ReportParse: A Unified NLP Tool for Extracting Document Structure and Semantics of Corporate Sustainability Reporting

Gaku Morio\textsuperscript{1,2}, Soh Young In\textsuperscript{3}, Jungah Yoon\textsuperscript{4}, Harri Rowlands\textsuperscript{5}, Christopher Manning\textsuperscript{1}

\textsuperscript{1}Stanford University, United States
\textsuperscript{2}Hitachi America, United States
\textsuperscript{3}Korea Advanced Institute of Science and Technology, South Korea
\textsuperscript{4}University of Otago, New Zealand
\textsuperscript{5}InfluenceMap, United Kingdom

gaku@stanford.edu, si2131@kaist.ac.kr, isabella.yoon@otago.ac.nz, harri.rowlands@influencemap.org, manning@stanford.edu

Abstract

We introduce ReportParse, a Python-based tool designed to parse corporate sustainability reports. It combines document structure analysis with natural language processing (NLP) models to extract sustainability-related information from the reports. We also provide easy-to-use web and command interfaces. The tool is expected to aid researchers and analysts in evaluating corporate commitment to and management of sustainability efforts.

1 Introduction

As societal awareness of sustainability grows, corporations are increasingly disclosing their sustainability actions, both mandatorily and voluntarily. These disclosures often take the form of annual sustainability reports [Rouen et al., 2022; Bosi et al., 2022] (simply ‘reports’ hereafter), which are textual documents spanning hundreds of pages. Researchers and practitioners analyze these reports to examine corporate commitment to sustainability goals, such as decarbonization and energy transition [Morio and Manning, 2023], and to investigate the potential for ‘greenwashing’ [Kang and Kim, 2022].

The reports often contain complex, unstructured data [In et al., 2019b] and are typically distributed in PDF format. The varying layouts, designs, and disclosed items across companies complicate the automation of information extraction using NLP techniques. Researchers have independently developed methods for report analysis [Li et al., 2022; Kang and Kim, 2022; Gutierrez-Bustamante and Espinosa-Leal, 2022; Polignano et al., 2022]. While Ni et al. [2023] provided a QA-based tool, there is no well-standardized tool for both structure and semantic analysis as far as we know. The lack of a standardized, open tool not only burdens researchers with implementation of a report analysis tool but also leads to reproducibility issues and methodological robustness concerns.

We introduce ReportParse\textsuperscript{1}, a unified tool for parsing both the document structure and semantics of the reports. The concept of extracting document structure and semantics is inspired by a unified scientific paper parsing tool [Lo et al., 2023]. However, our work is tailored for the reports. Here, ‘document structure’ refers to explicit units such as titles, blocks, and sentences, while ‘semantics’ involve annotations on these structure units for the reports. For instance, ReportParse can extract sentences (i.e., part of document structure) tagged as environmental claims (i.e., semantics), thereby simplifying the process of layout analysis, text extraction, and integration of NLP models. We describe each functionality of the tool and demonstrate the effectiveness of ReportParse through discussions.

2 Related Work

A widely used method for extracting document structure from a report involves tools like PyMuPDF [2024] to extract text from PDF files [Kang and Kim, 2022]. For information extraction in sustainability reports, CHATREPORT [Ni et al., 2023] offers a QA-based interface utilizing large language models (LLMs). Tools using LLMs, such as CHATREPORT, offer flexibility in extracting information. However, ReportParse distinguishes itself by providing users with a platform from which they can choose among various document struc-
ture analysis and NLP methods according to their needs.

Technically, our work is most similar to PaperMage [Lo et al., 2023], which offers a tool for analyzing scientific papers through layout analysis and public NLP models. However, our focus is on corporate sustainability reports, which lack a standard format. Additionally, we utilize NLP models that are specifically tailored for the climate change and sustainability domain.

3 System Overview of REPORTPARSE

The system pipeline is detailed in Figure 1. Our system is built upon a Python codebase. For a given PDF report, a ‘reader’ identifies the document structure, while ‘annotators’ use NLP models to assign semantics in relation to the structure. We also integrate command line and web interfaces.

Reader. The reader identifies the document layouts and semantic units. This is similar to ‘Parser’ of PaperMage. We use deepdoctection [2024] to identify elements like titles, text blocks, and lists. This structure, along with associated bounding boxes, is stored in an internal format and fed into the annotators. Text is tokenized into sentences using spaCy [Honnibal et al., 2020]. Users have the option to use PyMuPDF or to integrate their custom reader.

Annotators. The annotators assess the semantics related to the document structure, benefiting from valuable third-party models. The annotator is similar to the ‘predictor’ of PaperMage. Table 1 lists the example third-party annotators. Users can select any annotators suitable for their needs. These annotators can extract various sustainability-related details, such as environmental claims [Stammbach et al., 2023]. Each annotator assigns labels to specific document structures, e.g., an annotator assigns ‘risk’ labels of Bingler et al. [2024] for text blocks. Users can integrate a custom annotator.

Python Command Line Tool. This interface processes an input PDF file and outputs a JSON or CSV file with the analysis results. For instance, the following command employs deepdoctection as the reader and uses the model of Bingler et al. as the annotator to transform ‘filename.pdf’ into a JSON file in the current directory:

```
python -m reportparse.main -i filename.pdf -o ./
--reader "deepdoctection"
--annotators "climate_commitment"
```

Web Interface. We provide a user-friendly web interface created with Gradio [Abid et al., 2019], designed to visualize document structures and associated labels. Figure 2 shows an example of the analysis results. The top left panel illustrates the bounding boxes for both document structures and labels, helping users grasp the analysis process. The top right panels present statistics related to these structures, such as the distribution of annotated labels.

<table>
<thead>
<tr>
<th>Method / Model</th>
<th>Extracted information</th>
<th>Example label and text in sustainability reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Bingler et al., 2024]</td>
<td>Climate commitments / actions</td>
<td>Yes: We use the reduction rate of CO2 emissions per unit ...</td>
</tr>
<tr>
<td>[Stammbach et al., 2023]</td>
<td>Environmental claims</td>
<td>Yes: ... 35 cases of investment in energy-saving equipment ...</td>
</tr>
<tr>
<td>[Deng et al., 2023]</td>
<td>Renewable energy</td>
<td>Yes: ... we will promote wider use of renewable energy through ...</td>
</tr>
<tr>
<td>[Bingler et al., 2024]</td>
<td>Climate sentiments</td>
<td>Risk: As for climate-related business risks, we have followed ...</td>
</tr>
<tr>
<td>[Schimanski et al., 2023]</td>
<td>Net zero or reduction targets</td>
<td>Net-zero: ... an additional goal of realizing carbon neutrality ...</td>
</tr>
<tr>
<td>[Mukherjee, 2020]</td>
<td>ESG-related texts</td>
<td>Air quality: ... control and reduction of chemical substances ... one of the ...</td>
</tr>
<tr>
<td>[DistilBERT-SST2, 2022]</td>
<td>Sentiments</td>
<td>Positive: DEI is also at the core of our sustainability strategy ...</td>
</tr>
</tbody>
</table>

Table 1: Example third-party annotators supported by REPORTPARSE. These are available on Hugging Face Transformers [Wolf et al., 2020].
We present a study on how our tool can be used. Different readers / annotators may need to be considered to increase the robustness of research claims. Thus, REPORTPARSE is useful to improve reproducibility of studies and the robustness of analyses.

## 4 Problem Scenarios

While acknowledging the limitations of the tool, we show small pilot studies and discussions, providing insight into how our tool can be used.

### 4.1 For Analysts

Applying NLP methods to the reports offers significant advantages for practitioners such as analysts. The detailed data analysis afforded by NLP modules could aid analysts in supporting data-driven arguments.

**Pilot Study – Hypothesis Generation.** We present a study that may be useful for analysts in generating hypotheses for sustainability trends. We gathered 2,480 reports from various formats (ESG, sustainability, and responsibility reports) in the energy and basic materials sectors, including major oil and gas companies. Of those, we analyzed the corporate stances on climate change from 2014 to 2020 using REPORTPARSE. We integrate an annotator named ‘climate_policy’ based on the model from [Morio and Manning, 2023], which can predict corporate climate policy engagement for multiple aspects. Using this annotator, we first investigated the average number of pages related to IPCC or UN climate policies, categorizing them as ‘supportive’ or ‘not supportive’ (including no or mixed positions.) Figure 4 (a) indicates a recent trend towards positive stances on IPCC and UN policies. However, the effectiveness of these claims is questionable, as shown in Figure 4 (b), where we did not find a significant positive stance in corporate engagement with energy transition policies (like carbon capture and storage, and transportation decarbonization). While it is not possible to determine whether this case indicates greenwashing trends, it does provide useful insights for hypothesis generation for further analyses.

**Survey Study.** We conduct a survey with an analyst from a non-profit organisation (NPO), who specializes in analyzing corporate climate policy engagement within the reports. We randomly selected four reports of NYSE-listed companies in 2021, sampled from the collected reports. The analyst was asked to complete our survey, which included questions about the usefulness of the web interface and the correctness of the output generated by the annotator models. Table 2 shows a part of our survey results. The analyst confirmed the usefulness of the tool in reducing reading cost of the reports for assessing corporate climate policy engagement. However, the output from the model is not rated as perfect, and the importance of the human analyst still remains. Although the survey results cannot be generalized and the role of human analysts remains crucial, the integration of this tool into the analytical framework of NPOs could improve efficiency.

### 4.2 For Sustainable Finance Researchers

Corporate sustainability has become a pivotal factor in investment decision-making, complementing conventional financial metrics like firm size and growth potential [In et al., 2019a; Bolton and Kapeczyk, 2021]. REPORTPARSE can assess the consistency of a firm’s communications on a specific topic within its sustainability reports over time. However, the interpretability require careful consideration [Sautner et al., 2023]. By addressing these challenges and emphasizing the contributions of our tool, we may improve decision-making processes, enhance transparency, and generate value for stakeholders within sustainable finance ecosystems. Again, REPORTPARSE can be used to investigate the reproducibility and robustness of studies. In future work, we plan to focus on the quantification of corporate sustainability reporting using our tool and on addressing interpretability issues in relation to corporate sustainability and NLP [In et al., 2024].

## 5 Conclusion and Demonstration Scenario

REPORTPARSE facilitates systematic analysis of sustainability reports, promoting open and reproducible research in this field. During the conference session, we will showcase the web interface, allowing users to interact with it. Visitor feedback will inform potential enhancements and the addition of new annotation models to REPORTPARSE.
**Ethical Statement**

We acknowledge that errors in layout analysis and model output from the use of this tool could raise ethical concerns when applied to real applications. For example, if a researcher uses a tool without examining the erroneous output in detail, it will lead to erroneous hypothesis generation and erroneous conclusions. We intend to make the tool available only to analysts and researchers, but its use by investors and general users will lead to incorrect labeling of companies. For example, a particular company might be falsely accused of greenwashing. Conversely, a company might use this tool to unfairly enhance its own reputation. We encourage users to be transparent and to use our tool only as a supplementary tool for humans.

In this study, a simple survey was conducted. The survey did not contain sensitive (personal or harmful) questions.

**Acknowledgments**

We would like to thank Cleo Rank at InfluenceMap, who completed our survey and shared her thoughts. Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used. GM did this research within the Stanford Data Science (SDS) Affiliates Program. GM receives financial supports from Hitachi America to conduct this study. We used OpenAI ChatGPT in part of our paper to translate text, correct grammatical errors, and improve the writing. We declare that the text is original.

**References**


[In et al., 2024] Soh Young In, Gaku Morio, Jungah Yoon, and Christopher D. Manning. When do firms oversell or undersell their environmental sustainability? An empirical analysis of corporate sustainability communications. *Available at SSRN 3264923*, 2024.


