omitted firm size category, *dConsPdct*, Consumer products, is the omitted industry category. *DummyYears* refers to the control variables for the assessment years. Compared to the omitted variables *d2015&d2019* only *d2016* is significant for all the regressions
EnergyEff represents the transformation ln (EnergyEff+11).
Agri represents the transformation ln (Agri+11).
ManTransSaf represents the transformation ln((ManTransSaf+5) *10 + 1).
CommInd represents the transformation ln (CommInd+1).

similarity approach, we maintain a consistent representation of the labels proposed by B-Lab. This capability, potentially extendable across various fields and research inquiries, paves the way for real-time analysis. Such a tool could provide policymakers with a dynamic overview of specific issues, offering a preliminary assessment of policy impacts ahead of more traditional evaluation techniques. This represents a significant stride towards mitigating the challenges posed by conventional questionnaire-based surveys, providing early insights into the efficacy of environmental policies and strategies.

7. Limitation and future works

Despite the promising results, the methodology used in this study is nonetheless subject to multiple limitations. The first limitation is inherent to the ZSTC task, which is considered the most challenging task for NLP models (Brown et al., 2020). The model only has access to the label and the text, without any examples or further explanations, which forces it to interpret everything by itself. It also increases the misinterpretation and ambiguity of the already complex natural language. A second implicit limitation arises the nature of natural language itself. By

definition, language is inherently ambiguous, presenting challenges in the study of semantics. This is what we are doing in this study: attempting to understand the meaning behind a paragraph to be able to label it.

Another limitation is related to the labels used. We chose the labels directly from the items that B-Lab uses to evaluate the environmental culture of a company. However, these items may not be precise enough for the concept we are seeking, and contacting B-Lab could help us obtain more appropriate labels to better target certain topics. A further limitation arises from the NLP model used in the ZSTC task. The model represents words in vectors based on the knowledge acquired during training, which is not specific to the SI problem addressed in this research. As mentioned in Section 3.2, this knowledge is derived from English Wikipedia and BooksCorpus.

Finally, there are some limitations that stem from the data. The research conducted is an exploratory study about the capacity of the applied methods to assess the environmental readiness of the companies. For our study, we chose only Canadian and American companies. This limits our research and our results since this sample, as mentioned in the data description, was chosen to be homogeneous mitigating the impact of different legal, cultural, and linguistic backgrounds. Additionally, we lost data because we could not find the companies' websites using the Wayback Machine. Because we needed to gather the websites of data near the certification date, we needed to recover the old website of the companies. For these reasons, our sample was downsized.

Building from these limitations, new avenues to improve this research are possible. First, expanding the sample by adding other countries will help to understand if the results can be generalizable for the B-Corp data. Moreover, using other datasets than the website of the B-Corp companies would be needed to increase the reliability of our results. Second, our literature review revealed that researchers have uncovered intriguing findings about the environmental strategies employed by companies. These findings were obtained by analyzing the companies' entire websites. Indeed, some researchers have found that sometimes the companies dedicate a full page to their environmental initiatives and strategies (Calabrese et al., 2021; Fernández-Vázquez and Sancho-Rodríguez, 2020). Additionally, by analyzing images on the website using their descriptions, researchers may better understand the messages that companies are conveying (Fernández-Vázquez and Sancho-Rodríguez, 2020). Thus, it could be interesting to compare our results obtained from analyzing only the homepage with those obtained from analyzing the full website including the images. This analysis could enhance our understanding of the results. Third, the NLP model used in this research is not specialized. Thus, creating a specialized model could be beneficial, especially when investigating more granular concepts in a precise topic. Indeed, advanced NLP models, like BART offer the possibility to be fine-tuned to different tasks. In other words, BART can be trained to accomplish tasks that have never been done before. Moreover, one can use the fine-tuning process to specialize BART to a particular topic. For example, BioBERT (Lee et al., 2020) was built to be used for several NLP tasks in the biomedical domain, outperforming the previous models. Taking the lead from this research, it is possible to create a specific model for SI that might have a greater comprehension of certain specific concepts. Lastly, we used the algorithm for the ZSTC, which, to the best of our knowledge, is the best to perform this task. However, the continued expansion and increasing work done in the field of NLP will possibly bring more advanced techniques that can better tackle this task.

CRediT authorship contribution statement

Pietro Cruciata: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Davide Pulizzotto:** Writing – review & editing, Software, Formal analysis, Data curation, Conceptualization. **Catherine Beaudry:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis, Conceptualization.

Data availability

Data will be made available on request.

Acknowledgement

The authors acknowledge the financial support from the Social Sciences and Humanities Research Council of Canada (SSHRC) partnership grant: Partnership for the Organization of Innovation and New Technologies (4POINT0) [grant number 435-2019-0111], the Natural Sciences and Engineering Research Council of Canada (NSERC) partnership grant: Canada Excellence Research Chair in Data Science for Real-Time Decision-Making [grant number CERC-2012-00002], and the Canada Research Chairs Program [grant number CRC-2020-00062].

The authors would also like to express their gratitude to the organizers and participants of the DRUID 2023 conference, CARMA 2023 (International Conference on Advanced Research Methods and Analytics), the Atlanta Conference on Science and Innovation Policy, and the Association Francophone pour le Savoir (ACFAS) for their valuable insights during the early stages of this research.

The authors are indebted to the editors and reviewers for their thoughtful guidance during the revision process. Finally, the authors thank Carl St-Pierre for his supervision during the statistical analysis and Mikaël Héroux-Vaillancourt for his contributions to data collection and processing.

Appendix 1. Variable description

Table 10
Variable description.

Variables	Type	Description	Transformation
Dependent variable			
BLabEnvIndex	Continuous	It assesses the company's environmental practices and commitment to sustainability, with scores ranging from 0 for companies performing poorly to a maximum of 66.1	
Control variables			
CommInd	Continuous	It assesses the company's influence on the external communities where it operates, encompassing aspects such as diversity, economic contributions, civic participation, and the impact on the supply chain, with scores ranging from 6.5 for companies performing poorly to a maximum of 115.2	$\ln(CommInd + 1)$

(*continued on next page*)

Table 10 (*continued*)

Variables	Type	Description	Transformation
GovernInd	Continuous	It assesses the company's overall mission, ethics, accountability and transparency, with scores ranging from 4.1 for companies performing poorly to a maximum of 24.3	
d2015&2019	Dummy (omitted)	Dummy variable taking the value 1 if the firm has done the assessment test in the year 2015 or 2019, and 0 otherwise.	
d2016	Dummy	Dummy variable taking the value 1 if the firm has done the assessment test in the year 2016, and 0 otherwise.	
d2017	Dummy	Dummy variable taking the value 1 if the firm has done the assessment test in the year 2017, and 0 otherwise.	
d2018	Dummy	Dummy variable taking the value 1 if the firm has done the assessment test in the year 2018, and 0 otherwise.	
d2020	Dummy	Dummy variable taking the value 1 if the firm has done the assessment test in the year 2020, and 0 otherwise.	
d2021	Dummy	Dummy variable taking the value 1 if the firm has done the assessment test in the year 2021, and 0 otherwise.	
Web-based environment culture indicators (from the PCA analysis)			
FinMan	Continuous	Financial and Management aspects related to the web-based environmental culture of the firm	
WatLand	Continuous	Web-based environmental culture indicator of the company's impact on water and land	
EnergyEff	Continuous	Web-based environmental culture indicator of the company's energy efficiency	$\ln(EnergyEff + 11)$
ComImp	Continuous	Web-based environmental culture indicator of the company impact on the community	
Agri	Continuous	Web-based environmental culture indicator of the company impact on agriculture practice and process	$\ln(Agri + 11)$
ManTransSaf	Continuous	Web-based environmental culture indicator of the company process put in place to reduce the impact on the environment of manufacturing and transportation process	$\ln((ManTransSaf + 5) *10 + 1)$
Other independent variables			
dCanada	Dummy	Dummy variable taking the value 1 if the firm is located in Canada, and 0 otherwise (it is an American company)	
dmicro	Dummy (omitted)	Dummy variable with value 1 if the firm has 0 to 9 employees, and 0 otherwise.	
dsmall	Dummy	Dummy variable taking the value 1 if the firm has 10 to 49 employees, and 0 otherwise.	
dmedium	Dummy	Dummy variable taking the value 1 if the firm has 50 to 250 employees, and 0 otherwise.	
dlarge	Dummy	Dummy variable taking the value 1 if the firm has >250 employees, and 0 otherwise.	
dAgricul	Dummy	Dummy variable taking the value 1 if the firm in the Agriculture industry category, and 0 otherwise.	
dBuild	Dummy	Dummy variable taking the value 1 if the firm in the Building industry category, and 0 otherwise.	
dConsPdct	Dummy (omitted)	Dummy variable taking the value 1 if the firm in Consumer Products & Services industry category, and 0 otherwise.	
dEducationTr	Dummy	Dummy variable taking the value 1 if the firm in the Education & Training Services industry category, and 0 otherwise.	
dEnergyEnv	Dummy	Dummy variable taking the value 1 if the firm operates in the Energy & Environmental Services industry category, and 0 otherwise.	
dFinLegserv	Dummy	Dummy variable taking the value 1 if the firm operates in the Financial &Legal services industry category, and 0 otherwise.	
dHealthHuman	Dummy	Dummy variable taking the value 1 if the firm operates in the Health & Human Services industry category, and 0 otherwise.	
dMedRestHosp	Dummy	Dummy variable taking the value 1 if the firm operates in the Media, Restaurant, Hospitality & Travel industry category, and 0 otherwise.	
dRetTransLog	Dummy	Dummy variable taking the value 1 if the firm operates in the Retail, Transportation & Logistics industry category, and 0 otherwise.	
dBusinessPro	Dummy	Dummy variable taking the value 1 if the firm operates in the Business Products & Services industry category, and 0 otherwise.	

Appendix 2. Correlations and descriptive statistics

Table 11
Descriptive statistics.

Variables	Mean	Median	Std. Dev.	Min	Max	Skewness	Kurtosis
ln (*CommInd* + 1)	3.308	3.245	0.429	2.015	4.755	0.2703664	2.76969
CommInd	29.061	24.650	14.012	6.500	115.200	1.396835	5.568
GovernInd	14.343	14.800	3.826	4.100	24.300	−0.308	2.790
d2016	0.022	0	0.145	0	1	6.584	44.355
d2017	0.159	0	0.366	0	1	1.863	4.472
d2018	0.260	0	0.439	0	1	1.095	2.199
d2020	0.198	0	0.399	0	1	1.517	3.301
d2021	0.054	0	0.226	0	1	3.948	16.590
d2015&2019 (omitted group)	0.307	0	0.461	0	1	0.839	1.703
BLabEnvIndex	16.841	12.100	12.587	0	66.100	0.985	3.253
dmicro (omitted group)	0.458	0	0.498	0	1	0.170	1.029
dsmall	0.332	0	0.471	0	1	0.716	1.512
dmedium	0.158	0	0.365	0	1	1.879	4.530
dlarge	0.053	0	0.224	0	1	3.984	16.87
dAgricul	0.027	0	0.162	0	1	5.833	35.028
dBuild	0.036	0	0.186	0	1	4.979	25.787
dConsumPdt (omitted group)	0.283	0	0.451	0	1	0.964	1.930
dEducationTr	0.032	0	0.177	0	1	5.279	28.867
dEnergyEnvir	0.048	0	0.213	0	1	4.242	18.994
dFinLegservic	0.128	0	0.334	0	1	2.228	5.964
dHealthHuman	0.037	0	0.189	0	1	4.910	25.112
dMedRestHosp	0.031	0	0.172	0	1	5.448	30.679
dRetTransLog	0.028	0	0.165	0	1	5.730	33.835
dBusinessPro	0.350	0	0.477	0	1	0.627	1.393
dCanada	0.176	0	0.381	0	1	1.705	3.905
FinMan	0	−0.328	1	−2.190	4.516	1.416	4.976

(*continued on next page*)

Table 11 (*continued*)

Variables	Mean	Median	Std. Dev.	Min	Max	Skewness	Kurtosis
WatLand	0	−0.115	1	−3.062	5.536	0.753	4.415
EnergEff	0	−0.113	1	−4.776	6.992	1.217	9.368
ComImp	0	−0.025	1	−3.006	3.576	0.145	3.346
Agri	0	−0.164	1	−3.816	6.235	1.845	9.658
ManTransSaf	0	−0.154	1	−3.092	5.746	0.924	5.938
ln (*EnergEff* + 11)	2.394	2.388	0.089	1.828	2.890	0.292	8.505
ln (*Agri* + 11)	2.394	2.383	0.086	1.972	2.847	1.175	7.729
ln ((*ManTransSaf* + 5) *10 + 1)	3.913	3.901	0.193	3	4.686	−0.130	4.806

Note: Number of observations = 1110.

Table 12
Correlation matrix.

Variables		1	2	3	4	5	6	7	8	9	10	11	12	13
Impact area environment	1	1												
CommInd	2	−0.034	1											
GovernInd	3	−0.170 *	−0.021	1										
d2016	4	0.005	0.034	0.032	1 *									
d2017	5	0.035	0.068 *	−0.117 *	−0.065 *	1								
d2018	6	0.009	0.158 *	−0.105 *	−0.088 *	−0.258 *	1							
d2020	7	0.002	−0.106 *	0.146 *	−0.074 *	−0.216 *	−0.294 *	1						
d2021	8	−0.016	0.012	0.102 *	−0.036	−0.104 *	−0.142 *	−0.119 *	1					
d2015&2019	9	−0.032	−0.129 *	0.006	−0.099 *	−0.289 *	−0.394 *	−0.330 *	−0.159 *	1				
dmicro	10	−0.135 *	0.309 *	−0.015	0.000	0.005	−0.009	−0.003	0.020	−0.002	1			
dsmall	11	0.072 *	−0.178 *	0.008	0.014	0.028	−0.008	−0.024	−0.025	0.014	−0.647 *	1		
dmedium	12	0.076 *	−0.127 *	0.003	−0.013	−0.026	0.036	0.002	−0.016	−0.003	−0.397 *	−0.305 *	1	
dlarge	13	0.025	−0.106 *	0.011	−0.008	−0.026	−0.022	0.054	0.032	−0.018	−0.218 *	−0.167 *	−0.103 *	1
dAgricul	14	0.225 *	−0.007	−0.078 *	0.013	−0.027	0.040	−0.041	−0.040	0.034	−0.042	0.071 *	−0.042	0.010
dBuild	15	0.192 *	−0.037	−0.047	0.005	−0.045	0.040	0.001	−0.004	−0.003	−0.032	0.059 *	−0.017	−0.024
dConsPdct	16	0.362 *	0.154 *	−0.120 *	−0.025	0.043	−0.017	0.014	0.036	−0.040	−0.075 *	−0.009	0.052	0.101 *
dEducationTr	17	−0.147 *	−0.083 *	0.057	0.043	0.031	−0.062 *	−0.002	0.001	0.022	−0.025	−0.010	0.046	0.002
dEnergyEnvir	18	0.348 *	−0.116 *	−0.090 *	−0.033	0.006	0.002	0.026	−0.054	0.007	−0.070 *	0.094 *	0.008	−0.053
dFinLegservic	19	−0.292 *	−0.111 *	0.095 *	0.017	−0.012	0.025	0.006	−0.020	−0.015	0.054	−0.018	−0.047	−0.007
dHealthHuman	20	−0.133 *	−0.003	0.030	−0.029	0.006	0.004	−0.038	0.038	0.015	−0.017	−0.047	0.046	0.060 *
dMedRestHosp	21	−0.051	0.042	0.049	0.046	0.008	0.026	0.003	−0.019	−0.039	0.015	−0.047	0.052	−0.019
dRetTransLog	22	0.058	0.057	−0.023	−0.025	−0.014	−0.013	−0.029	0.032	0.042	0.053	−0.050	−0.013	0.009
dBusinessPro	23	−0.339 *	−0.002	0.088 *	0.008	−0.021	−0.014	0.009	0.000	0.020	0.083 *	−0.012	−0.048	−0.081 *
dCanada	24	0.001	0.022	0.001	−0.020	−0.052	−0.026	0.032	0.089 *	0.001	0.027	0.007	−0.044	−0.004
FinMan	25	0.527 *	0.012	−0.112 *	−0.047	0.070 *	0.014	0.013	−0.016	−0.057	0.051	0.023	−0.033	−0.107 *
WatLand	26	−0.196 *	0.002	0.052	0.012	0.033	−0.056	0.001	0.038	0.004	0.108 *	−0.044	−0.086 *	−0.009
EnergEff	27	0.062 *	−0.104 *	−0.004	0.018	−0.024	−0.037	0.041	−0.014	0.020	0.012	0.037	−0.059	−0.010
ComImp	28	−0.057	0.031	−0.044	−0.028	−0.035	0.069 *	0.005	0.011	−0.039	−0.031	−0.004	0.041	0.012
Agri	29	0.156 *	0.111 *	−0.079 *	0.009	−0.028	−0.004	−0.028	0.021	0.037	0.025	0.012	−0.063 *	0.020
ManTransSaf	30	0.001	0.048	0.031	−0.075 *	−0.011	−0.075 *	0.072 *	0.057	0.015	−0.066 *	0.029	−0.005	0.095 *

Notes: $^{*}p \leq 0.05$.
EnergEff represents the transformation ln (EnergEff+11).
Agri represents the transformation ln (Agri+11).
ManTransSaf represents the transformation ln((ManTransSaf+5) *10 + 1).
CommInd represents the transformation ln (CommInd+1).

Table 13
Correlation table (continued).

Variables		14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
dAgricul	**14**	1																
dBuild	**15**	−0.032	1															
dConsPdct	**16**	−0.105 *	−0.121 *	1														
dEducationTr	**17**	−0.031	−0.035	−0.115 *	1													
dEnergyEnvir	**18**	−0.037	−0.043	−0.141 *	−0.041	1												
dFinLegservic	**19**	−0.064 *	−0.074 *	−0.241 *	−0.0701 *	−0.086 *	1											
dHealthHuman	**20**	−0.033	−0.038	−0.123 *	−0.0359	−0.044	−0.075 *	1										
dMedRestHosp	**21**	−0.030	−0.034	−0.112 *	−0.0325	−0.040	−0.0681 *	−0.035	1									
dRetTransLog	**22**	−0.028	−0.033	−0.107 *	−0.031	−0.038	−0.0649 *	−0.033	−0.030	1								
dBusinessPro	**23**	−0.122 *	−0.142 *	−0.461 *	−0.1345 *	−0.165 *	−0.2813 *	−0.144 *	−0.131 *	−0.125 *	1							
dCanada	**24**	−0.062 *	−0.026	−0.001	−0.0043	−0.026	−0.0421	0.010	−0.013	0.008	0.073 *	1						
FinMan	**25**	0.051	0.016	0.233 *	−0.0974 *	0.295 *	−0.1322 *	−0.147 *	0.011	0.055	−0.211 *	0.042	1					
WatLand	**26**	−0.103 *	0.062 *	−0.210 *	−0.0218	−0.178 *	0.036	−0.048	0.025	−0.041	0.295 *	0.079 *	0	1				
EnergEff	**27**	−0.086 *	0.194 *	−0.208 *	−0.0469	0.335 *	0.0733 *	−0.066 *	−0.105 *	−0.050	0.047	−0.014	−0.048	0.000	1			
ComImp	**28**	−0.035	0.079 *	−0.116 *	0.1672 *	0.097 *	0.0615 *	0.098 *	0.021	−0.023	−0.097 *	−0.004	0.000	0.000	−0.003	1		
Agri	**29**	0.386 *	0.007	0.144 *	−0.075 *	−0.188 *	0.0008	−0.048	−0.103 *	−0.002	−0.101 *	−0.024	−0.038	0.010	0.035	0.002	1	
ManTransSaf	**30**	−0.062 *	−0.073 *	0.246 *	−0.0743 *	−0.171 *	0.0094	−0.011	−0.091 *	0.043	−0.063 *	0.019	−0.051	0.003	0.026	0.008	−0.007	1

Notes: *p ≤ 0.05.
EnergyEff represents the transformation ln (EnergyEff+11).
Agri represents the transformation ln (Agri+11).
ManTransSaf represents the transformation ln((ManTransSaf+5) *10 + 1).
CommInd represents the transformation ln (CommInd+1).

Appendix 3. Robustness check

Table 14
Base regression and robustness check.

Variables	Complete		Vce (Robust)		Tobit	
CommInd	−1.613	**	−1.613	**	−1.579	**
	(0.648)		(0.737)		(0.737)	
GovernInd	−0.050		−0.050		−0.050	
	(0.068)		(0.065)		(0.065)	
dsmall	0.885		0.885		0.936	
	(0.595)		(0.583)		(0.583)	
dmedium	3.276	***	3.276	***	3.353	***
	(0.756)		(0.760)		(0.755)	
dlarge	3.088	***	3.088	***	3.168	***
	(1.182)		(1.140)		(1.129)	
dCanada	0.794		0.794		0.770	
	(0.656)		(0.663)		(0.660)	
dAgricul	7.407	***	7.407	***	7.413	***
	(1.702)		(2.589)		(2.561)	
dBuild	7.430	***	7.430	***	7.474	***
	(1.495)		(2.097)		(2.087)	
dEducationTr	−12.385	***	−12.385	***	−12.635	***
	(1.561)		(1.171)		(1.227)	
dEnergyEnvir	9.457	***	9.457	***	9.430	***
	(1.504)		(1.818)		(1.800)	
dFinLegservic	−13.002	***	−13.002	***	−12.984	***
	(0.931)		(0.772)		(0.765)	
dHealthHuman	−10.176	***	−10.176	***	−10.384	***
	(1.438)		(1.177)		(1.210)	
dMedRestHosp	−7.773	***	−7.773	***	−7.767	***
	(1.545)		(1.895)		(1.873)	
dRetTransLog	−1.736		−1.736		−1.724	
	(1.552)		(1.565)		(1.548)	
dBusinessPro	−9.056	***	−9.056	***	−9.076	***
	(0.747)		(0.802)		(0.793)	
FinMan	4.144	***	4.144	***	4.159	***
	(0.287)		(0.392)		(0.389)	
WatLand	−0.664	**	−0.664	**	−0.657	**
	(0.271)		(0.324)		(0.320)	
EnergEff	2.515		2.515		2.645	
	(3.245)		(4.484)		(4.464)	
ComImp	−0.588	**	−0.588	**	−0.593	**
	(0.260)		(0.283)		(0.281)	
Agri	15.548	***	15.548	***	15.537	***
	(3.363)		(3.862)		(3.818)	
ManTransSaf	1.249		1.249		1.231	
	(1.384)		(1.603)		(1.587)	
DummyYears	yes		yes		yes	
Constant	−21.839	*	−21.839		−22.202	
	(11.614)		(14.758)		(14.633)	
Nb obs.	1110		1110		1110	
F	59.083	***	61.566	***	62.570	***
R^2	0.587		0.587			
Adjusted R^2	0.577		0.577			
Pseudo R^2					0.112	
Kurtosis						
Durbin-Watson:						
d_l						
d_u						
4-du						
4-dl						
Breusch-Pagan						
Limits (left-censored)					0 ($N = 7$)	
(right-censored)					66.1 ($N = 1$)	

Notes: ***p ≤ 0.001, **p ≤ 0.05, *p ≤ 0.1. The Breusch-Pagan test is a χ^2 with 1 degree of freedom. “dl” and “du” are the lower and upper critical values of the Durbin-Watson test. Since ‘2.004’ falls between “du” and “4-du” there is no autocorrelation.
dMicro, very small firms, is the omitted firm size category, *dConsPdct*, Consumer products, is the omitted industry category.
DummyYears refers to the control variables for the assessment years. Compared to the omitted variables *d2015&d2019* only *d2016* is significant for all the regressions

but Basic Reg.
EnergyEff represents the transformation ln *(EnergyEff + 11)*.
Agri represents the transformation ln *(Agri + 11)*.
ManTransSaf represents the transformation ln ((ManTransSaf + 5) *10 + 1).
CommInd represents the transformation *ln (CommInd + 1)*.

References

Aguilar-Fernández, M.E., Otegi-Olaso, J.R., 2018. Firm size and the business model for sustainable innovation. Sustainability 10 (12), 4785.
Aguilera-Caracuel, J., Ortiz-de-Mandojana, N., 2013. Green innovation and financial performance: an institutional approach. Organ. Environ. 26 (4), 365–385.
Axenbeck, J., Breithaupt, P., 2021. Innovation indicators based on firm websites—which website characteristics predict firm-level innovation activity? PloS One 16 (4), e0249583.
Blasi, S., Sedita, S.R., 2022. Mapping the emergence of a new organisational form: an exploration of the intellectual structure of the B Corp research. Corp. Soc. Respon. Environ. Manag. 29 (1), 107–123. https://doi.org/10.1002/csr.2187.
Blazquez, D., Domenech, J., 2018. Web data mining for monitoring business export orientation. Technol. Econ. Dev. Econ. 24 (2), 406–428.
Bossle, M.B., de Barcellos, M.D., Vieira, L.M., Sauvée, L., 2016. The drivers for adoption of eco-innovation. J. Clean. Prod. 113, 861–872.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020. Language models are few-shot learners. Adv. Neural Inf. Proces. Syst. 33, 1877–1901.
Calabrese, A., Costa, R., Ghiron, N.L., Tiburzi, L., Pedersen, E.R.G., 2021. How sustainable-orientated service innovation strategies are contributing to the sustainable development goals. Technol. Forecast. Soc. Chang. 169, 120816.
Callison, C., 2003. Media relations and the internet: how fortune 500 company web sites assist journalists in news gathering. Public Relat. Rev. 29 (1), 29–41.
Cao, K., Gehman, J., Grimes, M.G., 2017. Standing out and fitting. In: Charting the Emergence of Certified B Corporations by Industry and Region. In *Hybrid Ventures*. Emerald Publishing Limited, pp. 1–38. https://www.emerald.com/insight/content/doi/10.1108/S1074-754020170000019001/full/html.
Connelly, B.L., Certo, S.T., Ireland, R.D., Reutzel, C.R., 2011. Signaling theory: a review and assessment. J. Manag. 37 (1), 39–67.
Cormier, D., Magnan, M., 2015. The economic relevance of environmental disclosure and its impact on corporate legitimacy: an empirical investigation. Bus. Strateg. Environ. 24 (6), 431–450. https://doi.org/10.1002/bse.1829.
Cowan, K., Guzman, F., 2020. How CSR reputation, sustainability signals, and country-of-origin sustainability reputation contribute to corporate brand performance: an exploratory study. J. Bus. Res. 117, 683–693.
de Azevedo Rezende, L., Bansi, A.C., Alves, M.F.R., Galina, S.V.R., 2019. Take your time: examining when green innovation affects financial performance in multinationals. J. Clean. Prod. 233, 993–1003.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv: 1810.04805*.
Doran, J., Ryan, G., 2012. Regulation and firm perception, eco-innovation and firm performance. Eur. J. Innov. Manag. 15 (4), 421–444.
Eccles, R.G., Ioannou, I., Serafeim, G., 2014. The impact of corporate sustainability on organizational processes and performance. Manag. Sci. 60 (11), 2835–2857.
Fernández-Vázquez, J.-S., Sancho-Rodríguez, Á., 2020. Critical discourse analysis of climate change in IBEX 35 companies. Technological Forecasting and Social Change 157, 120063.
Goczol, J., Scoubeau, C., 2003. Corporate communication and strategy in the field of projects. Corp. Commun. Int. J. 8 (1), 60–66.
Gök, A., Waterworth, A., Shapira, P., 2015. Use of web mining in studying innovation. Scientometrics 102 (1), 653–671.
Hair, J., Tatham, R., Anderson, R., Black, W., 1998. Multivariate Data Analysis Prentice-Hall London, 5th ed. UK.
Hermundsdottir, F., Aspelund, A., 2021. Sustainability innovations and firm competitiveness: a review. J. Clean. Prod. 280, 124715.
Héroux-Vaillancourt, M., Beaudry, C., Rietsch, C., 2020. Using web content analysis to create innovation indicators—what do we really measure? Quantitative Science Studies 1 (4), 1601–1637.
Hoehn-Weiss, M.N., Karim, S., 2014. Unpacking functional alliance portfolios: how signals of viability affect young firms' outcomes. Strateg. Manag. J. 35 (9), 1364–1385.
Jiang, C., Yin, C., Tang, Q., Wang, Z., 2023. The value of official website information in the credit risk evaluation of SMEs. J. Bus. Res. 169, 114290.
Jung Moon, S., Hyun, K.D., 2014. Online media relations as an information subsidy: quality of *fortune 500* Companies' websites and relationships to media salience. Mass Commun. Soc. 17 (2), 258–273. https://doi.org/10.1080/15205436.2013.779716.
Ketata, I., Sofka, W., Grimpe, C., 2015. The role of internal capabilities and firms' environment for sustainable innovation: evidence for G ermany. R&D Manag. 45 (1), 60–75.
Kim, S., Schifeling, T., 2016. Varied incumbent behaviors and mobilization for new organizational forms: The rise of triple-bottom line business amid both corporate social responsibility and irresponsibility. *Available at SSRN*, 2794335. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2794335.
Kinne, J., Axenbeck, J., 2020. Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics 125 (3), 2011–2041.
Kinne, J., Lenz, D., 2021. Predicting innovative firms using web mining and deep learning. PloS One 16 (4), e0249071.
Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240.
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv Preprint arXiv:1910.13461*.
Libaers, D., Hicks, D., Porter, A.L., 2016. A taxonomy of small firm technology commercialization. Ind. Corp. Chang. 25 (3), 371–405.
Liute, A., De Giacomo, M.R., 2022. The environmental performance of UK-based B Corp companies: an analysis based on the triple bottom line approach. Business Strategy and the Environment 31 (3), 810–827.
Magnusson, P., Westjohn, S.A., Zdravkovic, S., 2011. "What? I thought Samsung was Japanese": accurate or not, perceived country of origin matters. Int. Mark. Rev. 28 (5), 454–472.
Mavlanova, T., Benbunan-Fich, R., Koufaris, M., 2012. Signaling theory and information asymmetry in online commerce. Inf. Manag. 49 (5), 240–247.
Mazzei, A., 2010. Promoting active communication behaviours through internal communication. Corporate Communications: An International Journal 15 (3), 221–234.
Paelman, V., Van Cauwenberge, P., Vander Bauwhede, H., 2020. Effect of B Corp certification on short-term growth: European evidence. Sustainability 12 (20), 8459.
Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving Language Understanding by Generative Pre-Training.
Romi, A., Cook, K.A., Dixon-Fowler, H.R., 2018. The Influence of Social Responsibility on Employee Productivity and Sales Growth: Evidence from Certified B Corps. Sustainability Accounting, Management and Policy Journal.
Tariq, A., Badir, Y.F., Tariq, W., Bhutta, U.S., 2017. Drivers and consequences of green product and process innovation: a systematic review, conceptual framework, and future outlook. Technol. Soc. 51, 8–23.
WCED, S.W.S., 1987. World commission on environment and development. Our Common Future 17 (1), 1–91.
Weale, A., 1992. The new politics of pollution. Manchester University Press. https://books.google.com/books?hl=it&lr=&id=99JRAQAAIAAJ&oi=fnd&pg=PR7&dq=weale+1992+triple+bottom+line&ots=hLEGGwKUrS&sig=WHlkV_gswckdQTeezJ1_u11xr-o.
Yildiz, H., Tahali, S., Trichina, E., 2023. The adoption of the green label by SMEs in the hotel sector: a leverage for reassuring their customers. J. Enterp. Inf. Manag.
Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv Preprint arXiv:1909.00161*.
Zhu, J., Hua, W., 2017. Visualizing the knowledge domain of sustainable development research between 1987 and 2015: a bibliometric analysis. Scientometrics 110 (2), 893–914.

Pietro Cruciata is a PhD candidate at Polytechnique Montreal. Prior to this, he earned a master's degree in Statistics from the University of Bologna. His research focuses on developing and validating innovation indicators by leveraging information from companies' websites and the methodological advancements in the field of Natural Language Processing.

Davide Pulizzotto has a doctorate in semiology (University of Quebec at Montreal, 2020). He is a postdoctoral researcher at the Interuniversity Research Center on Science and Technology (CIRST) at the University of Quebec at Montreal and a research associate at the Partnership for the Organization of Innovation (4POINT0) at Polytechnique Montreal.

Catherine Beaudry holds a master's degree and PhD in economics from the University of Oxford. She is a full professor at the Department of Mathematics and Industrial Engineering and holds the Tier 1 Canada Research Chair in Management and Economics of Innovation (Chair Innovation) at Polytechnique Montréal, and leads the Partnership for the Organization of Innovation and New Technologies n4POINT0). She is a Fellow of the Academy of Social Sciences of the Royal Society of Canada and has been awarded the Prix Acfas Jacques-Rousseau 2022.