Text-analysis of CSR-reporting

in

Financial sector

Abstract

The aim of this paper is to use text analysis on a sample of CSR-reports concerning the finance sector collected from the GRI-database to provide legislators with information about what dimensions are in focus in finance CSR-reports. To do this a methodology inspired by Pencle & Malaescu (2016) is used.

We find that the word count, dictionary and cluster topic analysis indicate that there is more focus in the finance sector on employee and customer sides of CSR while the topic analysis of the legislators specific phrases (greenhouse gas emissions, diversity, employee health & safety, and customer welfare) indicate that diversity is the most reported phrase followed by greenhouse gas emissions. The reports generally report in a positive sentiment about CSR.

Keywords: Corporate Social Responsibility (CSR), Finance sector, Text analysis, Topic analysis, Sentiment analysis, Greenhouse gas emissions, Diversity, Employee health, Consumer welfare.

1. Introduction

The purpose of CSR-reporting is to provide non-financial information to stakeholders. Stakeholders can be many different parties, such as shareholders, customers, legislators, employees' etcetera.

There have been many attempts to define CSR (Corporate social responsibility) but it is still not clear how the concept should be defined (see for example, Dahlsrud 2008 and Sheehy 2015). Although there is still no clear and unified definition there is according to Dahlsrud many similarities between the 37 definitions examined in his 2008 paper. Dahlsrud concludes that the "…confusion is not so much about how CSR is defined, as about how CSR is socially constructed in a specific context".

Sheehy describes that CSR is not a singular concept but rather a mix of different concepts from various stakeholders with various motives and approaches that supposedly address a multitude of social and environmental concerns that are not so well defined. Sheehy argues that a clear definition would foster transparency by making it possible to challenge those that do not follow the definition but also for companies to deal with complaints of for example greenwashing by showing how they follow the standards.

Others argue that there is no need for a definition (for example Okoye 2010). Okoye argues that CSR is an essentially contested concept and that as it deals with ever changing circumstances it is important that it is not defined since that offers a flexibility to address how companies should handle CSR issues and change over time.

Pencle & Malaescu (2016) offer a way forward in capturing and analyzing the elusive CSR subject using text-analysis. They made a dictionary of CSR words for four different dimensions of CSR (Employee, Environment, Human Rights and Social and Community). In their paper they only present a sample of the words in the dictionary so we cannot directly use that. They have a link to the full dictionary, but it is not currently working (http://www.catscanner.net/dictionaries.php). We can however try to use a very simplified method inspired by theirs (although not as exhaustive due to time constraints) to make our analysis.

1.1 Aim

The aim of this paper is to use text analysis on a sample of CSR-reports concerning the finance sector collected from the GRI-database to provide legislators with information about what dimensions are in focus in finance CSR-reports. To do this a methodology inspired by Pencle & Malaescu (2016) is used.

2. Methods

2.1 Data and data processing

As our sample we have been working with a GRI (Global Reporting Initiative) database that holds circa 19000 CSR or Sustainability reports from various sectors of industry and countries from the years 1999 to 2017. We chose the finance sector since its impact on CSR issues is often less clear than that of for example the manufacturing sector and therefore interesting to look in to in terms of what the sector writes about in its CSR reports.

The database contained circa 2500 pdfs for the finance sector and out of these 1522 were chosen randomly (from all countries, regions and languages). This was done to get a more generalizable result than if we would have worked with only one country, region or language. However, it also opened us up to issues with translation.

The PDFs were categorized and translated using a two-step Python scripting process. Initially, the "langdetect" package was employed for language recognition, enabling the sorting of PDFs into two distinct folders: one for English-language documents and another for documents in languages other than English. Subsequently, a separate Python script utilized the Google Translate API, "googletrans", to systematically translate all non-English PDFs. This iterative translation process was applied across the entire folder containing non-English documents. Upon completion of the translations, the two folders were merged, forming a consolidated and clean database for the project. This streamlined database serves as the foundation for subsequent analyses and data manipulation.

After translation we loaded the files into R and used the lapply function to read text from each PDF file using pdf_text to store it in a list. This did not work, and we had to manually remove 105 pdfs and were then left with 1417 pdfs from which we could read the text using the lapply function. We used R (TermDocumentMatrix) to clean up the data (removing punctuation, numbers, common stop words, disabling stemming and converting to lowercase) and load the terms into a term documents matrix for further analysis.

2.2 Data analysis

Following Pencle & Malaescu (2016) we did both a deductive and an inductive process for the wordlist. We started with the inductive by making R count all words that appeared at least three times.

For the deductive part, unlike Pencle & Malaescu we did not have the time or access to experts to evaluate content validity. Instead, we manually inspected the top 1000 most common results and selected the keywords related to CSR. To this list we then added relevant words according to CSR-literature (Dahlsruds example phrases, the sample list in Pencle & Malaescu (2016) but also Jizi, Salama, Dixon, Stratling (2014) and Kuzey, Uyar, Nizaeva, Karaman (2021)) as well as from our own general knowledge about the subject.

We then carried out topic analysis on word clusters and on the key phrases according to the legislators from the assignment (greenhouse gas emissions, diversity, employee health & safety, and customer welfare). We then continued to make sentiment analysis to see if the reporting was mostly positive or negative.

2.3 Data distribution

We compiled an Excel spreadsheet containing the titles of PDF documents sourced from the finance sector. The Excel sheet includes columns detailing the size, sector, publication year, and Global Reporting Initiative (GRI) type. This compilation serves as the foundation for our material analysis, providing us with an overview of the content structure that we will be analyzing. By systematically organizing and categorizing this information, we aim to gain insights into the key themes and trends prevalent in the finance sector.

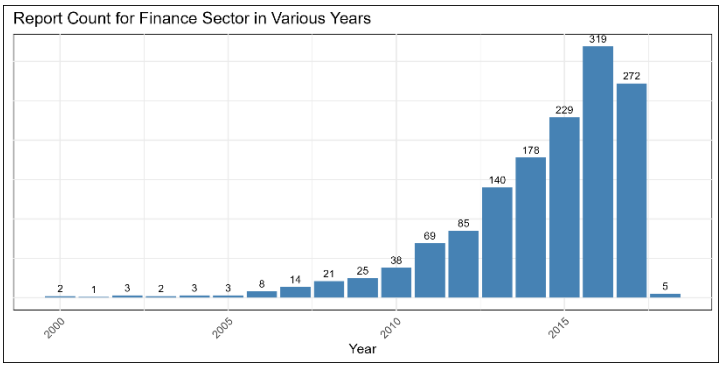

Figure 1. Distribution of reports per year.

The distribution of reports predominantly spans the years 2015-2017.

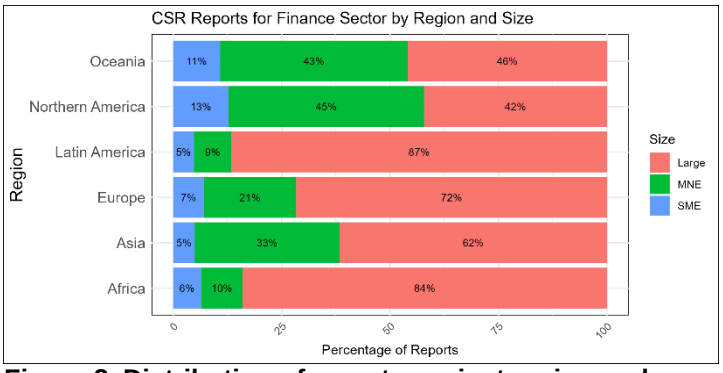

Figure 2. Distribution of reports against region and company size.

The reports primarily emanate from large companies within the finance sector, with a significant representation hailing from the Asia and Europe regions.

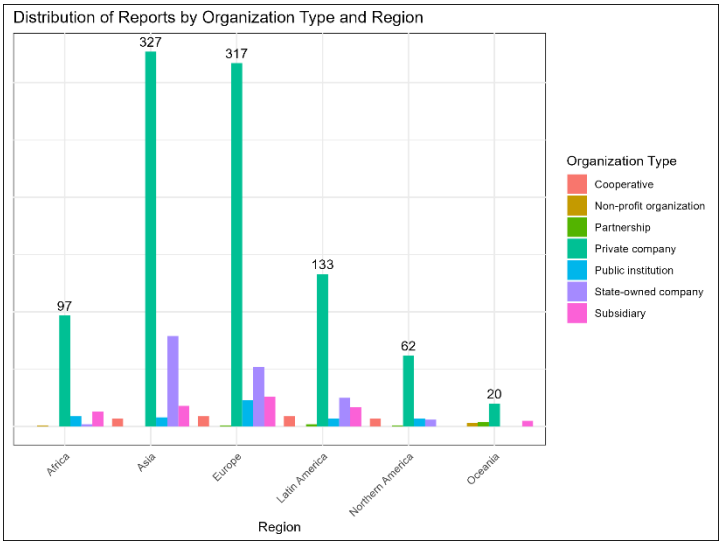

Figure 3. Distribution of Reports by Organization Type and Region.

In all regions of our sample, the most common organization type is private company.

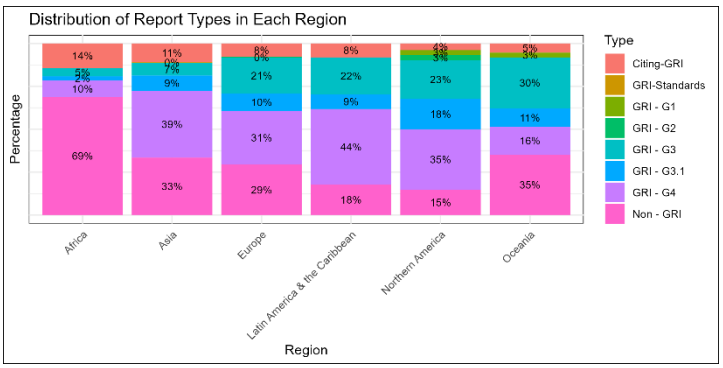

Figure 4. Distribution of GRI Reporting Standards for each region.

The analysis showed that the majority of reports from the Africa region do not follow the GRI Standards for sustainability reporting.

3. Results and discussion

3.1 Inductive wordlist

In the top ten words with the highest count in our dataset it is noteworthy that none of the words are what you would describe as CSR-words. But rather they are all typical finance words such as financial and bank (the top two results) or general words such as year and total. Most words in this list are therefore not so useful for our analysis.

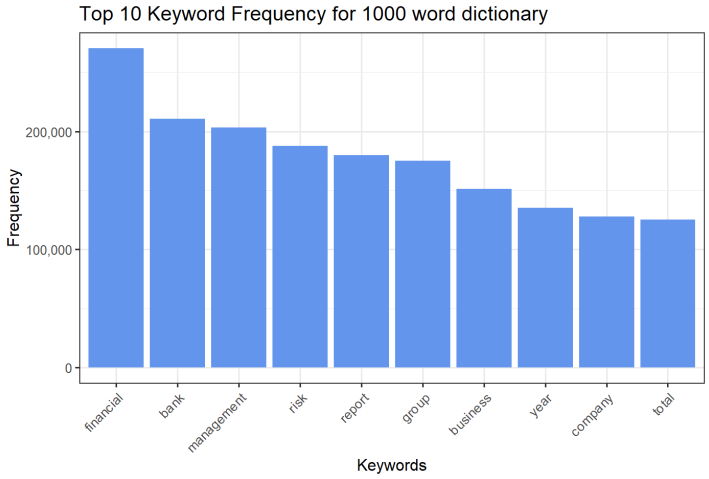

Figure 5. Top ten keywords in inductive dictionary

Figure 6. CSR Reports Word cloud.

Figure 6 illustrates the most frequently utilized terms across our entire dataset.

3.2 Deductive wordlist

The inductive list of the 1000 most common words was then checked for CSR relatability and non-CSR words were removed and CSR-keywords were added to the search. This resulted in a list of 95 words. The top ten words:

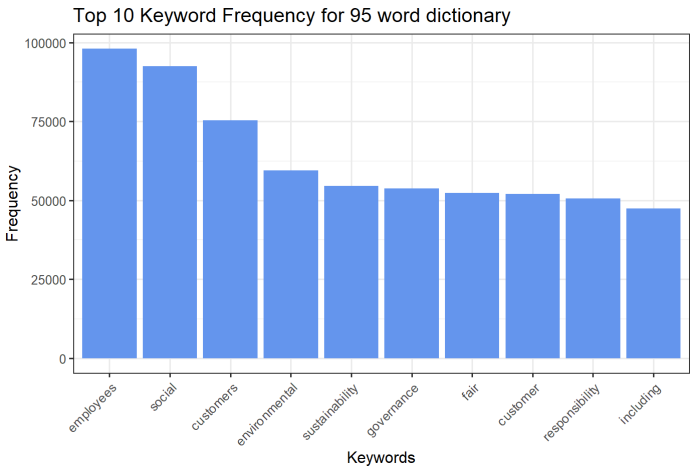

Figure 7. Top ten keywords in 95 word deductive dictionary.

This indicates that in the finance sector CSR, is first of all associated with employees and social indicating that these are important words in financial sector CSR-reporting. However, adding up customers and customer will result in a count of more than 120000 words, hinting at some of the problem with word validity and synonyms.

The 95 words were then checked manually for synonyms and added up resulting in a list of 79 words. The top ten being:

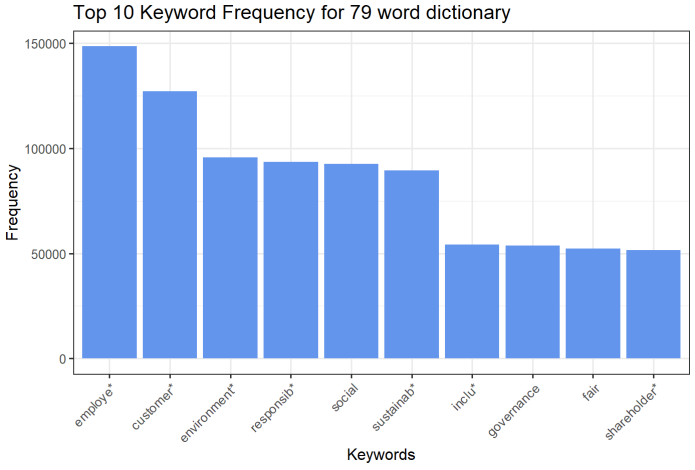

Figure 8. Top ten keywords in 79 word deductive dictionary.

This change puts employe* (the sum of employees 97981 + employee 38424 + employment 12262) on top. But customer*, environment* and responsib* have all passed social, putting social at 5th instead of 2nd place. The top two results have over 125000 counts. Indicating that for CSR reporting in the financial sector issues relating to employees and to customers are more prevalent. Emissions for example doesn't appear until place 27 in the list (and gas on place 54) indicating that greenhouse gas emissions is not as prevalent as other topics and other dimensions in the CSR-reporting of companies in the financial sector.

3.3 Topic modelling analysis

Many of the words in CSR reporting can fit into different categories or are general categories such as CSR and ESG. However with inspiration from how Dahlsruds and Pencle & Malaescu (2016) clustered their words and phrases we tried to put the 79 dictionary words into the four groups greenhouse gas emissions (greenhouse), diversity, employee health & safety (employee), and customer welfare (customer) all words that did not fit on category went in the general category. We did the analysis and got the following result:

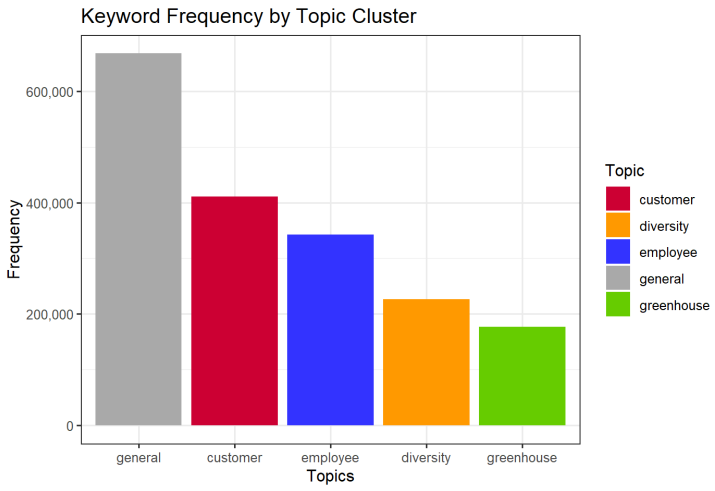

Figure 9. Keyword Frequency by Topic Cluster

We then proceeded to test how the unique phrases themselves "greenhouse gas emissions", "diversity",

"employee health & safety", and "customer welfare" to see how they were distributed in the data.

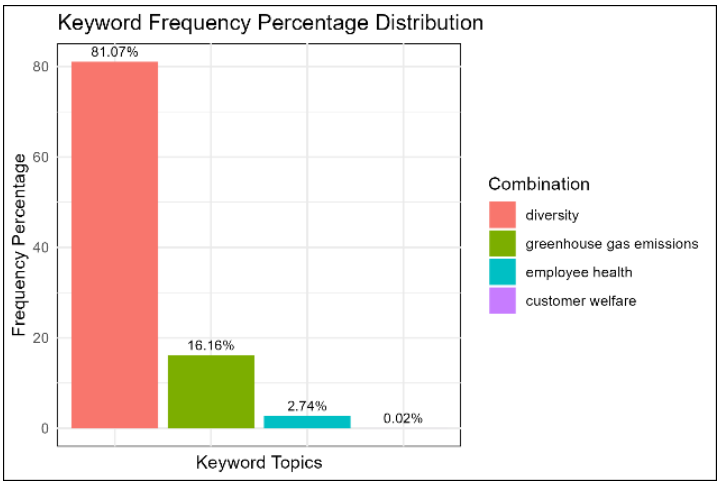

Figure 10. Keyword count distribution by topic.

This analysis showed that diversity is by far the most common of the four phrases (81,07 %). Trailing quite far behind as number two is the phrase greenhouse gas emissions (16,16 %). Employee health and customer welfare barely register with (2,74 % and 0,02 %).

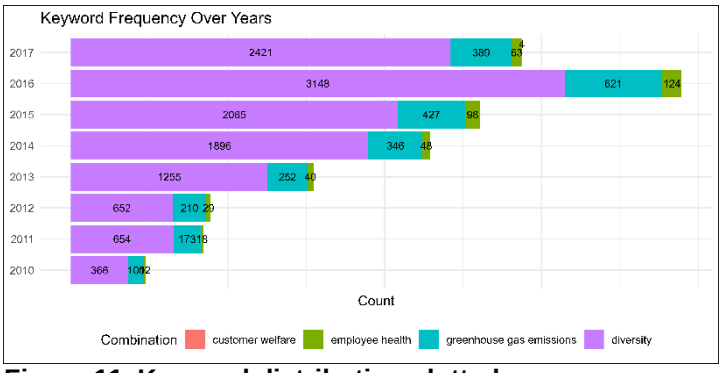

Figure 11. Keyword distribution plotted over years (2010-2017)

The analysis reveals that the frequency of the keyword "diversity" peaked between 2014 and 2016, gradually declining in 2017. Interestingly, there was no mention of the CSR issue "customer welfare" until 2016, but its occurrence started to increase gradually in 2017, indicating a growing awareness and recognition of the importance of addressing customer-centric issues.

We then wanted to see how the words were connected and proceeded to make a bigram graph.

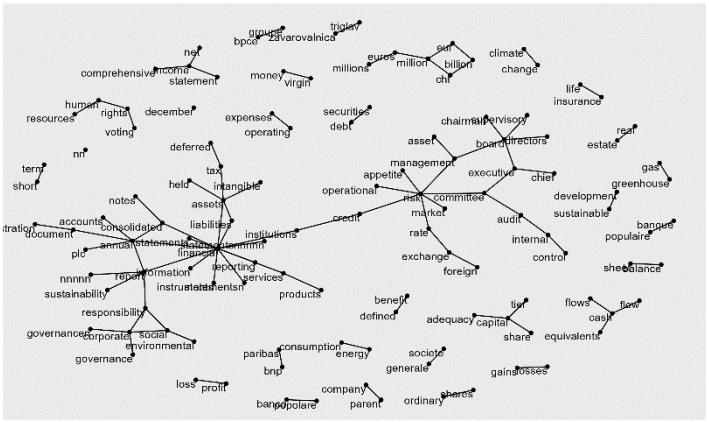

Figure 12. Co-occurrence network analysis

We see for example that greenhouse is related to gas and that human, rights, voting and resources form a network as does responsibility, governance, corporate, social and environmental.

3.4 Sentiment analysis

The last step of our analysis was to test for sentiment in our data.

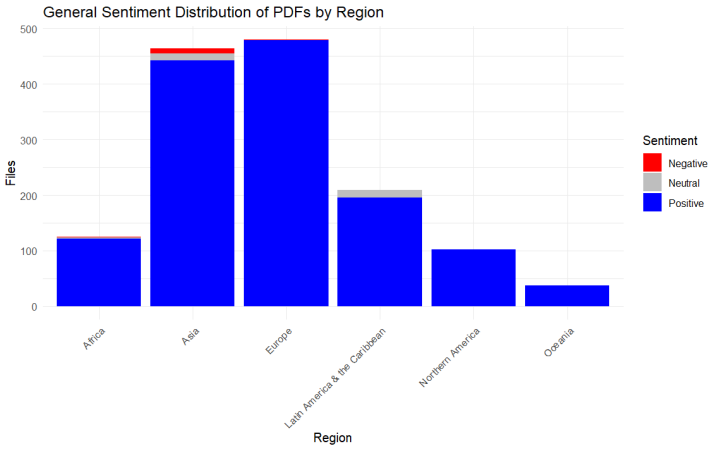

Figure 13. General sentiment distribution of PDFs by region

The results show a very positive sentiment in the CSR reporting. Even with a value of minimum 0.5 for positive sentiment and value below -0.05 for negative sentiment the results were still overwhelmingly positive. The remaining pdfs were classified as neutral.

4. Conclusions

The word count, dictionary and cluster topic analysis indicate that there is more focus in the finance sector on employee and customer while the topic analysis of the specific phrases indicate that "diversity" is the most reported phrase followed by "greenhouse gas

emissions". This contradiction in the results could be a result of many things. For example, diversity itself being a word could make it more common than a whole phrase since single words are usually more common than whole phrases. It could also be that the phrases employee health and customer welfare do not appear so frequently but that those topics are instead discussed using other terms.

The co-occurrence analysis gives some support to the clustering of terms used in the deductive topic creation. It also indicates that some words are connected, for example that greenhouse is related to gas and that human, rights, voting and resources form a network as does responsibility, governance, corporate, social and environmental. This could suggest that topics are discussed but perhaps not using the phrases we used in our phrase based topic analysis.

The sentiment analysis of the CSR reports is very positive indicating that in the finance sector reporting sentiment on CSR issues is very positive.

The observed CSR priorities within the financial sector may, in part, stem from the translation process of the PDF documents. Throughout this analysis, it becomes evident that words with dual meanings or nuanced translations may not always convey the intended message accurately in English. There are also issues with subjectivity and validity when doing deductive analysis. Words can be misinterpreted or overlooked. Therefore the dictionaries we created can serve as inspiration for future research but due to their limitations they should not be copied directly. The main contribution of this paper is that it gives a general overview of what dimensions are in focus in finance sector CSR-reports from the GRI-database.

To sum up some information to the legislator: We find that the word count, dictionary and cluster topic analysis indicate that there is more focus in the finance sector on employee and customer sides of CSR while the topic analysis of the legislators specific phrases (greenhouse gas emissions, diversity, employee health & safety, and customer welfare) indicate that diversity is the most reported phrase followed by greenhouse gas emissions. The reports generally report in a positive sentiment about CSR.

Reference list

Dahlsrud, A. (2008), How corporate social responsibility is defined: an analysis of 37 definitions. Corp. Soc. Responsib. Environ. Mgmt, 15: 1-13. https://doi.org/10.1002/csr.132

Sheehy, B. (2015), Defining CSR: Problems and Solutions. Journal of business ethics, 10/2015, Volume 131, Issue 3

Okoye, A. (2010). Theorising corporate social responsibility as an essentially contested concept: Is a definition necessary? Journal of Business Ethics, 89(4), 613–627. doi:10.1007/s10551-008 0021-9.

Pencle, N. and Mălăescu, I. (2016) What's in the Words? Development and Validation of a Multidimensional Dictionary for CSR and Application Using Prospectuses. Journal of emerging technologies in accounting, 09/2016, Volume 13, Issue 2

Jizi, M. I., Salama, A., Dixon, R., & Stratling, R. (2014). Corporate Governance and Corporate Social Responsibility Disclosure: Evidence from the US Banking Sector: JBE. Journal of Business Ethics, 125(4), 601-615. https://doi.org/10.1007/s10551-013-1929-2

Kuzey, C., Uyar, A., Nizaeva, M., Karaman, A., (2021) CSR performance and firm performance in the tourism, healthcare, and financial sectors: Do metrics and CSR committees matter?, Journal of Cleaner Production, Volume 319, https://doi.org/10.1016/j.jclepro.2021.128802.