5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16463

0-shot text classification for web-based environmental indicators: Pilot study on B-Corp data

Pietro Cruciata[1], Davide Pulizzotto[1], Mikaël Héroux-Vaillancourt[1], Catherine Beaudry[1]
[1]Polytechnique Montréal, Canada

Abstract

This paper proposes a tool that uses web-based information to generate a proxy for the environmental culture indicator developed by B-Lab. The tool is based on recent advances in Natural Language Processing (NLP), such as pre-trained language models like BART that better capture the semantic facets of natural language. The algorithm and data provide several advantages, including real-time analysis, minimal building cost, granularity, and a large sample size, making it appealing. The Zero-shot text classification task is used to create an indicator of companies' environmental culture, which was chosen due to the urgency created by recent climatic events, pushing for increased environmental protection and sustainability culture promotion. The tool was tested on the B-CORP dataset, which provides scores on environmental performance. Results indicate that scores for certain environmental topics generated by the tool are correlated with B-Lab's environmental indicator. This research open door to the possibility of predicting the environmental readiness of the companies base on web-based indicators.

Keywords: *Natural Language Processing, Zero-shot text classification, Sustainable Innovation*

1. Introduction

Governments acknowledge that innovation is a critical driver of economic growth, and thus, they allocate funds to support companies' research and development (R&D) projects. If the government can allocate these funds efficiently, it can lead to accelerated economic development. Typically, policy makers and governments rely on administrative data and questionnaire-based surveys to assess the quantitative impact of these R&D investments. While these sources of information may serve their intended purpose, they often have significant limitations. For instance, connecting particular policy instruments to alterations in firm performance, as assessed by administrative data, poses a challenge. Furthermore, surveys reliant on questionnaires (particularly those on a large-scale, such as the biennial European CIS or the annual MIP) are inadequate in terms of regional granularity, scope, timeliness, and conducting such surveys incurs significant costs. (Axenbeck and Breithaupt 2021). Due to all these factors, conventional indicators of innovation seldom offer a comprehensive view of the effects of policy combinations (Kinne and Lenz 2021). Alternative or complementary to these sources are web-based unstructured textual data. Among their advantages, the rapidity of their evolution, their increasing quantity, variety, and availability opened new possibilities for policy makers and researchers (Gök, Waterworth, and Shapira 2015). As policy makers now turn their attention to adaptation to climate change, mitigation of its effect, and generally a better socio-environmental impact of their policies, sustainable innovation is perceived as a key solution.

This paper proposes a method that employs web-based information to create an environmental culture indicator proxy of the real environmental culture indicator developed by B-Lab. Indeed, the environmental impact of the companies plays a crucial role in the triple bottom line framework established by John Elkington in the 1990s, which highlight the equal importance of social, environment and economics goals to pursue sustainable innovation. Moreover, the main part of the external communication of companies relies on their websites. Assuming that corporate websites are written with the intention of highlighting the 'best' qualities of the firm, our intuition is that there will be a correalation between the web-based environmental culture indicator and the real environmental culture indicator developed by B-Lab.

2. Data and Methodology

2.1. Data

To create and test the tool we use two types of data:

1. The full-text of the companies' websites that are B-Corp certified.
2. The B-Corp data.

B-Lab publicly releases the dataset with all the companies certified and the scores received. The scores include one main indicator, "overall score", which is an aggregation of five other indicators evaluating specific dimensions: governance, customers, workers, community, and environment. These dimensions are in turn divided into several items. In this paper, we focus solely on the B-Corp indicator concerning the "impact area environment" and as this is a pilot study we use only a subset of the B-Corp data limited to the Canadian and American companies. To create a corpus for each company, we identified URLs from the B-Corp data and downloaded text only from their homepages using the Wayback Machine. We used the Wayback Machine to download the pages as it was crucial to retrieve websites close to the certification date. We found 1741 company websites using the Wayback Machine. Next, we filtered the websites, choosing only English webpages and the most recent audit. Since a company can be certified more than once, we removed duplicates. Thus, the final sample has a total of 1110 firms, with 82% of companies from the US and 18% from Canada.

2.2. Methodology

Once the data has been prepared, the first step of the analysis consists in understanding the text of the corporate websites. Instead of counting specific keywords about predetermined topics, like most of the literature in social science, we use the Zero-shot text classification method which is a Natural language processing (NLP) task that is designed to answer the question: "Is this text about label X?" The answer to this question is an indicator of the confidence that the given text is about label X. The labels that we used for the purpose are the names of the items that compose the B-Corp environmental certification. Using the NLP model BART with the ZSTC, we aim to extrapolate the importance of a label. Then, our second task is simply calculating the Pearson correlations to measure which different items, among the ones included by B-Lab to evaluate the environmental impact certification, detected by BART in the text of the corporate websites are good proxies for the environmental score obtained by these firms.

The core of the tool is the Natural Language Processing (NLP) model "Bidirectional and Auto-Regressive Transformers" (BART)(Lewis et al. 2019), a transformed-based deep learning model for NLP developed by Facebook AI combining the most important characteristics of BERT and GPT. BART was pre-trained on English Wikipedia and BooksCorpus, using a two steps process: first, the text is changed by adding a noise factor (e.g., changing the words randomly), then, the model learns to reconstruct the original text. This new approach allowed BART to reach state-of-the-art performances in several NLP challenges.

We build the tool using BART on the Zero-shot text classification (ZSTC) task. ZSTC is a challenging task on the realms of the Natural Language Understanding problems, which require 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.

Performing the ZSTC requires choosing the labels and the corpus. Since we want to create a web-based environmental culture indicator, we use as labels the items that compose the 'impact area environment' index in B-Corp data (Table 1). After trying several settings, we decided to split each website into groups of 3 sentences to create our corpus. Indeed, we notice that the ZSTC performs better when the input is a text longer than a single sentence and smaller than the full website. Therefore, for each website, we perform the ZSTC on each group of sentences. It is important to highlight that the ZSTC produces a score among the several classes using cosine similarity metrics computation between the word-embedding vectors created by BART representing the label and the word embedding representation of the target corpus. The score is in a range from 0 to 1 and can be the same for more than one label. Then, to prepare the results for the Pearson correlation test, we take the average scores of each label for each website. In this way, for each website, we have the averaged results of the ZSTC for all the labels in Table 1 and the B-Corp data.

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
- land office plant
- Environmental education information
- Environmental management
- Environmentally innovative agricultural process
- Environmentally innovative manufacturing process
- Environmentally innovative wholesal process
- green investing
- green lending
- inputs
- land life
- landwildlife 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(2023)

3. Results

3.1. Zero-shot text classification

Table 2 shows a sample of the ZSTC results, which range from 0 to 1, representing the average score of each label for all the 1110 websites. Considering the mean score, "designed to conserve wholesale process", "inputs" and "safety" are the labels with the higher average. In other words, in the full text of the website, on average there are more groups of sentences that, according to the model, refer to these labels. All labels have a minimum score of around 0, but the maximum values are not uniform. 17 out of the 31 labels have a maximum score over 0.90, meaning that at least once, each label is predominant on a website according to the model. On the other hand, the labels "clear burning energy," "green lending," "community," "designed to conserve manufacturing process," and "land wildlife conservation" have the lowest maximum value in the table, with the last label having a maximum score of less than 0.5. This suggests that our sample does not generally include websites where the "land wildlife conservation" label is more prominent than the other labels. Additionally, "land wildlife conservation," "environmentally innovative manufacturing process," and "green investing" have low average scores, indicating that the model rarely finds groups of sentences that correspond to these labels.

Generally, the minimum score value is closer to the mean than the maximum value, except for the scores of the labels "designed to conserve wholesale process" and "inputs." This suggests that while the other labels are frequently found on websites with low scores or not found at all, the model only considers the "designed to conserve wholesale process" and "inputs" labels when their scores are significantly higher than the others. Additionally, the scores in the table do not follow a normal distribution, as evidenced by the mean and median being unequal and the third quartile being closer to the minimum value, indicating possible outliers.

Table 2: Sample of ZSTC results averaged for all the companies

Labels	**mean**	**std**	**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
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

3.2. Correlation results

Table 3 shows the Pearson correlation results between each web-based environmental indicator and environmental indicator of B-Corp. As aforementioned, we ensure the normality of all the variables transforming them and testing skewness and kurtosis.

Table 3: Pearson Correlation results

Labels	r	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 labesl with ** are transformed with the formula ln((label *10)+1)
The labels with *** are transformed with the formula ln((label *100)+1)

To ease the interpretation of the results we divide Table 3 in 3 parts. The lower box of the table contains the variables that have either negative or a null correlation with the B-Corp variable. Only the last 3 have p-values < 0,005 with the last two presenting p-value<0,001. Additionally, the last two labels have a score that is weakly inversely related. The label that are not in the two box are the one less important. The variables with a positive correlation have a p-value less than 0.001, and three of them exhibit a correlation close to 0.30. However, when the correlation decreases and approaches zero, the p-value becomes insignificant. Finally, the most interesting is box on the highest part of Table 2. In this box, we find the labels that have the highest correlation with the environmental variable of B-Corp. They all have a p-value< 0,001 and a correlation higher than 0,30. Among these labels, the first four have a correlation higher than 0,4 with green investing that reaches almost 50% (0.497).

4. Conclusion

The goal of the research is to verify whether the ZSTC task can be used to create environmental indicators that are correlated with the real environmental indicators developed by B-Lab. Once we perform the ZSTC, we find significant correlations between most of the ZSTC scores and the environmental index measured by B-Lab. Specifically, we find that the scores of the topics: green investing, environmentally innovative wholesale processes, resource conservation, and green lending are the most correlated with the "impact area environment" indicator developed by B-Lab. This means that companies with a higher score on the environmental indicator created by B-Lab are likely to talk about the aforementioned topics.

Despite the promising results, our research presents three main limits. The first limit is inherent to the ZSTC task, which is considered one of the most challenging tasks for NLP models (Brown et al. 2020). This approach resembles an unsupervised method, which makes it generalizable. However, 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. This can increase the misinterpretation and ambiguity of the already complicated natural language. Although this limit is intrinsic to the methodology, using a different state-of-the-art model instead of BART could potentially yield better results. It may be worth exploring other pre-trained language models such as LLAMA(Touvron et al. 2023) and PALM(Chowdhery et al. 2022) which have shown excellent performance in various NLP tasks. These models may have better capabilities to interpret complex natural language and reduce misinterpretation and ambiguity. The second limit 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 leaving to the model the interpretation of certain concepts. To overcome this limit, we could contact B Lab to obtain more appropriate labels that are specifically designed for the purpose of targeting certain environmental topics to reduce ambiguity and improve the accuracy of the results. Also, collaborating with domain experts, such as environmental scientists or sustainability practitioners, could also provide valuable insights for selecting appropriate labels. Finally, the third limit is connected to the correlation results. Indeed, the Pearson correlation partially explains the correlation between the items and the environmental indicator. For the third limit, including control variables in a regression analysis can provide further insight into the relationship between the ZSTC score and the B-Lab variable, thereby enhancing our understanding of the observed correlation. Additionally, the regression analysis will allow us to predict the score that B-Lab can give to the companies based on their website text. For instance, it could be possible to examine companies that sought certification but were unable to obtain it, thus distinguishing between greenwashing and genuine green compliance.

References

Axenbeck, Janna, and Patrick Breithaupt. 2021. "Innovation Indicators Based on Firm Websites—Which Website Characteristics Predict Firm-Level Innovation Activity?" *PloS One* 16(4):e0249583.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33:1877–1901.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, and Sebastian Gehrmann. 2022. "Palm: Scaling Language Modeling with Pathways." *ArXiv Preprint ArXiv:2204.02311.*

Gök, Abdullah, Alec Waterworth, and Philip Shapira. 2015. "Use of Web Mining in Studying Innovation." *Scientometrics* 102(1):653–71.

Kinne, Jan, and David Lenz. 2021. "Predicting Innovative Firms Using Web Mining and Deep Learning." *PloS One* 16(4):e0249071.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. "Bart: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension." *ArXiv Preprint ArXiv:1910.13461.*

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar. 2023. "Llama: Open and Efficient Foundation Language Models." *ArXiv Preprint ArXiv:2302.13971.*