CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17800

Structuring and extracting sustainability information from corporate websites SMEs: A pilot test on textile firms

Francisco Javier Rodríguez-Ruiz[1], Ana Garcia-Bernabeu[2]

[1]Department of Textile and Paper Engineering, Universitat Politècnica de València, Spain, [2]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain.

How to cite: Rodríguez-Ruiz, F. J.; Garcia-Bernabeu, A. 2024. Structuring and extracting sustainability information from corporate websites SMEs: A pilot test on textile firms. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17800

Abstract

In recent years, heightened awareness of environmental, social, and governance (ESG) issues has spurred a growing demand for sustainability-related data. While large corporations progress towards disclosing non-financial information, small and medium-sized enterprises (SMEs) face limitations due to the absence of standardized frameworks for reporting sustainable data. This paper aims to elucidate the process of developing a sustainability indicator framework by utilizing web-available information, encompassing the collection, processing, and analysis of ESG indicators. The structured extraction of ESG information has been assessed within a sample of textile SMEs in the Valencian Community over two years, aiming to provide an initial diagnosis of the quantity of sustainability reported information. The primary conclusion drawn is that companies are progressively incorporating such information into their web platforms, albeit without consistent coverage across all ESG dimensions and sub-dimensions.

Keywords: *Sustainability reporting; Web-based information; SMEs (Small and Medium-sized Enterprises); Sustainability indicator framework; Textile*

1. Introduction

In the past few years, there has been a significant increase in awareness regarding environmental, social and governance (ESG) concerns, leading to a substantial surge in the demand for sustainability data. Unlike larger corporations, Small and Medium Enterprises (SMEs) may not have standardized reporting frameworks for sustainability disclosure. The absence of universal metrics and reporting standards makes it challenging to compare and benchmark sustainability performance across different SMEs (Pranugrahaning et al., 2021, Martins et al., 2022).

Editorial Universitat Politècnica de València

In the lack of a standardized framework for assessing corporate sustainability, many companies, including SMEs, have chosen to report their engagement in sustainable practices through their corporate website (Cruciata et al. 2023; Palma et al. 2022; Wanderley et al., 2008; Lodhia, 2010). This digital footprint, when monitored for sustainability performance, becomes a powerful tool for gaining insights into company behaviors and practices. With recent advancements in Natural Language Processing (NLP) technologies, the ability to extract meaningful information from this digital trail has been greatly improved (Blazquez and Domènech, 2008, Luccioni et al., 2020).

The aim of this paper is to propose a framework of sustainability indicators for SMEs that permits the extraction of web-based information on their environmental, social and good governance practices. The selected indicators have been obtained by adapting the proposal of the Bank of Spain's BELAb project for large companies (Fernández-Rosillo San Isidro et al., 2023), as well as other reference reports focused on the case of SMEs. Once the indicators have been defined, an initial diagnosis will be made of the information disclosed through the web in a sample of textile companies in two periods, 2021 and 2024, to see how the disclosed sustainability information via their websites has evolved. This proposal is an innovative approach to adapt a sustainability indicator framework originally designed for large companies to SMEs. The use of NLP to extract and analyze web-based sustainability information is a creative and practical solution to the problem of non-standardized information in SMEs.

The paper is organized as follows. Section 2 presents the proposed monitoring framework by exploring and identifying relevant sustainability indicators for SMEs. Section 3 explains the methodology used to extract the information about the selected indicators. Next, in Section 4, we present an application of extracting and structuring ESG information in a sample of SME textiles companies in Spain. Finally, the paper ends with conclusions and future line of research.

2. Exploring and determining the relevant indicators for SMEs

Regulation of non-financial reporting differs by country and jurisdiction, but there has been a growing interest around the world in promoting non-financial disclosure. In the European Union, for example, Directive 2014/95/EU on non-financial disclosure (European Union, 2014) and diversity for large companies was implemented, requiring certain companies to report on environmental, social and personnel, human rights, and anti-corruption issues.

Although progress has been made in the disclosure of non-financial information currently, SMEs are not obligated to submit a Non-Financial Information Statement (NFIS). Nevertheless, regulatory frameworks are undergoing changes, and SMEs may be mandated to do so in the future. Our analysis encompassed the list of preliminary indicators proposed by Fernández-Rosillo San Isidro et al., (2023) which has been complemented taking into account several international ESG standards, with a particular focus on delving into the technical documentation of the Global Reporting Initiative (GRI). This focus on GRI was driven by the significant number of SMEs companies that choose to report according to this standard. Table 1 presents an initial compilation of 37 ESG indicators distributed across three main ESG dimensions, each further classified into ten subtypes. The environmental dimension encompasses distinct groups such as Energy, Water, Greenhouse Gases, Waste, and Environmental Policies. Within the social dimension, indicators are structured into three groups: Employees, Diversity, and Society. Lastly, the governance dimension is characterized by a set of specific indicators related to corporate governance and corruption and bribery.

3. Materials and methods

3.1. Methodology

The process followed for information extraction is divided into three stages.

First stage: Definition of keywords associated to each indicator. Following the proposal suggested in Section 2, a keyword dictionary derived from the stem of words for each ESG indicator is built to track the presence of the text on the company's website.

Second stage: Data extraction process with a procedure like the one described in Blázquez et al. (2018) and Crosato et al. (2021). To analyze the disclosure of information about an indicator this stage is designed to answer the question: "Is this text about the label X in the ESG indicator included in the web? The answer to this question is an indicator of the company's awareness about the ESG indicator. Next, the number of instances in which a label appears in association with an indicator in the companies' web pages is transformed into values of "1" if it appears, or "zero" otherwise.

Third stage: A Wilcoxon signed-rank test was employed to evaluate differences between the years 2021 and 2024 in the median frequency of keywords appearing on company websites. This non-parametric approach was chosen because it makes no assumptions about the underlying data distribution. Null Hypothesis (H_0): There is no statistically significant difference in the median frequency of keyword occurrence between the years 2021 and 2024; Alternative Hypothesis (H_1): There is a statistically significant difference in the median frequency of keyword occurrence between the years 2021 and 2024.

Table 1. Selection of ESG indicators for sustainability reporting in SMEs. ***Note.*** Developed based on Corral-Lage et al. (2021) & Fernández-Rosillo San Isidro et al. (2023).

Type	Subtype	Identifier	Indicator
E	Energy	E01_ENCO	Energy consumption
		E02_ROEC	Reduction of energy consumption
		E03_RETE	Percentage of renewable energy relative to total energy consumed
	Water	E04_WACO	Water consumption
	Green House Gases	E05_GHEI	GHG emissions intensity
		E06_GHER	GHG emissions reduction
	Waste	E07_WAGE	Waste generated
		E08_HAWW	Hazardous waste
		E09_MAWA	Managed waste
		E10_REWA	Reused waste
	Environmental Policies	E11_CIEC	Circular economy
		E12_ENPO	Environmental policy
		E13_ISCO	Regulatory compliance
S	Employees	S01_EMTR	Employee training
		S02_DISA	Disability
		S03_JOST	Job stability
		S04_EMTU	Employee turnover
		S05_NUOD	Number of dismissals
		S06_WLBP	Work-life balance policies
		S07_OCRI	Occupational risk
		S08_JOTE	Job seniority
		S09_EMPL	Employees
	Diversity & equality	S10_GEDI	Gender diversity
		S11_AGPG	Average gender pay gap
		S12_EQPL	Equality plan
		S13_DIPL	Diversity plan
	Society	S14_HASP	Health and safety policy
		S15_HURP	Human rights policy
		S16_SUPA	Supplier payments
		S17_SUCH	Supply chain
G	Corporate Governance	G01_AVBR	Average board remuneration
		G02_BOME	Board Meetings
		G03_GDIT	Gender Diversity in the Board
	Corruption and bribery	G04_COAB	Corruption and bribery
		G05_CRPP	Crime prevention policy
		G06_NCAB	Number of corruption and bribery reports
		G07_WHCH	Whistleblower channel

Table 2. Example of labels associates to subtype dimension "Energy".

Type	Subtype	Dictionary of labels (Spanish)
E	Energy	Consumo de energía; GRI 302; Consumo energético; Economía Circular; Uso de agua; Huella hídrica, Residuos; Reciclaje; Energía Renovable, …

3.2. Data

The sample for this study covers 215 textile SMEs located in the Comunidad Valenciana region in Spain, with data for the years 2021 and 2024 considered. The sample of official websites of companies was retrieved from the SABI (Iberian Balance Sheet Analysis System) database after a selection and filtering process of active companies in the Comunidad Valenciana textile sector in both years and with a single URL. As for the sustainability variables they were extracted from the websites of the companies after crawling the complete website in both years.

4. Results

The Wilcoxon test revealed that the p value obtained (0.001) is less than the significance level ($\alpha = 0.05$), leading to the rejection of the null hypothesis. Therefore, it is concluded that there is statistical evidence to affirm that the medians of the frequency of appearance of the keyword between the years 2021 and 2024 are different. This result suggests a significant change in the disclosure of sustainability information on company websites during that period, indicating an increase in concern about sustainability.

Looking at the disclosed information on sustainability in our sample for both years in Table 3, the first conclusion to be drawn is that there has been a 27,72% increase in information in 2024 compared to 2021.

Table 3. Summary of ESG information retrieved by type.

Type	2021	2024	Increase
E	217	283	30,4%
S	306	366	19,6%
G	36	65	80,6%
Total	559	714	27,7%

Figure 1 shows the distribution of disclosed information by type (environmental, social or governance). Notice that, while there is an increase in the percentage of disclosed information about environmental and governance dimensions, there is a small decrease in the percentage of reporting of social dimension.

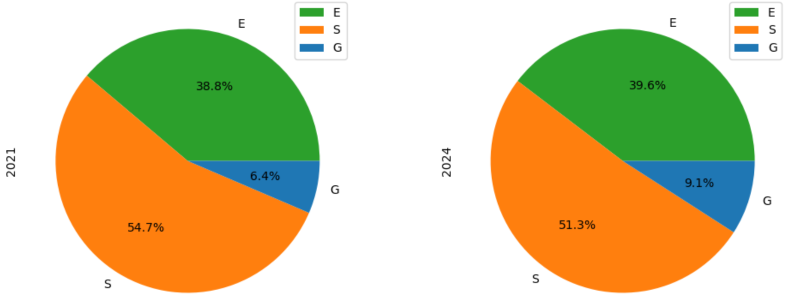

Figure 1. Benchmarking of information presented on the web by type (ESG).

Regarding the subtype of information included in the group of environmental indicators (see Figure 2), there is a greater presence of information contained in the subcategories of energy policies, energy and water. On the other hand, there is almost no information on gas emissions and water consumption. As for the set of social indicators, information on employee welfare, equality and diversity issues is predominant, while website disclosed information on the company's impact on society is less present. Finally, in the group of governance indicators, there is a growing interest in disclosure information related to whistleblowing and corruption channel.

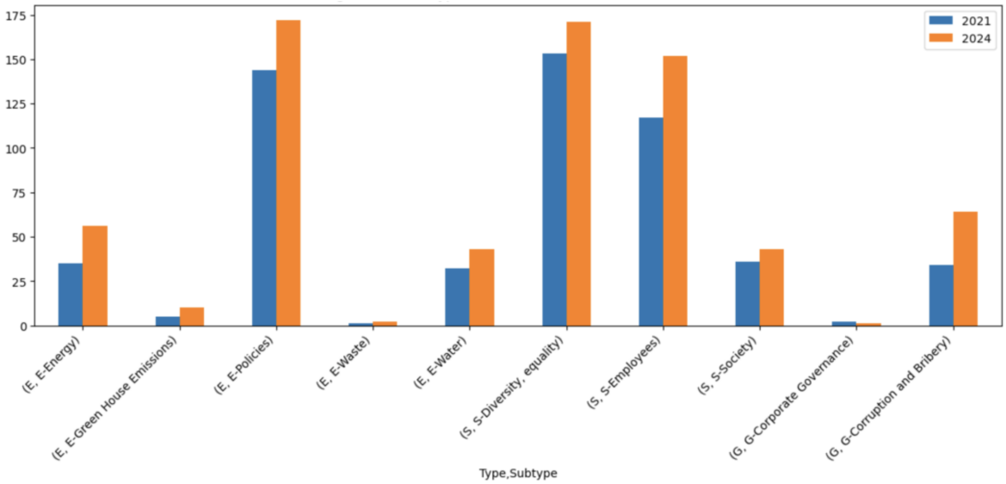

Figure 2. Benchmarking of ESG information presented by subtype.

5. Conclusions

In this paper we have developed a framework for analyzing complementary sustainability information for SMEs based on unconventional data extracted from company's websites. The proposed framework, adapted from the Bank of Spain's BELAb project and other relevant reports, serves as a valuable tool for analyzing sustainability indicators tailored to the unique characteristics of SMEs. The research successfully leverages Natural Language Processing (NLP) technologies to extract meaningful sustainability information from the digital footprint left by companies on their websites.

The distribution of disclosed information by subtype provides a more detailed overview, emphasizing the importance of certain indicators within each dimension. For instance, the paper highlights the prominence of data related to energy policies, energy, and water within environmental indicators, as well as the growing interest in governance indicators related to whistleblowing and corruption channels.

The findings underscore the progress made by textile SMEs in incorporating sustainability information into their digital presence, while also pointing towards areas where improvements and standardization are necessary for a more transparent and comprehensive disclosure landscape. Future research should build upon these insights. Furthermore, this exploration is encouraged to encompass a more extensive sample of companies in other regions and industries, spanning multiple years, to provide a comprehensive understanding of evolving practices. In addition, future research could enhance the process of keyword extraction for ESG indicators by using machine learning techniques.

References

Blazquez, D. and Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99–113.

Corral Lage, J., García Delgado, S., Ipiñazar Petralanda, I., Peña Miguel, N., Saitua Iribar, A. 2021. Guía para la emisión y verificación de información sostenible a través de indicadores medioambientales, sociales y de gobernanza para PYMEs. BNFIX GLOBAL, S.L. pp. 66. Retrieved March 03, 2024, from https://www.bnfix.com/wp-content/uploads/2021/11/GUIA_BNFIX_impreso-1.pdf

Crosato, L., J. Domènech, and C. Liberati (2021). Predicting SME's default: Are their websites informative? *Economics Letters*, 204:109888.

Cruciata, P., Pulizzotto, D., Héroux-Vaillancourt, M., & Beaudry, C. (2023). 0-shot text classification for web-based environmental indicators: Pilot study on B-Corp data. *5th International Conference on Advanced Research Methods and Analytics (CARMA2023)*. http://dx.doi.org/10.4995/CARMA2023.2023.16463

European Union. (2014). Directive as regards disclosure of non-financial and diversity information by certain large undertakings and groups, 2014/95/EU. Retrieved March 03, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014L0095

Fernández-Rosillo San Isidro, B., Koblents Lapteva, E., & Morales Fernández, A. (2023). Micro-database for sustainability (ESG) indicators developed at the Banco de España (2022). *Notas estadísticas/Banco de España, 17*.

Lodhia, S. K. (2010). Research methods for analysing world wide web sustainability communication. *Social and Environmental Accountability Journal*, *30*(1), 26–36.

Luccioni, A., Baylor, E., & Duchene, N. (2020). Analyzing sustainability reports using natural language processing. *arXiv preprint arXiv:2011.08073*.

Martins, A., Branco, M. C., Melo, P. N., & Machado, C. (2022). Sustainability in small and medium-sized enterprises: A systematic literature review and future research agenda. *Sustainability*, 14(11), 6493.

Palma, M., Lourenço, I. C., & Branco, M. C. (2022, October). Web-based sustainability reporting by family companies: the role of the richest European families. In *Accounting Forum* (Vol. 46, No. 4, pp. 344-368). Routledge.

Pranugrahaning, A., Donovan, J. D., Topple, C., and Masli, E. K. (2021). Corporate sustainability assessments: A systematic literature review and conceptual framework. *Journal of Cleaner Production, 295, 126385.*

Wanderley, L. S. O., Lucian, R., Farache, F., and de Sousa Filho, J. M. (2008). CSR information disclosure on the web: a context-based approach analysing the influence of country of origin and industry sector. *Journal of business ethics*, 369–378.