University of Bari
Department of Computer Science

Analysis of Global Reporting Initiative Sustainability Reports with NLP techniques

Student name: *Sergio Caputo*

Course: *Fundamentals of Artificial Intelligence* – Professor: *Stefano Ferilli*

Contents

Abstract

This paper aims to investigate different Natural Language Processing (NLP) approaches for retrieving relevant information from sustainability reports in accordance with the Global Reporting Initiative framework. The proposed workflow allows to identify and locate references to the various sustainability topics discussed by the reports, to extract the contexts of each reference and to provide a summary able to underline the most relevant discussed aspects.

1. Introduction

In recent decades corporate sustainability has become increasingly prominent in politics, management practice, reporting, and business science (Kolk, 2003 [1]; Barkemeyeret al., 2009 [2]). Customers, shareholders and investors ask for a greater transparency and regular disclosure of non-financial performance of companies.

Consumers and investors want to make informed choices and rational investment and that is the reason why they demand the disclosure of reliable information from enterprises (Epstein, 2008 [3]).
An organization can inform all interested parties about its economic, social and environmental activities by developing and issuing a sustainability report. CSR (Corporate Social Responsibility) reports are a voluntary business communication tool, which aim is to communicate the company attitudes towards assumptions of the CSR concept.

For the purpose of this paper sustainability reporting is considered as the practice of providing information to external and internal stakeholders on the economic, environmental and social results achieved by an organization. In literature and business practice most frequently used terminology for these kinds of practices are: sustainability reporting, corporate social responsibility reporting, non-financial disclosures.

Sustainability reporting is now a global standard. From year to year the number of reporting enterprises are growing. The European Union is the most active region in the world in terms of sustainability reporting. According to GRI statistics (GRI 2010), 45 % of published worldwide sustainability reports in 2010 year came from Europe.

2. Literature standards review

2.1. Global Reporting Initiative Standards

The Global Reporting Initiative Standards (GRI) represent global best practices for sustainability reporting of businesses, companies and institutions of any size anywhere in the world. The standards allow organizations to uniquely and uniformly measure their impact on planet Earth and make it public in a format that even non-experts in the field can understand.

It was originally a division of CERES (Coalition for Environmentally Responsible Economies) created to develop a sustainable accounting system that would allow companies to track their environmental impact. This would make it easier for them to pursue goals within broader social responsibility. The GRI department was then recognized as an independent body in 2002, when UNEP (United Nations Environment Programme) shared its principles for member nations to follow.

Sustainability reporting based on the Standards provides information about an organization's positive or negative contribution to sustainable development and allows it to report on its economic, environmental and social impacts, thereby increasing transparency on their contribution to sustainable development.

In addition to reporting companies, the Standards are highly relevant to many stakeholders - including investors, policymakers, capital markets, and civil society.

The GRI Standards, which are modular and inter-correlated, are primarily designed to be used as a set, to prepare a sustainability report focused on material issues, their related impact and how they are managed. (Fig.1)

The GRI Standards are a modular system comprising three series of Standards:

1. **GRI Universal Standards**:
   - *GRI 101 Foundation* explains how to use the Standards and how to draft the report. It also specifies the principles – such as accuracy, balance, and verifiability – fundamental to good-quality reporting.
   - *GRI 102 General Disclosures* contains disclosures relating to details about an organization's structure and reporting practices; activities and workers; governance; strategy; policies; practices and stakeholder engagement. These give insight into the organization's profile and scale, and help in providing a context for understanding an organization's impacts.
   - *GRI 103 Material Topics* explains the steps by which an organization can determine the topics most relevant to its impacts, its material topics, and describes how the Sector Standards are used in this process. It also contains disclosures for reporting its list of material topics; the process by which the organization has determined its material topics and how it manages each topic.
2. **GRI Sector Standard**: intends to increase the quality, completeness, and consistency of the reports of the organizations. Standards are developed for 40 sectors, starting with those with the highest impact. The standards list topics that are

likely to be relevant for most organizations in a given industry, and indicate the important information to report on these topics.

3. **GRI Topic Standards**: contain disclosures for providing information on topics. Examples include Standards on waste, occupational health and safety, and tax. Each Standard incorporates an overview of the topic and disclosures specific to the topic and how an organization manages its associated impacts.

Each Standard begins with a detailed explanation of how to use it. The companies may use all or part of some GRI Standards to report specific data. This can both be seen as the strength as well as the weakness of the GRI guidelines. In the years to come, the GRI has developed more and more, expanding the number of activities its principles address.
There exist three different sets of Topic Standards used in the referenced reports for this work, they cover respectively: Economy (GRI 200), Environment (GRI 300) and Social (GRI 400). Each of these Topic Standards is structured with a list of relevant disclosures about specific topics and each of these has fine grained informative to be reported.

These metrics gather accountability information, enabling businesses to identify potential risks and address them, possibly turning them into opportunities or strengths. In practice, the GRI Standards not only provide an opportunity for the business to change old polluting habits, but also analyze waste to reduce costs and increase efficiency in all production, storage and distribution processes. GRI's objective is to drive positive change and have a real impact on the social well-being of companies, keeping the focus on opportunities for better work for employees, more sustainability for the planet and the abolition, once and for all, of all forms of human exploitation.

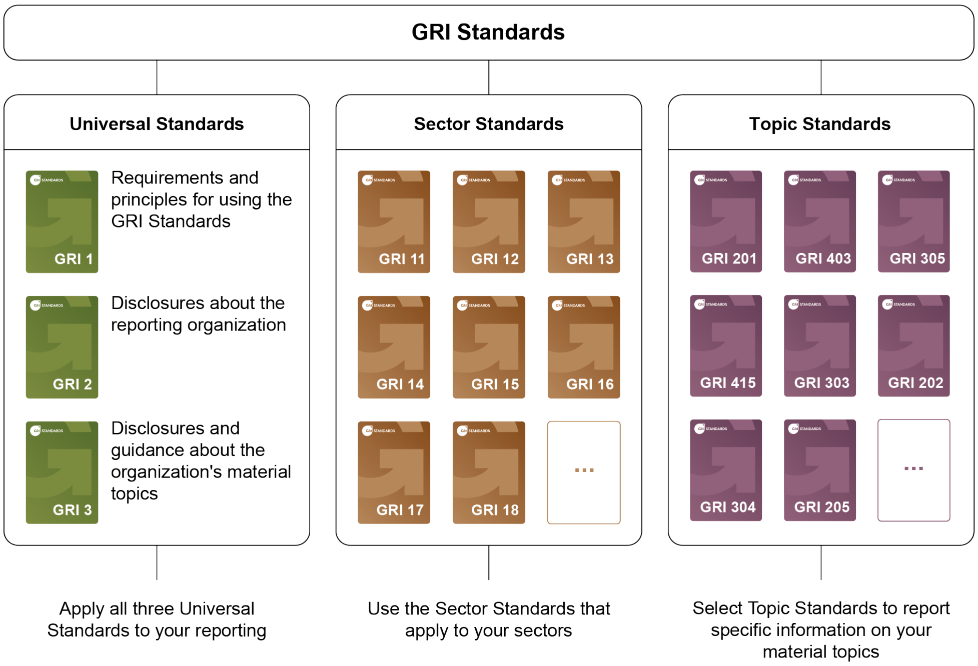

GRI Standards
Universal Standards
Sector Standards
Topic Standards
GRI 1
Requirements and principles for using the GRI Standards
GRI 2
Disclosures about the reporting organization
GRI 3
Disclosures and guidance about the organization's material topics
GRI 11
GRI 12
GRI 13
GRI 14
GRI 15
GRI 16
GRI 17
GRI 18
...
GRI 201
GRI 403
GRI 305
GRI 415
GRI 303
GRI 202
GRI 304
GRI 205
...
Apply all three Universal Standards to your reporting
Use the Sector Standards that apply to your sectors
Select Topic Standards to report specific information on your material topics

Figure 1: GRI Standard modules.

2.2. Non-financial Reporting Directive

Non-financial Reporting Directive (NFRD) is an amendment to the Accounting Directive (Directive 2013/34/EU) and was adopted in 2014. The disclosure of non-financial information is considered as vital for managing change towards a sustainable global economy by combining long-term profitability with social justice and environmental protection. The objective of the NFRD is therefore to raise the transparency of the social and environmental information provided by undertakings in all sectors to a similarly high level across all Member States and thus to improve the disclosure of non-financial information by certain large undertakings.

2.3. EU Corporate Sustainability Reporting Directive

The EU Corporate Sustainability Reporting Directive (CSRD), proposed on 21 April 2021, has a significantly extended scope than the Non-Financial Reporting Directive, applying to all large or listed companies operating in the EU. With a stated aim of bringing sustainability reporting on a par with financial reporting, it would help ensure both have equal weight and rigor.

It extends the reporting obligation to all large companies, all banks and all European insurance companies, whether quoted or unquoted, as well as all quoted companies, with the sole exception of quoted micro-companies. The reporting obligation is also extended to all groups, which will have to produce a consolidated sustainability report. In addition to listed micro-businesses, economic activities that are part of the consolidated reporting of the parent company, which is required to comply with the rules of European standards, are also excluded from the reporting obligation.

In numerical terms, applying the new inclusions and exclusions of the Directive, estimates predict that the 11,700 companies in Europe that are currently subject to the reporting obligation will increase to 49,000 economic activities.

3. GRI usage

A growing number of companies refers to the GRI as having inspired their reporting. Slightly over half (52.5%) of the Fortune Global 250 in the 2005 survey mentioned GRI in the report.
This, however, does not mean that companies follow the guidelines strictly, fully or consistently – they frequently take some components of the extensive set.

In 2005, 29% of the reports was specific about what parts of GRI were used and 6% declared to be in accordance with GRI. Companies sometimes also use the GRI guidelines to select the issues they include in their reports; this applied to 40% of the Fortune Global 250, and it turned out to be the most frequently mentioned tool to do this ('stakeholder consultation' in general was second with a score of 20%). Another way is which GRI turns out to play some role is in external verification –9% mentioned that the guidelines were part of this in one way or the other.

Studies within the available literature have pointed out that organizations have both positive and negative motivations to create sustainability reports. The positive motivations are linked to transparency and accountability, whereas the negative motivations tend to be linked to superficial aspects as enhancing an organization's image and decision-making direction sense, without substantive change. Different kind of problems have been found in the reporting model proposed by the GRI, it allows companies to report the positive facts only, while omitting the negative information. Therefore, legitimacy with no real transparency would be possible, allowing the report to be turned into a mere device to simulate sustainable positioning and to improve company's image. (Quilice al. 2018 [4])

4. Reporting Analysis

The automatic analysis of sustainability-related textual documents was the primary focus of this project. In particular, we wanted 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 and then use the associated descriptions to produce a summary for each of the disclosures. Since the GRI Standards are by far the most widely used, we concentrated our efforts in this work on sustainability reports that adhere to them. This task, which might seem simple, is instead made complex by the heterogeneity of layouts and the writing styles of sustainability documents.

4.1. Dataset

In order to accomplish the proposed goal we created an ad hoc dataset of reports. The collection of PDF files was created collecting annual reports made publicly available by 134 Italian companies from 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 European norms, which state that non-financial disclosures are mandatory for large companies.
A large amount of companies publish environmental reports at least once per year, some other also quarterly depending on time-based objectives and goals to reach.

Companies are not forced to use a predefined layout design standard in drafting of reports: different designing choices are taken for characters style, sections design and disposition, graphs, images, number of pages, etc. Most of the reports are divided into 5 main sections, including management information, environment and climate change, environmental performance review, listing of verifiable environmental claims and green initiatives and declarations about environmental compliance.
In addition, also information about internal organization'structure, departments responsible, employees' roles in sustainability activities are provided in the reports.

5. Workflow

In order to accomplish the final summarizing task, a specific and detailed procedure was defined (Figure 2). Different sub-tasks were necessary to be fulfilled before reaching the described goal (Figure 3).

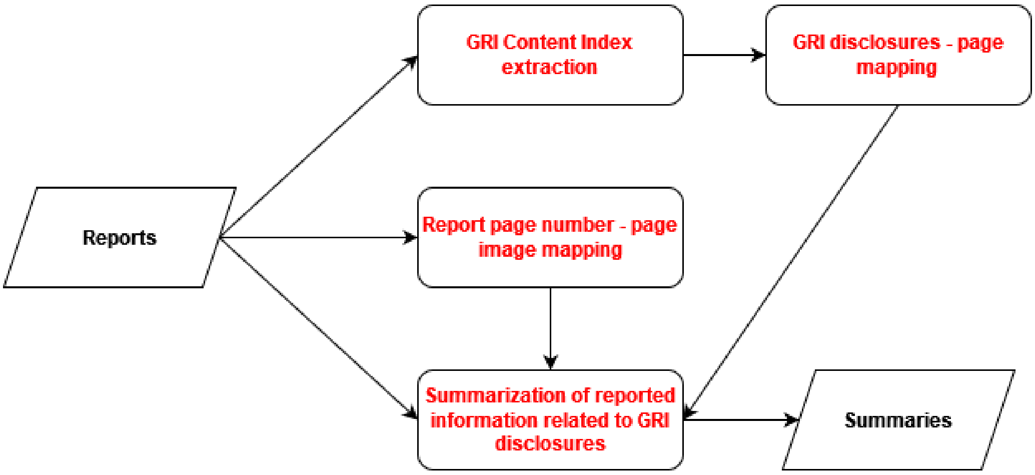

GRI Content Index extraction
GRI disclosures - page mapping
Reports
Report page number - page image mapping
Summarization of reported information related to GRI disclosures
Summaries

Figure 2: General workflow.

5.1. GRI Content Index extraction

Sustainability reports written with respect to the GRI Standards must contain a GRI Content Index. The content index makes reported information traceable and increases the report's credibility and transparency. The content index provides an overview of the organization's reported information and helps stakeholders navigate the report at a glance. It specifies the GRI Standards that the organization has used. The index also lists the location, such as a page number or URL, for all disclosures that the organization has used to report on its material topics. The content index can also help a stakeholder understand what the organization has not reported. The organization must specify in the content index if a 'reason for omission' is being used. In addition, the disclosure or the requirement that the organization cannot comply with, together with an explanation, must be listed in the content.

First, we had to locate the GRI Content Index for each report with the aim of detecting all the reported GRI disclosures and the related page numbers of the report in which the company described the details about them. So as to fulfill this task, it was necessary to generate the Table of Contents in order to locate the pages containing the GRI content indexes tables.

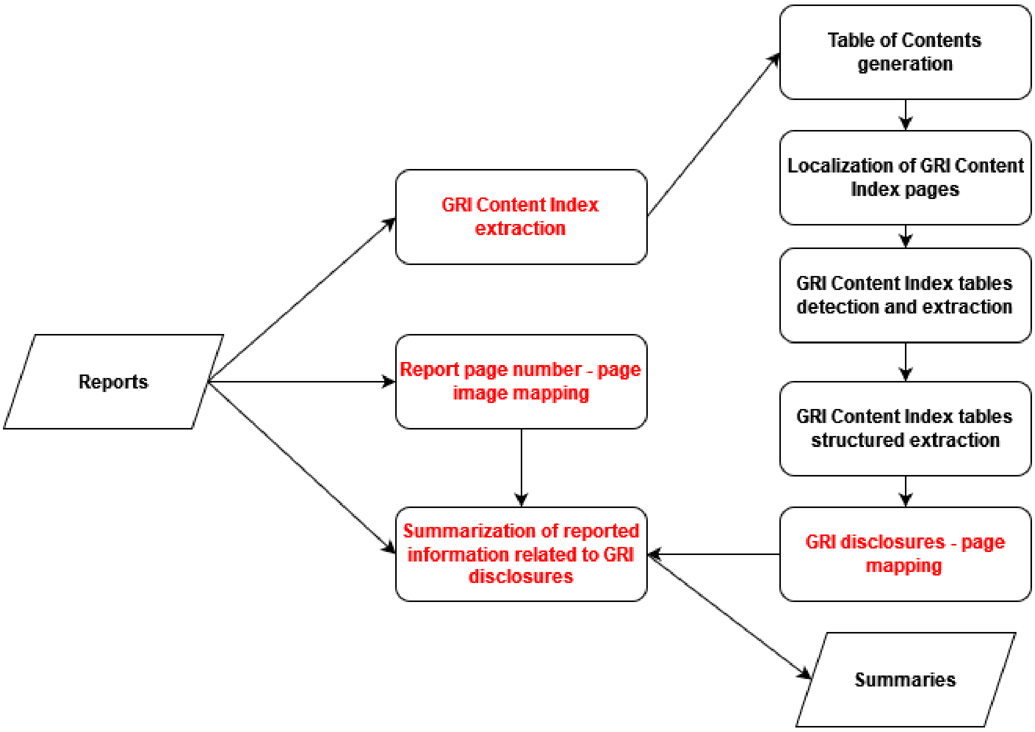


Table of Contents generation
Localization of GRI Content Index pages
GRI Content Index tables detection and extraction
GRI Content Index tables structured extraction
GRI disclosures - page mapping
GRI Content Index extraction
Reports
Report page number - page image mapping
Summarization of reported information related to GRI disclosures
Summaries

Figure 3: Detailed workflow provided with sub-tasks.

5.1.1. Table of Contents generation

Starting from the initial reports collection of PDFs files, each of these had to be converted into a set of images that could be processed by an OCR (Optical Character Recognition) tool in order to extract a textual representation of the pages inside the report.
First of all we had to make a suitable input for the OCR tool. For this purpose we used Poppler, 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.
Considering the collection of images analogous to each report, these were given as input to the model made available by Detectron2 trained over PubLayNet dataset.
Detectron2 was built by Facebook AI Research (FAIR) to support rapid implementation and evaluation of novel computer vision research. It includes implementations for the object detection and segmentation.
For this specific sub-task, Detrctron2 goal was concerning the detection of titles in reports pages. Afterwards detection was performed, the text contained in the detected boxes was extracted adopting Tesseract, 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.

Given the titles extracted and the images from which extractions were performed, we were able to generate a ToC containing the mapping between the title and the related image for each report.

5.1.2. Localization of GRI Content Index pages

Since the final goal of this sub-task was about extracting the pages pertinent to the GRI Content Index, the Tables of Content previously generated were used to locate the pages in which these indexes were contained.
Reports provided by companies, organizations and institutions often specify the titles associated to the GRI Content Index sections in different ways, for this reason, we conducted an analysis of the titles used in our dataset to introduce the index tables. All the retrieved titles were considered for extracting the associated GRI Content Index tables in the reports. (Table 1)

Introductory titles to GRI Content Index
GRI Content Index
INDICE DEI CONTENUTI GRI
TABELLA DI CORRELAZIONE GRI STANDARD
TABELLA DI CORRELAZIONE TRA TEMI MATERIALI, GRI STANDARD E SDGs
CORRISPONDENZA ASPETTI E INDICATORI GRI CONTENT INDEX
CORRELAZIONE TEMI MATERIALI – GRI STANDARD
TAVOLA DEI CONTENUTI GRI
TABELLA DI RACCORDO CON IL GRI
Index of GRI contents
Tabella degli standard GRI
GRI STANDARDS CONTENT INDEX
Tavole di riepilogo degli indicatori GRI
Indice GRI
CONTENT INDEX GRI
Tabella di raccordo GRI Standard
GRI Standard Content Index
GRI index
GRI contents reference table
CONFORMITÀ ALLO STANDARD GRI E AL D.LGS. 254/2016
Tabella di correlazione con gli indicatori GRI Standards
TABELLA DI CORRISPONDENZA STANDARD GRI
GRI Standards utilizzati nel presente documento
GRI TABLE OF CONTENTS
GRI CONTENT TABLE
Tabella GRI-Referenced
Tabella di correlazione al D. Lgs. 254/16 e al GRI Referenced
STANDARD CONTENTS AND GRI INDICATORS
Summary of GRI indicators
GRI standards
TABELLA DEGLI INDICATORI GRI

Table 1: Titles used to introduce GRI Content Index tables

5.1.3. GRI Content Index tables detection and extraction

At this point, given the indications about the pages containing the tables of interest about GRIs, table detection and extraction could be achieved.
Table detection is performed to identify all regions in images that contain tables, while Table Structure Recognition involves identifying their components, i.e. rows, columns, and cells, to finally identify the entire table structure.
Table detection was fulfilled by adopting the Multi-stage Pipeline for Table Detection and Table Structure Recognition proposed by Pascal Fischer et al. [5] In this work they specified two modules: for Table Detection the approach used was fully data-driven based on a Convolutional Neural Network (CNN) trained on the TableBank dataset, able to localize tables inside images and forward them to Table Structure Recognition. For the latter, they used a deterministic algorithm in order to address three table types (unbordered tables, bordered tables and partially bordered tables).

Since the second module proposed by this work was computationally too expensive to be run on a Google Colab environment, we implemented a runnable Table Structure Recognition module in Colab.

5.1.4. GRI Content Index tables structured extraction

Tables extracted as images in the previous phase were given as input to PaddleOCR, an OCR framework or toolkit which provides multilingual practical OCR tools that help the users to apply and train different models in a few lines of code. PaddleOCR offers a series of high-quality pre-trained models. This contains three types of models to make OCR highly accurate and close to commercial products. It provides Text detection, Text direction classifier and Text recognition.

In our case, we used PaddleOCR to process every extracted table image and obtain as output:

- Coordinates of the bounding boxes containing detected text
- Detected text contained in the boxes
- Probability that defines the degree of correctness of the detection

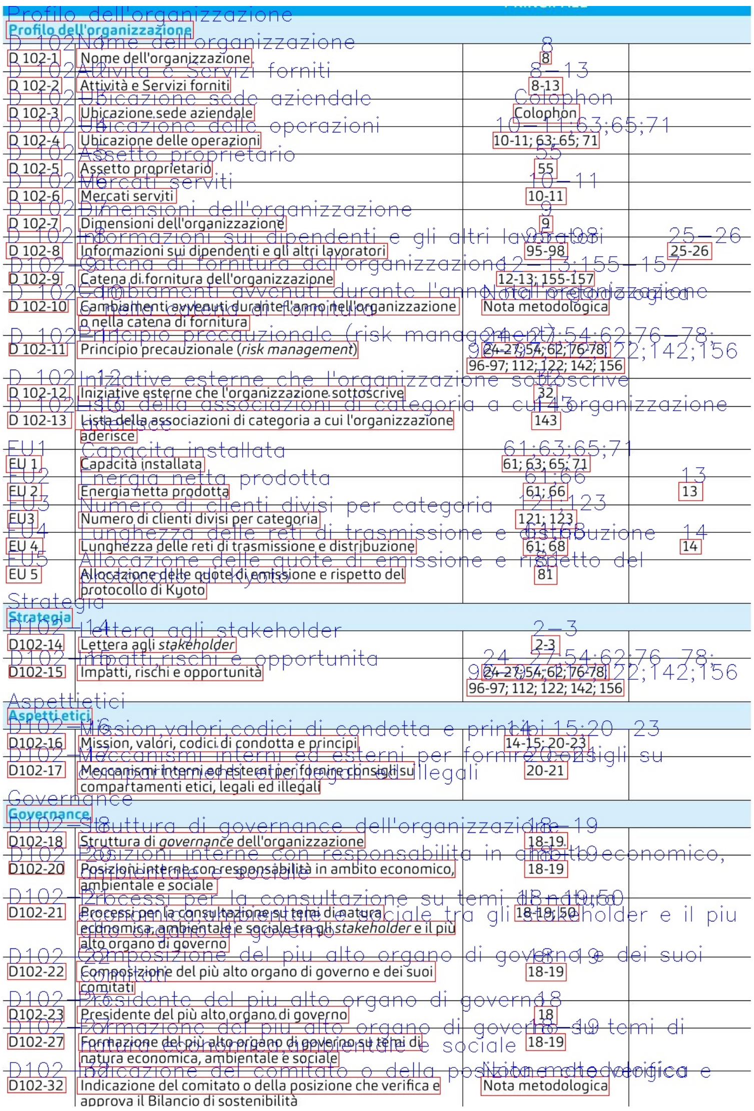
Considering the bounding boxes detected (Fig.4) we could start the process aimed at the text reconstruction performed maintaining the original table layout.

1. Each bounding box was extended right to the edges of the table (horizontal lines). Since boxes that are in the same horizontal line of the table occupy the same region if extended, we were interested in finding out the extended boxes that had the greatest intersection score.
2. Same procedure was performed for the vertical lines, each bounding box was extended vertically right to the bottom of the table.

3. In order to find out the best extended bounding box for boxes in the same horizontal/vertical line, Non Max Suppression algorithm was used to return the box with the highest IoU (Intersection over Union) and with maximum probability (Fig.5).

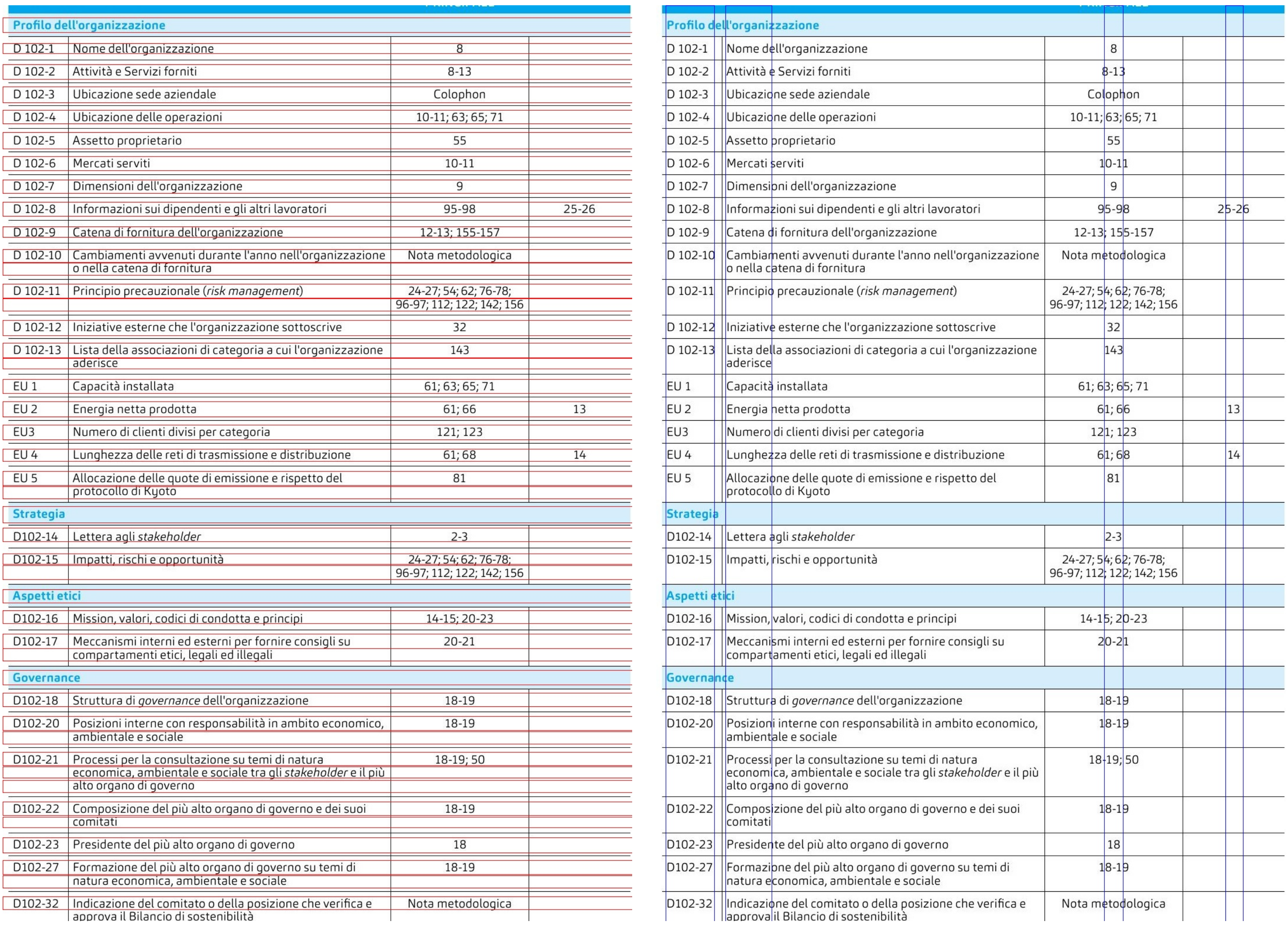
4. At this point we retrieved all the main vertical and horizontal line boxes (columns and rows of the table). Intersections computed between vertical and horizontal boxes were used to get the boxes that coincided the most with the intersection ones, also in this case IoU was used defining a threshold of 0.1. These were the corresponding cells in the table which contained text.

5. Text retrieved was stored in a matrix and used to create a textual representation of the table with CSV (Comma Separated Values) format file.



Profilo dell'organizzazione			
D 102-1	Nome dell'organizzazione	8	
D 102-2	Attività e Servizi forniti	8-13	
D 102-3	Ubicazione sede aziendale	Colophon	
D 102-4	Ubicazione delle operazioni	10-11; 63; 65; 71	
D 102-5	Assetto proprietario	55	
D 102-6	Mercati serviti	10-11	
D 102-7	Dimensioni dell'organizzazione	9	
D 102-8	Informazioni sui dipendenti e gli altri lavoratori	95-98	25-26
D 102-9	Catena di fornitura dell'organizzazione	12-13; 155-157	
D 102-10	Cambiamenti avvenuti durante l'anno nell'organizzazione o nella catena di fornitura	Nota metodologica	
D 102-11	Principio precauzionale (*risk management*)	24-27; 54; 62; 76-78; 96-97; 112; 122; 142; 156	
D 102-12	Iniziative esterne che l'organizzazione sottoscrive	32	
D 102-13	Lista della associazioni di categoria a cui l'organizzazione aderisce	143	
EU 1	Capacità installata	61; 63; 65; 71	
EU 2	Energia netta prodotta	61; 66	13
EU3	Numero di clienti divisi per categoria	121; 123	
EU 4	Lunghezza delle reti di trasmissione e distribuzione	61; 68	14
EU 5	Allocazione delle quote di emissione e rispetto del protocollo di Kyoto	81	
Strategia			
D102-14	Lettera agli *stakeholder*	2-3	
D102-15	Impatti, rischi e opportunità	24-27; 54; 62; 76-78; 96-97; 112; 122; 142; 156	
Aspetti etici			
D102-16	Mission, valori, codici di condotta e principi	14-15; 20-23	
D102-17	Meccanismi interni ed esterni per fornire consigli su compartamenti etici, legali ed illegali	20-21	
Governance			
D102-18	Struttura di *governance* dell'organizzazione	18-19	
D102-20	Posizioni interne con responsabilità in ambito economico, ambientale e sociale	18-19	
D102-21	Processi per la consultazione su temi di natura economica, ambientale e sociale tra gli *stakeholder* e il più alto organo di governo	18-19; 50	
D102-22	Composizione del più alto organo di governo e dei suoi comitati	18-19	
D102-23	Presidente del più alto organo di governo	18	
D102-27	Formazione del più alto organo di governo su temi di natura economica, ambientale e sociale	18-19	
D102-32	Indicazione del comitato o della posizione che verifica e approva il Bilancio di sostenibilità	Nota metodologica	

Figure 4: Bounding boxes detected by PaddleOCR.



Profilo dell'organizzazione			
D 102-1	Nome dell'organizzazione	8	
D 102-2	Attività e Servizi forniti	8-13	
D 102-3	Ubicazione sede aziendale	Colophon	
D 102-4	Ubicazione delle operazioni	10-11; 63; 65; 71	
D 102-5	Assetto proprietario	55	
D 102-6	Mercati serviti	10-11	
D 102-7	Dimensioni dell'organizzazione	9	
D 102-8	Informazioni sui dipendenti e gli altri lavoratori	95-98	25-26
D 102-9	Catena di fornitura dell'organizzazione	12-13; 155-157	
D 102-10	Cambiamenti avvenuti durante l'anno nell'organizzazione o nella catena di fornitura	Nota metodologica	
D 102-11	Principio precauzionale (*risk management*)	24-27; 54; 62; 76-78; 96-97; 112; 122; 142; 156	
D 102-12	Iniziative esterne che l'organizzazione sottoscrive	32	
D 102-13	Lista della associazioni di categoria a cui l'organizzazione aderisce	143	
EU 1	Capacità installata	61; 63; 65; 71	
EU 2	Energia netta prodotta	61; 66	13
EU3	Numero di clienti divisi per categoria	121; 123	
EU 4	Lunghezza delle reti di trasmissione e distribuzione	61; 68	14
EU 5	Allocazione delle quote di emissione e rispetto del protocollo di Kyoto	81	
Strategia			
D102-14	Lettera agli *stakeholder*	2-3	
D102-15	Impatti, rischi e opportunità	24-27; 54; 62; 76-78; 96-97; 112; 122; 142; 156	
Aspetti etici			
D102-16	Mission, valori, codici di condotta e principi	14-15; 20-23	
D102-17	Meccanismi interni ed esterni per fornire consigli su compartamenti etici, legali ed illegali	20-21	
Governance			
D102-18	Struttura di *governance* dell'organizzazione	18-19	
D102-20	Posizioni interne con responsabilità in ambito economico, ambientale e sociale	18-19	
D102-21	Processi per la consultazione su temi di natura economica, ambientale e sociale tra gli *stakeholder* e il più alto organo di governo	18-19; 50	
D102-22	Composizione del più alto organo di governo e dei suoi comitati	18-19	
D102-23	Presidente del più alto organo di governo	18	
D102-27	Formazione del più alto organo di governo su temi di natura economica, ambientale e sociale	18-19	
D102-32	Indicazione del comitato o della posizione che verifica e approva il Bilancio di sostenibilità	Nota metodologica	

Figure 5: Horizontal and vertical lines obtained with Non Max Suppression.

5.2. GRI disclosures - page mapping

After obtaining a textual reconstruction of the tables, we had to define a mapping between the GRI disclosures and the related page numbers indicated in the tables. The collection of this information was necessary because it was essential for retrieving for each disclosure the pages on which the company reported the situation considering the sustainable topic of the disclosure.

For this scope, a JSON dictionary was created for each report. Keys specified were the GRI codes of interest defined by ”GRI <code>” or just ”<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 302-4” or ”GRI 203” or ”306-4” are examples of GRI codes.

Every key related to a disclosure code found in the extracted table was associated in the dictionary to the page numbers retrieved for that GRI code.

5.3. Report page number - page image mapping

Since each report is designed differently from the others, layouts had to be studied to perform the page number extractions.
Considering that each image was generated by converting a single PDF page to JPG format, we had to deal also with layouts in which for each PDF page there were actually two pages. Furthermore, layouts differed also for the placement of the page numbers on the page, there could be observed reports that illustrated these in different positions. According to these details emerged during the analysis, we decided to take into account a strategy that could deal with the detection and extraction of page numbers located in the headers and footers, the two most commonly used layout sections for this purpose.
Since PDF reports were not equipped with information about headers and footers, each page in JPG format was handled to extract them. In our case, we stated that the areas reserved for headers and footers were placed in the upper and lower 10% of the pages. Considering this, the two sections were cut out for each JPG page and Tesseract OCR engine was adopted to extract the text contained in these areas (page numbers).
Text extracted related to page numbers was then stored in JSON files which allowed the mapping between the page in JPG format and the corresponding page number used in the report.

5.4. Summary of reported information related to GRI disclosures

Lastly, summaries about the reported GRI disclosures were produced. This final task was achieved considering: the collection of page images for each report, the produced mapping between each GRI disclosure and page in which it is detailed (Section 5.2) and the mapping between each image related to a page of the report and its extracted page number (Section 5.3).

First of all we standardized the pages patterns in the mapping between GRI disclosure and reference pages used to indicate the set of pages related to a GRI. As mentioned previously in the pertinent section, each detected GRI inside the GRI Content Index tables was collected in the JSON mapping file and associated to the pages in which the GRI disclosure discussion was reported. This set of pages could be defined by a list of single pages related to that argument, or ranges of pages or set defined by both single and ranges. These patterns were standardized in order to obtain lists of single pages easy to handle for text extraction and subsequent summarization.
Afterwards, each page in the list of pages was searched among the mapping file contents (page image - page number mapping) to individuate which were the page images to be considered for text extraction. Considering the images recovered for each GRI disclosure, these were fed to Tesseract OCR engine to extract all the text contained in those pages. The extracted text was used to perform the final task related to the summarization.

5.4.1. Summarization approach adopted

For the last step, the text summarization approach exploited was the one proposed by Rossiello al. [6]. Summaries produced rely on the word embeddings, thus also sentences that contain words with the same meaning as the most relevant words (centroid) are selected even if the words are different. The embedding model is based on Word2vec and it was trained on the text that would have been summarized. Suddenly, after splitting the text into sentences and performing cleanup operations, the resulting sentences were used to find the most relevant words (centroid) of the text with the TF-IDF. Centroid words vectors were summed up to obtain the embedding representation of the centroid. In order to compare sentences to the centroid embedding, the algorithm calculated the embedding representation of each sentence and cosine similarity was used to calculate the similarity between the centroid embedding and the sentence embedding to obtain a similarity score. Sentences were selected according to their score. The number of sentences selected was limited by how many words the summary had to contain. When too similar sentences were part of the summary, redundancy was handled comparing the sentences already in the summary with the candidate sentence using cosine similarity. If the chosen sentence was too similar to one in the summary, it would not be added to the final text.

6. Results

6.1. Evaluation with ROUGE metrics

The performance of the automatic generated summaries with GRI disclosures texts was evaluated using the ROUGE measure. It compares an automatically generated summary with a reference text or summary (which is typically created by a human expert) and calculates the overlap between the two. The reference summary is treated as the "ground truth" and the automatically generated summary is evaluated based on how well it matches the reference.

In our case ROUGE-2 was used to compare the original text and the system generated summary considering overlapping bigrams. On average, the summaries generated had a Rouge score of 0.49, indicating that they contained a high overlap of bigrams with the reference text due to precision equal to 0.99. Precision expressed how much of the system summary was in fact relevant or needed. Instead, recall average value was 0.37, this indicated how much of the reference text the system summary was recovering or capturing.
This suggests that the workflow was able to effectively identify and extract the most important information contained for each GRI disclosures while making the summary non-verbose selecting useful words.

$$ROUGE2_{Precision} = \frac{number\_of\_overlapping\_bigrams}{total\_bigrams\_in\_system\_summary}$$

$$ROUGE2_{Recall} = \frac{number\_of\_overlapping\_bigrams}{total\_bigrams\_in\_reference\_text}$$

$$ROUGE2 = 2 * \frac{ROUGE2_{precision} * ROUGE2_{recall}}{ROUGE2_{precision} + ROUGE2_{recall}}$$

6.2. Qualitative evaluation

In order to determine whether the workflow proposed in this paper extracted, processed, and produced summaries consistent with the reported GRI codes, a qualitative evaluation was performed manually.

Due to the large number of reports, those considered for this analysis were the top five reports which obtained the best Rouge scores. For each of these, we conducted an analysis aimed at verifying whether the referenced pages were correctly extracted from the GRI Content Indexes and whether the extracted pages were properly mapped to the corresponding page images.
In total, the cases analyzed among the five reports were 18, each of these coincided with a GRI disclosure code for which we had all the information about the extracted pages, page images, text contained in these pages and the produced summary.

In 83% of cases, the pages mentioned into the tables in the GRI Content indexes sections were correctly extracted. The remaining 17% were subject to incorrect extractions due to indications in the table related to paragraphs and sections of the report rather than pages, or even changes in the layout of the GRI Content Index tables that offset the index related to the column used to indicate the reference page numbers for each GRI. Consequently, these wrong extractions for the page numbers generated a mapping to the images of the mismatched pages.

For all the cases with correct pages extraction, 67% provided also exact mapping to the corresponding page images. For the other 33%, page numbers were incorrectly mapped to mismatched page images. The results obtained underlined how the proposed module for mapping page numbers and page images could definitely affect the general performance of the designed workflow.
Dealing with a multitude of reports with different layouts wherein various design choices were carried out by the companies, impacted the last mapping module. The absence of a rigid standardized layout, or at least, of general strict rules to write reports with respect of a reference framework, made the mapping task complex. We had to develop a general solution based on OCR for page numbers extraction that could consider page numbers specified in both the headers and footers, but of course, the text extracted in these areas could be affected by other graphical elements, footer notes, sentences, chapter/section titles, etc. The presence of these superfluous elements caused noise within the extracted text, so cleaning and preprocessing procedures were implemented, but these could not completely solve the problem and the results proved it.

7. Conclusions and future works

The proposed work turned out many companies base their sustainability reports on GRI standards or other reporting standards, but despite this, reports are poorly structured and thus complex to read and analyze. In this work we addressed the problem by proposing a supporting system for the analysis of sustainability reports, specifically designed to identify the topics/disclosures discussed within GRI compliant reports. We proposed different Natural Language Processing and Information Retrieval approaches able to deal with closed format files, i.e. PDFs, in a specific workflow. This may help stakeholders to analyze and use the report in an efficient and faster way than usual, providing disclosure complaint summaries generated from pages related to each GRI code. Firstly, the documents were employed to extract the GRI Content Index. This preliminary task was necessary to establish which disclosures and related reference pages were reported. In order to fulfill this task, a sequence of sub-tasks was performed: Table of Contents generation, localization of GRI Content Index pages, detection and structured extraction of the tables in the GRI Content Index. Afterward, the second task was about mapping the GRI disclosures to the reference pages. Then, the mapping between the actual page numbers and the image of the page was generated in order to access the correct pages related to the disclosures. Finally, given the two previous mappings, summaries concerning specific GRI disclosures were produced. The adopted summarization approach relied on centroids of word embeddings. The obtained results showed that the workflow was able to produce high-quality summaries that accurately captured the key points contained in the original texts. A qualitative analysis underlined the good performances of the method designed for extracting the reference pages from the GRI Content Index tables, on the other hand, the module for mapping the extracted pages to corresponding page images performed slightly worse due to biased page numbers extractions. Overall, the results obtained with the proposed workflow indicate that it is a promising approach for summarizing GRI disclosure texts and could be a valuable tool for stakeholders seeking to quickly understand and use the report. Further studies and evaluations are needed to confirm and refine the effectiveness of the workflow considering a larger dataset of reports.

References

[1] Kolk, A. (2003), Trends in sustainability reporting by the Fortune Global 250. Bus. Strat. Env., 12: 279-291. https://doi.org/10.1002/bse.370

[2] Barkemeyer, Ralf & Figge, Frank & Hahn, Tobias & Holt, Diane. (2009). What the Papers Say: Trends in Sustainability. A Comparative Analysis of 115 Leading National Newspapers Worldwide. Journal of Corporate Citizenship. 2009. 69-86. http://dx.doi.org/10.9774/GLEAF.4700.2009.sp.00009

[3] Epstein, M.J.: Making Sustainability Work. Best Practices in Managing and Measuring Corporate Social, Environmental, and Economic Impacts. Greenleaf Publishing, Sheffield (2008) http://dx.doi.org/10.4324/9781351276443

[4] Quilice, TF, Cezarino, LO, Alves, MFR, Liboni, LB, Caldana, ACF. Positive and negative aspects of GRI reporting as perceived by Brazilian organizations. Environ Qual Manage. (2018); 27: 19– 30. https://doi.org/10.1002/tqem.21543

[5] Pascal Fischer and Alen Smajic and Alexander Mehler and Giuseppe Abrami. Multi-Type-TD-TSR – Extracting Tables from Document Images using a Multistage Pipeline for Table Detection and Table Structure Recognition: from OCR to Structured Table Representations. (2021) https://arxiv.org/abs/2105.11021

[6] Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based Text Summarization through Compositionality of Word Embeddings. In Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, pages 12–21, Valencia, Spain. Association for Computational Linguistics. https://aclanthology.org/W17-1003