Proceedings of CSR-NLP I @LREC 2022, pages 1–8
Marseille, 25 June 2022

An NLP Approach for the Analysis of Global Reporting Initiative Indexes from Corporate Sustainability Reports

Marco Polignano §, Nicola Bellantuono °, Francesco Paolo Lagrasta *, Sergio Caputo §, Pierpaolo Pontrandolfo *, Giovanni Semeraro §
§University of Bari Aldo Moro, °University of Foggia *Polytechnic of Bari
§Dept. Computer Science *Dept. Mechanics, Matemathics and Management, Via E. Orabona 4, 70125, Bari
° Dept. of Agriculture, Food, Natural Resources and Engineering, Via Napoli 25, 71122, Foggia
{marco.polignano, giovanni.semeraro}@uniba.it
nicola.bellantuono@unifg.it
s.caputo34@studenti.uniba.it
{francescopaolo.lagrasta, pierpaolo.pontrandolfo}@poliba.it

Abstract

Sustainability reporting has become an annual requirement in many countries and for certain types of companies. Sustainability reports inform stakeholders about companies' commitment to sustainable development and their economic, social, and environmental sustainability practices. However, the fact that norms and standards allow a certain discretion to be adopted by drafting organizations makes such reports hardly comparable in terms of layout, disclosures, key performance indicators (KPIs), and so on. In this work, we present a system based on natural language processing and information extraction techniques to retrieve relevant information from sustainability reports, compliant with the Global Reporting Initiative Standards, written in Italian and English language. Specifically, the system is able to identify references to the various sustainability topics discussed by the reports: on which page of the document those references have been found, the context of each reference, and if it is mentioned positively or negatively. The output of the system has been then evaluated against a ground truth obtained through a manual annotation process on 134 reports. Experimental outcomes highlight the affordability of the approach for improving sustainability disclosures, accessibility, and transparency, thus empowering stakeholders to conduct further analysis and considerations.

Keywords: Natural Language Processing, Information Extraction, Sustainability Reporting, Global Reporting Initiative, Corporate Analysis

1. Introduction

The EU Corporate Sustainability Reporting Directive (CSRD), proposed on April 21, 2021[1], would significantly extend the scope of sustainability reporting legislation among European companies. With a stated aim of bringing sustainability reporting on a par with financial reporting, it would help to have equal weight and rigor. The currently in force Non-Financial Reporting Directive applies to some large companies operating in the EU. CSRD would extend the reporting obligation to all large companies, either listed or unlisted, as well as to all listed firms, with the sole exception of listed micro-companies. The reporting obligation would also be extended to all groups, which will have to produce a consolidated sustainability report. Estimates predict that the application of these new inclusion criteria will bring the number of sustainability reporting obliged companies from the current 11,700 to about 49,000.

The ways in which companies approach sustainability reporting are often varied and non-standard. While society and governments demand sustainable development, the efforts deployed by companies are often not adequate. Only recently, given the numerous initiatives towards environmental and social respect, some of the largest and best-known companies have decided to accept the request of national and supranational governments for more adequate reporting on these issues (Bowen, 2014). Nonetheless, economics still play a pivotal role for environmental decisions, and according to (Dyllick and Muff, 2016), companies' understanding of sustainability has been misguided resulting in most companies committed to reducing unsustainability rather than actually pursuing sustainability. Identifying relevant sustainability topics and disclosing related information seems then a quite challenging task even for responsible companies, resulting in lower communication efficacy and in turn accessibility by stakeholders (e.g. consumers, authorities). The identified issues to some extent depend on the difficulties in monitoring activities experienced by the competent bodies: sustainability reports are often complex, with customized layouts, long, and challenging to read, which make their analysis time consuming and costly. We propose to address such issues by means of the support of computer systems. We developed an approach based on Natural Language Processing (NLP) and Information Retrieval (IR) to support the review process of such documents. Our system is capable of analyzing documents in closed format, i.e., PDF, and

[1] https://ec.europa.eu/info/business-economy-euro/companyreportingandauditing/company-reporting/corporatesustainabilityreporting_en

extracting information potentially valuable for the review phase. In fact, referring to the Global Reporting Initiative (GRI) Standards[2], we searched for sustainability topics in the textual document in order to identify the context of use and the page where they are discussed. Our approach speeds up the operations of analysis, study, and review of corporate documents on sustainability.

1.1. Research Goals

In this work, we aim to address the issue of automatic analysis of textual documents concerning sustainability. In particular, we want to investigate the possibility of adopting NLP and IR techniques to be able to automatically extract relevant information for possible consultation and review by stakeholders. Specifically, it was considered that the preliminary analysis that could be done is to check whether specific sustainability topics or disclosures are discussed in the document. In this work, we focus on sustainability reports compliant with GRI Standards as the latter are by far the most widely adopted. This task, which might seem simple, is instead made complex by the heterogeneity of layouts and the writing style of sustainability documents. We believe that an automated system capable of detecting which topics are actually discussed within the sustainability reports could be a valuable aid for stakeholders as well as anyone involved in the process of reviewing corporate documents. The main contributions of this work are:

- An NLP and IR strategy for the automatic analysis of corporate sustainability reports;
- A system to automatically analyze GRI Standards compliant reports;
- An evaluation using real reports that focuses on GRI topics/disclosures and on the analysis of the extracted contexts.

2. Related Work

Describing how an organization deals with its economic, environmental, and social impacts is an articulated process, called sustainability reporting, whose deliverable is nowadays usually defined as sustainability report. Although the early attempts to describe social activities of companies date back to the 1970s whereas companies involved in environmentally sensitive industries began to publish their environmental reports in the subsequent decade, only in the mid-1990s the first periodic reports of activities, encompassing the three sustainability dimensions in a holistic perspective, were published. These reports, which at the beginning were almost entirely limited to bigger companies, timidly spread also among SMEs, institutions and no profit organizations (Hsu et al., 2013).

[2]https://www.globalreporting.org/standards/

More recently, the rapid increase of awareness on the responsibility that businesses play to achieve sustainable development shed a new light on the practice of sustainability reporting (Minutiello and Tettamanzi, 2022). Sustainability reports, indeed, become crucial for companies not only to communicate to stakeholders how they deal with sustainability, but also to reflect on their sustainable strategies and embrace them committedly. Reporting, in fact, "is not only a matter of communication nor a mere data gathering or compliance exercise. It helps organizations to set goals, measure performance, and manage change" (Global Reporting Initiative, 2013). Recent norms and initiatives on the compulsory release of non-financial disclosures have put a new emphasis on studies of sustainability reporting. Scholars, in particular, have investigated the plethora of schemes proposed to help report sustainability (Grewal and Serafeim, 2020), which often let companies discretionarily choose the shape to adopt for their reports, the content included and the way to identify and present it, at the price of making reports hardly comparable, so weakening the general audience's ability to retrieve information they need. Even when companies strictly follow the Global Reporting Intiative standards, which act as a de facto standard for sustainability reporting, they release heterogeneous kinds of documents, which hinder the identification of resemblances and dissimilarities among companies' sustainability strategies, priorities, and results.

Therefore sustainability reports are configured as complex sources, adopting non-standard layouts and different methods of information representation (e.g. text, tables, infographics), in which both unstructured and semi-structured data can be traced. This feature, combined with the amount of documentation produced annually by companies from all over the world, makes the investigation of sustainability reports through NLP and Text Mining (TM) techniques an interesting and potentially vastly impacting scientific challenge (Zhou et al., 2021). As a matter of fact, NLP techniques allow to analyze and automatically represent texts written in human languages to obtain a human-like language processing useful for subsequent analysis (Liddy, 2001). TM provides instead automatic processes aimed at extracting implicit knowledge from textual data (Jo, 2019), enabling inferences otherwise impractical for the human reader.

Companies non-financial disclosures have been investigated using NLP and TM techniques starting from 2009, when Bayesian analysis and TM were used to measure report content differences among multiple sectors (Modapothala and Issac, 2009). Since then, such techniques have been adopted to investigate different aspects related to non-financial documentation. In (Aureli et al., 2016) TM was used to assess changes in the quantity of sustainability disclosures delivered by companies before and after an industrial disaster. Changes over time were also tracked in (Székely

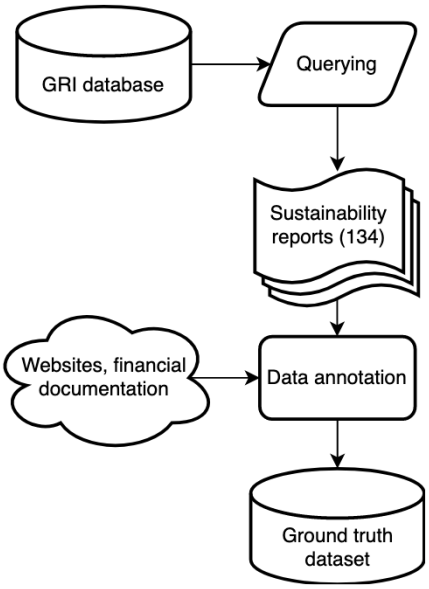

Figure 1: Ground truth dataset creation process.

and vom Brocke, 2017) where authors performed latent Dirichlet allocation (LDA, an NLP technique for topic modeling) to record the evolution over the span of 16 years (1999 - 2015) of the topics contained within sustainability reports. (Lindgren et al., 2021) adopted Bayesian machine learning and LDA, arguing that shareholders appear to be the implicit target users of sustainability reports. (Zhou et al., 2021) used LDA to explore container shipping companies sustainability disclosures. It is not an isolated case: many of the studies focusing on the application of NLP and TM on sustainability reports are devoted to specific industries in order to identify sector-dependent characteristics, trends and best practices (e.g. (Wang et al., 2020), (Uyar et al., 2021)).

This brief overview of the literature contributes to demonstrate how the use of these techniques could prolifically support the investigation of sustainability reports. Despite the research stream boasting a history of more than ten years, to the authors' knowledge, no study was aimed at developing tools intended for the sustainability disclosures (e.g., GRI Standards) information retrieval.

3. Resources and Annotation Process

The dataset used as ground truth to test the effectiveness of the developed tool contains data extracted from 134 sustainability reports, retrieved from the online GRI Sustainability Disclosure Database[3] in March 2021. Fig. 1 depicts the dataset creation process.

[3]Unfortunately GRI decommissioned the Sustainability Disclosure Database, which is no longer accessible. More information is available at https://www.globalreporting.org/how-to-use-the-gri-standards/register-your-report/

The reports were selected through a query that returned GRI Standards compliant documents published by Italian companies: for each of them, the latest available report was included in the dataset. The query design ensures homogeneity in terms of framework (GRI Standards) and national culture (Italian). On the other hand, no constraints were imposed on the size and sectors of the drafting companies: hence, the dataset contains reports ascribable to 27 different sectors (e.g. waste management, automotive, agriculture) published by micro (2/134), small (2/134), medium (4/134) and large (126/134) organizations. The unbalanced representation with respect to the drafting organizations size is due to the previously mentioned European norms, which state that non-financial disclosures are mandatory for large companies.

After report selection, the dataset was populated. Each report underwent a manual annotation phase aimed at structuring organizational, financial and sustainability-related data. The annotation process involved two researchers: the first populated the dataset while the second performed spot checks on the correctness of the information entered. The wide scope of the selected features allows the dataset to be leveraged as a knowledge base for testing new hypotheses and/or developing tools with a potentially different purpose than that of this work. The following organizational and financial variables related to the drafting companies were selected: company name, sector, number of employees, size, annual turnover and annual balance sheet total. As these data were not always contained within the reports, they were collected from different data sources, such as drafting companies websites and miscellaneous financial documentation. In order to collect data pertaining to sustainability-related aspects, we manually scanned the documents in full, extracting title, year, language, reporting option (GRI Standards allow two options - core or comprehensive- to be chosen), involved stakeholders, stakeholder engagement strategy and GRI disclosures. GRI disclosures were recorded through a dummy variables approach: each disclosure-related feature was valued 1 (yes) if it had been included in the report and 0 (no) otherwise. The resulting dataset contains 134 reports published by as many Italian companies. Each report is described by 150 features, 108 of whom are related to GRI sustainability disclosures.

4. The Natural Language Processing Pipeline

Most of the corporate sustainability reports are organized into five main sections, including management information, environment, and climate change, environmental performance review, a listing of verifiable environmental claims and green initiatives, and declarations about environmental compliance. In addition, information about the internal organization's structure, the departments responsible, and employees' roles in

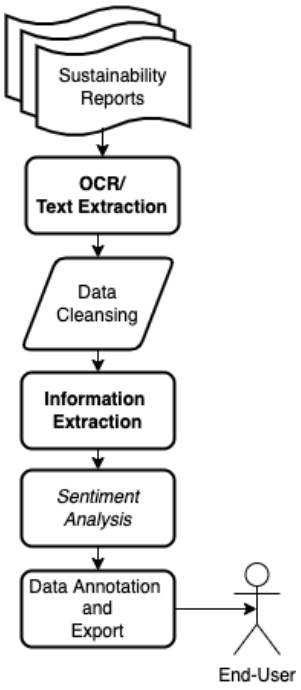

Figure 2: NLP Pipeline for processing the Sustainability Reports.

sustainability activities are provided in most reports. However, companies are not forced to use a predefined standard in the layout design of reports: different design choices are taken for characters style, sections design, and disposition, graphs, images, number of pages, etc.

In order to correctly analyze the documents of our dataset, we decided to implement a complete NLP pipeline (Fig. 2). The first step is the conversion of sustainability reports into a computable textual format. Considering the format of files in the dataset, we had to convert each PDF file into a set of images that could be processed by an OCR (Optical Character Recognition) tool to extract a textual representation of the pages a report consists of. The extracted text is consequently the starting point to perform a focused search on GRI disclosures presence. Before starting the process, we had to make a suitable input for the OCR tool. For this purpose, we used Poppler[4], a library for rendering PDF files and examining or modifying their structure. It was used to convert PDF reports into a collection of images, in which each image corresponded to a page in the report. Then, we adopted Tesseract (Smith, 2007), an OCR engine able to convert the text contained into an image obtained with scans, pictures, or photos into understandable characters for a word processor. The results are usually excellent as far as character recognition is concerned. Conversely, it lacks the ability to maintain page layout, particularly when tables or columns occur in the document. Initially limited to ASCII characters, since March 2022 Tesseract supports UTF-8 characters and recognizes more than 100 languages[5]. During this transformation step, in many cases we observed that the resulting text was not accurate enough to be considered a robust mapping of the content discernible by a human observer. This issue is mainly caused by the need to transform the document pages into images before applying the OCR tools. For this reason, we decided to enrich the textual representation we obtained with text extracted directly from the PDF format. For this purpose, PDFplumber[6] was exploited. It is a Python open-source package whose objective is directed to parsing PDFs, analyzing PDF layouts and object positioning, and extracting text. For our case, we used this library to extract text directly from PDF.

Intending to improve the quality of the text obtained using the two approaches, we decided to apply classical text cleaning techniques. In particular, we decided to remove stopwords as well as to replace carriage returns, tab characters, double or triple spaces with single spaces. In addition, we decided to remove every non-alphanumeric character and/or character of length less than or equal to two, including punctuation, except for the "-" character because it is used within the notation of GRI disclosures. We performed this task using the Spacy Library (Srinivasa-Desikan, 2018). In particular, it provides a complex text analysis pipeline for several languages, including Italian. It offers the possibility of dividing the text into tokens and checking if each of them is a stopword, its length, and the type of content.

Given the textual representation of the reports, we were able to perform a keyword search based on common GRI Standards' disclosures names in order to estimate the presence and absence of them within the document. The conversion of each PDF page into searchable textual content also allowed us to detect the portion of text where each of the GRI standard disclosure was located in the analyzed document. In particular, we can extract the page number and the paragraph containing the GRI Standards' disclosures names. We searched for 215 total keywords created by using the following two possible structures: "GRI <code>", "<code>". The element <code> is an integer number referring to GRI topics between 200 and 400 or their disclosures like 200-x, 300-x, and 400-x. Keywords like "GRI 300-4" or "GRI 203" or "306-4" are examples of of used search terms. Limitations of this approach include the possibility of identifying any numerical values in sustainability reports as GRI topics/disclosures. To overcome that limit, we decided to evaluate a second version of our search pipeline which makes use only of keywords obtained by using GRI disclosures, i.e., those containing the "-" symbol. This makes the pos-

[4]https://pypi.org/project/python-poppler/

[5]https://github.com/tesseract-ocr/tesseract

[6]https://github.com/jsvine/pdfplumber

sibility of false positives very unlikely. In that second approach, we considered a match for a GRI Standard if at least one of its disclosures was identified in the text. In both cases, we considered as the reference context of the specific GRI standard, the portion of the text that contains it, i.e., the one obtained by extracting the 25 terms before and after the match. As an output of this step of the proposed pipeline, we are able to obtain the possible match for each of the possible GRI topics, the reference context, and the page of the document where the match has been identified. This output can be exported for later use by the end-user or optionally processed through a sentiment analysis tool.

Sustainability reports should discuss aspects relevant to the drafting organization and to its stakeholders, highlighting both positive and negative outcomes. Notwithstanding this, companies might tend to report only positive information, neglecting to inform stakeholders about negative performances (Boiral, 2013). In order to verify if this misuse is commonly applied, we can conduct an analysis that takes into account also the sentiment of the context where the GRI disclosures were found. We expect inhomogeneous sentiments since, in principle, each company should make available information, whether positive or negative, especially in the sustainability context. In the literature many approaches for Sentiment Analysis have been proposed (Polignano et al., 2017b; Polignano et al., 2017a; Polignano et al., 2019). In particular, in this work, we decide to use two tools: TextBlob[7] and Sent-It (Basile and Novielli, 2014). TextBlob is a Python library for processing textual data with common NLP tasks. It was adopted to recognize the sentiment for contexts written in English. Sent-It is a sentiment analysis tool that identifies the sentiment for Italian texts. It is a system based on a supervised machine learning approach. In particular, for training, three different kinds of features based on keywords and microblogging properties of tweets, on their representation in a distributional semantic model, and on a sentiment lexicon have been exploited. Data provided for training are annotated according to the subjectivity/objectivity of the content. Moreover, each piece of text is categorized as positive, negative, or neutral. In our case, most of the disclosure's contexts were written in Italian, and we are able to obtain a score of polarity (i.e., positive, negative) and subjectivity/objectivity.
At the end of the analysis process, it is possible to export the results of each document in JSON format. It shows the reference context (if any), page number, polarity score, and subjectivity score for each GRI disclosure. The proposed system has been coded in the Python language and run on the Google Colab Environment[8]. The source code has been released through the GitHub platform[9].

	OCR	Text Extr.	GRI	Sub GRI	Sent.
OCR -all-GRI	✓		✓		
OCR -sub-GRI	✓			✓	
TE -all-GRI		✓	✓		
TE -sub-GRI		✓		✓	
OCR-TE -all-GRI	✓	✓	✓		
OCR-TE -sub-GRI	✓	✓		✓	
OCR-TE-SA -all-GRI	✓	✓	✓		✓
OCR-TE-SA -sub-GRI	✓	✓		✓	✓

Table 1: Configurations of the system we evaluated.

5. Evaluation

The evaluation phase aims to assess the effectiveness and robustness of the proposed system. In particular, we want to investigate the following research questions:

- **RQ1:** Is it possible to develop a robust and effective system for automatically search GRI topics from corporate sustainability reports?
- **RQ2:** How system performances are influenced by the granularity chosen for the keywords used while searching for GRI standards?
- **RQ3:** How the system performances are influenced by the tool used for the text extraction from PDF files?
- **RQ4:** Is it possible to use a sentiment analysis tool for evaluating if sustainability reports are discussing only positive aspects?

With the goal of answering the research questions posed, we ran our system using different configurations. In particular, following the configurations reported in Tab. 1, we used OCR, Text Extractor, or both as document processing tools and all possible or disclosure keywords for the search phase. Finally, we evaluated two configurations based on the use of the sentiment analysis tool, for which we considered a match for the search phase, only GRI topics with a neutral or positive context.
The results obtained from the experimental runs are shown in Tab. 2. It is possible to observe that the results obtained in terms of the F1 measure are promising for all the configurations discussed. In particular, it

[7] https://textblob.readthedocs.io/en/dev/index.html
[8] https://colab.research.google.com/
[9] https://github.com/marcopoli/GRI-Sustainability-Reports-Analysis

	Precision	Recall	F1
OCR -all-GRI	*0.95579*	0.80189	0.87210
OCR -sub-GRI	0.95103	0.89208	0.92061
TE -all-GRI	0.95752	0.80704	0.87586
TE -sub-GRI	0.95228	0.89616	0.92337
OCR-TE -all-GRI	0.95577	0.84152	0.89501
OCR-TE -sub-GRI	0.95014	*0.93725*	*0.94365*
OCR-TE-SA -all-GRI	0.95642	0.79034	0.86548
OCR-TE-SA -sub-GRI	0.95106	0.87259	0.91014

Table 2: Results obtained from the evaluation runs.

varies from the lowest value of 0.87210 obtained from the OCR-all-GRI configuration to 0.94365 found by performing the OCR-TE-sub-GRI configuration. What has been observed shows that in its simplicity, the analysis pipeline presented is highly effective, allowing to obtain results that can represent an excellent base of departure for end users. It is, in fact, clear that a value of F1 measure so close to the value 1 is an indication of the effectiveness of the discovery process and the reliability of the results obtained. This allows us to answer the **RQ1** positively.

Observing the fine-grained results we obtained, in some cases, lower Recall values have been obtained with respect to the average. This issue was caused by the absence of some disclosures we used in the search phase. Indeed using the keywords about the first level of the GRI standards caused many situations of mismatching where the manual annotator has considered the first level of the GRI standard found only because one of its disclosures has been found. Similarly, we performed some configurations by using only keywords obtained from GRI disclosures and considered a match for the first level of the GRI Standard if at least one of its disclosures was identified in the text. Configurations that contain the string "-sub-GRI" in the name are those that follow this approach. What can be observed is that in all cases, these configurations behave better than their "-all-GRI" counterparts. There is, in fact, an increase in performance that varies from 5.16% to 5.56%. The results obtained allow us to provide a clear answer to **RQ2**.

To avoid as many errors as possible due to the incorrect encoding of text resulting from PDF conversion operations, configurations using different text extraction techniques were performed. The results obtained show that using a text extractor succeeds in reducing some of the problems of OCR, particularly those in which the text was shown on colored backgrounds or in fonts that are difficult to interpret. On the contrary, the text contained in images is ignored. Indeed, we moved from an F1 measure value of 0.92061 for the OCR-based technique to 0.92337 for the one based on Text Extractor. Instead, the best performances are achieved when combining approaches. A GRI standard is considered identified if it is found in the text obtained by at least one of the two approaches. This process allowed us to obtain an F1 score of 0.94365, the highest among the results of our runs. These considerations allow us to provide a response to **RQ3**.

Putting our attention to the last two configurations posted in Tab.2, OCR-TE-SA-all-GRI and OCR-TE-SA-sub-GRI, we can observe that the performance of the proposed pipeline decreased comparing them with their counterpart without sentiment analysis applied. This would suggest that in the reports, there are also contexts in which there is a negative concept expressed about a GRI standard. Unfortunately, however, following a detailed analysis, it has been observed that the negative contexts identified are false positives. In fact, they are generated by the misclassification of the same by the sentiment analysis tool used. The presence of certain negative words could definitely affect the sentiment processed using the Sent-It tool. We detect that words such as "rischi", "corruzione", "malattia", "pericoloso" could heavily influence the final sentiment prediction especially if these were found in sentences that uses negations. The results in Tab. 3, show us the most common terms in case of misclassification. These represent words in the Italian language that express, if taken individually, a negative sentiment. Conversely, their use in a negated form or with an outlined context can lead to overall positive sentiment. This shows that corporate sustainability reports tend to present only the positive goals achieved. What was observed appears to be a valid response to **RQ4**.

6. Implications of Research, Limits and Challenges

The proposed approach to analyze sustainability reports is a first step towards the possibility of offering complete and reliable support to the interested stakeholders. In fact, it is common to observe documents that use heterogeneous layouts and different writing styles, sometimes even when adopting the same reporting standard/guideline. The analysis of such documents becomes a demanding, long and tedious task. Therefore, a computer system can be an essential support for analysis operations, as long as it is reliable and effective. The approach we propose is based on a straightforward methodology. The obtained results, though limited to a single PDF format, GRI Standards, and documents written in Italian or English, prove that the approach is viable. Further research could address such limitations by extending the scope of our approach to different formats, standards, and languages. These

Root word	% False Positive OCR	% False Positive Text Extr.
rischi	0,16	0,14
corruzione	0,14	0,14
rifiuti	0,08	0,09
discriminator	0,01	0,01
malatti	0,08	0,07
inquinant	0,01	0,01
spesa	0,01	0,01
emission	0,03	0,04
infortun	0,07	0,07
sanzion	0,05	0,05
incident	0,05	0,05
decess	0,03	0,03
pericolos	0,03	0,03
violazion	0,03	0,05
mort	0,02	0,02

Table 3: Percentage of contexts erroneously classified with negative sentiment containing the root word.

points are the challenges that we will face in future work, with the aim of making the system presented here as complete and reliable as possible. The sentiment analysis strategy presented here is the basis for in-depth analysis work that could be conducted on such reports. Elements such as subjectivity, writing style, and ease of reading could prove to be interesting information to assess the quality of such documents. This could have substantial managerial implications for firms willing to become aware of the actual interest of stakeholders in their sustainability reports.

7. Conclusion

Sustainability reporting should be considered as an impartial and transparent helpful tool to explain sustainable goals, objectives, and companies' activities to their stakeholders. These reports are an excellent resource for monitoring corporate best practices that address environmental, social, and economic sustainability. However, companies often produce reports which, even when they are compliant with GRI Standards or other reporting standards, are poorly structured and thus complex to read and analyze. Therefore, in this work we addressed the problem by proposing a system supporting the analysis of sustainability reports, specifically designed to identify the topics/disclosures discussed within GRI compliant reports. We propose a system based on a pipeline of Natural Language Processing and Information retrieval able to deal with closed format files, i.e. PDFs. The documents were transformed into a machine-readable textual format using OCR and Text Extraction tools. The text obtained here has been considered as the raw data over which to perform a keyword search operation. In particular, we considered keywords representative of the GRI topics and disclosures. This approach was repeated with different configurations of the system with the aim of optimizing the search process and, consequently, the final retrieval performances. The obtained results showed that the system we proposed is extremely performant on the considered dataset, showing a score of F1 measure equal to 0.94365. It was also observed that the text extraction strategy from the PDF format could strongly impact on the obtained results, suggesting a hybrid extraction mode to make up for the shortcomings of OCR and Text Extraction tools. The search strategy can also strongly affect performance. The design of the most correct keywords to be used has in fact proved to be a fundamental step in the implementation of the system. Finally, the sentiment analysis tool proved to be a useful component of the proposed system, by demonstrating that the examined reports seem to emphasize positive aspects rather than negative ones, which would contradict one of the GRI reporting principles (balance). A key future challenge is to enhance the system in terms of robustness, efficiency, effectiveness, and flexibility.

8. Acknowledgements

The work of Marco Polignano has been supported by Apulia Region, Italy through the project "Un Assistente Dialogante Intelligente per il Monitoraggio Remoto di Pazienti" (Grant n. 10AC8FB6) in the context of "Research for Innovation - REFIN". We would like to thank Francesco Mastrorosa who developed the dataset used in this work.

9. Bibliographical References

Aureli, S., Medei, R., Supino, E., and Travaglini, C. (2016). Sustainability disclosure after a crisis: A text mining approach. *International Journal of Social Ecology and Sustainable Development (IJSESD)*, 7(1):35–49.

Basile, P. and Novielli, N. (2014). Uniba at evalita 2014-sentipolc task predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. *UNIBA at EVALITA 2014-SENTIPOLC Task Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features*, pages 58–63.

Boiral, O. (2013). Sustainability reports as simulacra? a counter-account of a and a+ gri reports. *Accounting, Auditing and Accountability Journal*, 26(7):1036–1071.

Bowen, F. (2014). *After greenwashing: Symbolic corporate environmentalism and society*. Cambridge University Press.

Dyllick, T. and Muff, K. (2016). Clarifying the meaning of sustainable business: Introducing a typology from business-as-usual to true business sustainability. *Organization & Environment*, 29(2):156–174.

Global Reporting Initiative. (2013). Gri-g4 sustainability reporting guidelines—reporting principles and standard disclosures 2013. `http://www.globalreporting.org/resourcelibrary/`

GRIG4-Part2-Implementation-Manual.pdf. Accessed: 2022-04-01.

Grewal, J. and Serafeim, G. (2020). Research on corporate sustainability: Review and directions for future research. *Foundations and Trends® in Accounting*, 14(2):73–127.

Hsu, C. W., Wen-Hao, L., and Wei-Chung, C. (2013). Materiality analysis model in sustainability reporting: a case study at lite-on technology corporation. *Journal of Cleaner Production*, 57:142–151.

Jo, T. (2019). Introduction. In *Text Mining: Concepts, Implementation, and Big Data Challenge*, pages 3–17. Springer International Publishing, Cham.

Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of Library and Information Science*. Marcel Decker Inc., New York.

Lindgren, C., Huq, A. M., and Carling, K. (2021). Who are the intended users of csr reports? insights from a data-driven approach. *Sustainability*, 13(3).

Minutiello, V. and Tettamanzi, P. (2022). The quality of nonfinancial voluntary disclosure: A systematic literature network analysis on sustainability reporting and integrated reporting. *Corporate Social Responsibility and Environmental Management*, 29(1):1–18.

Modapothala, J. R. and Issac, B. (2009). Study of economic, environmental and social factors in sustainability reports using text mining and bayesian analysis. In *2009 IEEE Symposium on Industrial Electronics Applications*, volume 1, pages 209–214.

Polignano, M., Basile, P., Rossiello, G., de Gemmis, M., and Semeraro, G. (2017a). Learning inclination to empathy from social media footprints. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 383–384.

Polignano, M., Gemmis, M. d., Narducci, F., and Semeraro, G. (2017b). Do you feel blue? detection of negative feeling from social media. In *Conference of the Italian Association for Artificial Intelligence*, pages 321–333. Springer.

Polignano, M., Basile, P., de Gemmis, M., and Semeraro, G. (2019). A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68.

Smith, R. (2007). An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Székely, N. and vom Brocke, J. (2017). What can we learn from corporate sustainability reporting? deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLOS ONE*, 12(4):1–27, 04.

Uyar, A., Koseoglu, M. A., Kılıç, M., and Mehraliyev, F. (2021). Thematic structure of sustainability reports of the hospitality and tourism sector: A periodical, regional, and format-based analysis. *Current Issues in Tourism*, 24(18):2602–2627.

Wang, X., Yuen, K. F., Wong, Y. D., and Li, K. X. (2020). How can the maritime industry meet sustainable development goals? an analysis of sustainability reports from the social entrepreneurship perspective. *Transportation Research Part D: Transport and Environment*, 78:102173.

Zhou, Y., Wang, X., and Yuen, K. F. (2021). Sustainability disclosure for container shipping: A text-mining approach. *Transport Policy*, 110:465–477.

10. Language Resource References

Resource Type: Corpus
Resource Name: GRI-134-IT
Size: 134 documents
Resource Production Status: Newly created-finished
Language(s): Italian
Modality: Written
Use of the Resource: Information Extraction
Resource Availability: From Owner
License: Creative Commons rights reserved 4.0 - Noncommercial - Share Alike - International
Resource URL: https://github.com/marcopoli/GRI-Sustainability-Reports-Analysis/blob/master/GRI-134-IT.csv
Resource Description: This is a corpus of corporate sustainability reports that should be compliant with the Global Reporting Initiative standards.

11. Annexes

Contribution of authors:

- Marco Polignano: writing of Sections 1, 4, 5, 6, 7, process modeling and formalization
- Nicola Bellantuono: writing Section 2 (Related works)
- Francesco Paolo Lagrasta: writing Section 3 (Resources and Annotation Process)
- Sergio Caputo: process development, execution of experiments
- Pierpaolo Pontrandolfo: conceptualization, supervision
- Giovanni Semeraro: conceptualization, supervision