

An NLP Pipeline for detecting GRI Indexes from Sustainability Reports

Marco Polignano^{1,*}, Sergio Caputo¹, Nicola Bellantuono², Francesco Paolo Lagrasta³, Pierpaolo Pontrandolfo³, Giovanni Semeraro¹ and Stefano Ferilli¹

¹University of Bari, via E. Orabona 4, Bari, 70125, Italy

²University of Foggia, via Napoli 25, 71122, Foggia

³Polytechnic University of Bari, via E. Orabona 4, Bari, 70125, Italy

Abstract

Communicating information on a large scale about sustainable development has become an annual obligation in several nations for many types of businesses. Sustainability reports inform stakeholders about a company's commitment to sustainable development and its economic, social, and environmental sustainability policies. However, because norms and standards enable drafting companies to have some freedom, such reports are scarcely similar in style, disclosures, key performance indicators (KPIs), and so on. In this paper, we offer a system based on natural language processing and information extraction approaches for retrieving essential information from sustainability reports produced in Italian and English that are consistent with the Global Reporting Initiative Standards. The algorithm can specifically detect references to the many sustainability topics mentioned in the reports: which page of the document those references were found on, the context of each reference, and its summary. The system's output was then compared to ground truth derived by a manual annotation method on 134 reports. The experimental results demonstrate the approach's affordability for boosting sustainability disclosures, accessibility, and transparency, allowing stakeholders to perform additional analysis and considerations. Even with limited data, our approach can actively support stakeholders in successfully dealing with specific analysis tasks.

Keywords

Sustainability Reports, Natural Language Processing, GRI, Information Seeking

1. Introduction

Corporate sustainability has gained prominence in economics, management practice, marketing, and business science in recent decades. Consumers, stockholders, and investors want increased transparency and frequent disclosure of a company's non-financial performance [1, 2]. Consumers and investors want to make educated decisions and make sound investments, which is why they expect businesses to provide credible information. A sustainability report, developed and issued by an organization, can inform all interested parties about its economic, social, and environmental activities. CSR (Corporate Social Responsibility) reports are a voluntary corporate communication tool that attempts to express the company's views about the CSR concept's assumptions. For the purposes of this article, sustainability reporting is

defined as the practice of giving information to external and internal stakeholders about an organization's economic, environmental, and social outcomes. The most commonly used nomenclature in literature and business practice for these types of procedures is sustainability reporting, corporate social responsibility reporting, and non-financial disclosures. Sustainability reporting has become a worldwide norm. The number of reporting businesses is increasing year after year. Regarding sustainability reporting, the European Union is the most active area globally. Under the umbrella of Global Reporting Initiative (GRI) Standards, which are the most widespread guidelines in the field of sustainability reporting, there are three sets of Topic Standards utilized in the cited reports for this study, and they are as follows: Economic (GRI 200), Environment (GRI 300), and Social (GRI 400). Each of these Subject Standards is organized around a list of relevant disclosures concerning specific topics, and each includes fine-grained information to disclose. These indicators collect accountability data, allowing firms to detect and handle possible risks, perhaps turning them into opportunities or strengths. In fact, the GRI Standards allow businesses to modify old polluting behaviors and evaluate pollution to cut costs and boost efficiency in all production, storage, and distribution processes. GRI's objective is to drive positive change and have a tangible impact on the social well-being of companies, focusing on opportunities for better


Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

✉ marco.polignano@uniba.it (M. Polignano);
s.caputo34@studenti.uniba.it (S. Caputo);
nicola.bellantuono@unifg.it (N. Bellantuono);
francescopaolo.lagrasta@poliba.it (F. P. Lagrasta);
pierpaolo.pontrandolfo@poliba.it (P. Pontrandolfo);
giovanni.semeraro@uniba.it (G. Semeraro); stefano.ferilli@uniba.it (S. Ferilli)

ORCID 0000-0002-3939-0136 (M. Polignano)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

work for employees, more sustainability for the planet, and the abolition, once and for all, of all forms of human exploitation.

2. NLP Pipeline

A rising number of companies credit the GRI with inspiring the writing of their reports. This does not imply that businesses strictly, entirely, or continuously follow the principles. They typically take various components from the extensive set without strictly following the standard. 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 extract relevant information for possible stakeholder consultation and review automatically [3]. To achieve the intended purpose, we built an ad hoc dataset of reports. The PDF file collection was generated by gathering annual reports made publicly available by 134 Italian enterprises from 27 various industries, divided into micro (2/134), small (2/134), medium (4/134), and large (126/134). Many businesses issue environmental reports at least once a year, and some even quarterly, depending on time-based objectives and targets to achieve. In order to accomplish the final summarizing task, a specific and detailed procedure was defined (Figure 1). Starting from the initial collection of PDFs files, each had to be converted into a set of images that could be processed by an OCR (Optical Character Recognition) tool to extract a textual representation of the pages inside the report. Detectron2 [4] was used for the detection of titles in report pages. Given the titles extracted and the images from which extractions were performed, we generated a Table of Contents containing the mapping between the title and the related image for each report. Since the final goal of this sub-task was to extract 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. 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. 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 sustainability 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", "GRI 203" or "306-4" are examples of GRI codes. Lastly, summaries of the reported GRI disclosures were produced. This final task was achieved considering the collection of page images for each report and the mapping between each GRI disclosure and the page on which it is detailed. The text summarization approach exploited was the one proposed by Rossiello et al. [5]. The summaries produced rely on word embeddings. Thus, 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 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 word 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 many similar sentences were part of the summary, redundancy was handled by comparing the sentences already in the summary with the candidate sentence using cosine similarity. If the chosen sentence were too similar to one in the summary, it would not be added to the final text. 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 they contained a high overlap of bigrams with the reference text. Instead, the recall average value was 0.37, indicating how much the system summary was recovering or capturing the reference text. This suggests that the workflow was able to effectively identify and extract the most important information contained in each GRI disclosure while making the summary non-verbose by selecting valuable words.

3. Conclusion

In this work, we addressed the problem of Sustainability Reports Analysis by proposing a system to identify the topics/disclosures discussed within GRI-compliant reports. Overall, the results obtained indicate a promising approach for summarizing GRI disclosure as a valu-

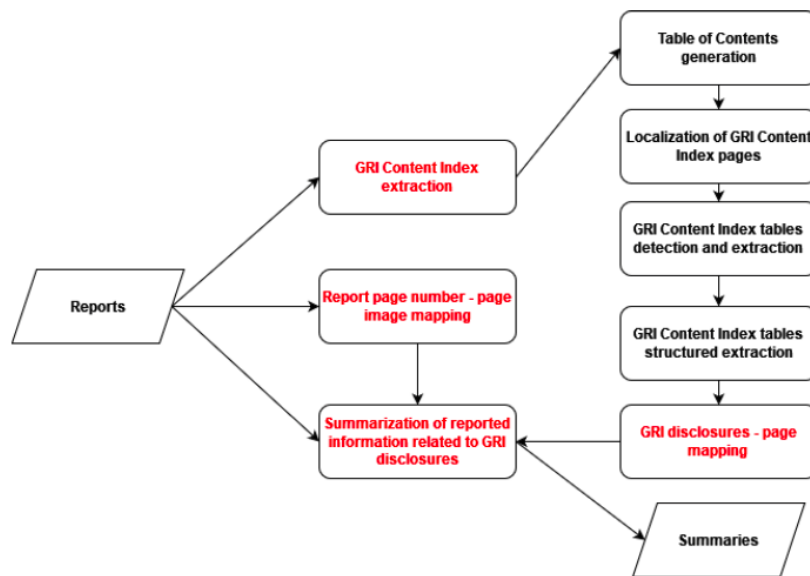


Figure 1: Workflow of the NLP pipeline.

able tool for stakeholders seeking to understand and use the report quickly. Further studies and evaluations are needed to confirm and refine the effectiveness of the workflow, considering a larger dataset of reports.

evaluation across source types and genres, 2017, pp. 12–21.

References

- [1] A. Kolk, Trends in sustainability reporting by the fortune global 250, *Business strategy and the environment* 12 (2003) 279–291.
- [2] R. Barkemeyer, F. Figge, D. Holt, T. Hahn, What the papers say: Trends in sustainability: A comparative analysis of 115 leading national newspapers worldwide, *Journal of Corporate Citizenship* (2009) 69–86.
- [3] M. Polignano, N. Bellantuono, F. P. Lagrasta, S. Caputo, P. Pontrandolfo, G. Semeraro, An nlp approach for the analysis of global reporting initiative indexes from corporate sustainability reports, in: *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference, 2022*, pp. 1–8.
- [4] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, *Detectron2*, <https://github.com/facebookresearch/detectron2>, 2019.
- [5] G. Rossiello, P. Basile, G. Semeraro, Centroid-based text summarization through compositionality of word embeddings, in: *Proceedings of the multiling 2017 workshop on summarization and summary*