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

Gaku Morio[1,2] , **Soh Young In**[3] , **Jungah Yoon**[4] , **Harri Rowlands**[5] , **Christopher Manning**[1]
[1]Stanford University, United States
[2]Hitachi America, United States
[3]Korea Advanced Institute of Science and Technology, South Korea
[4]University of Otago, New Zealand
[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**[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.

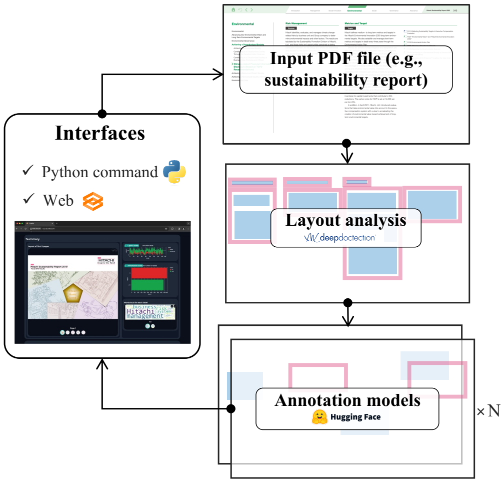


Figure 1: System overview of REPORTPARSE.

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-

[1] Project website: https://github.com/climate-nlp/reportparse

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

Table 1: Example *third-party* annotators supported by REPORTPARSE. These are available on Hugging Face Transformers [Wolf *et al.*, 2020].

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.

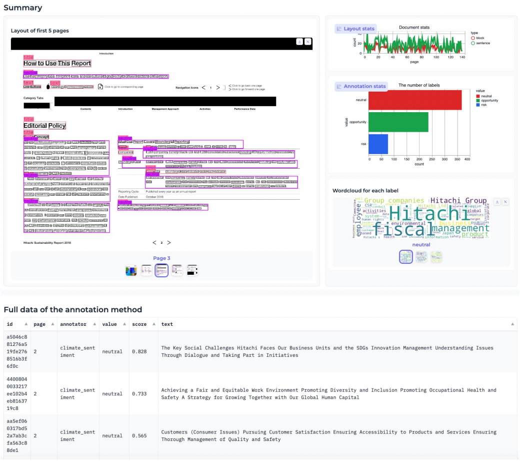
Reader. The reader identifies the document layouts and semantic units. This is similar to 'Parser' of PaperMage. We use deepdoctection [2024] to identify elements like *titles*, *text blocks*, and *lists*. This structure, along with associated bounding boxes, is stored in an internal format and fed into 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.



id	page	annotator	value	score	text
a5046c8 81276a5 19fe276 8516b3f 6f0c	2	climate_sent iment	neutral	0.828	The Key Social Challenges Hitachi Faces Our Business Units and the SDGs Innovation Management Understanding Issues Through Dialogue and Taking Part in Initiatives
4400804 0033217 ee102b4 eb81637 19c8	2	climate_sent iment	neutral	0.733	Achieving a Fair and Equitable Work Environment Promoting Diversity and Inclusion Promoting Occupational Health and Safety A Strategy for Growing Together with Our Global Human Capital
aa5ef06 0317bd5 2a7ab3c fa563c8 8de1	2	climate_sent iment	neutral	0.565	Customers (Consumer Issues) Pursuing Customer Satisfaction Ensuring Accessibility to Products and Services Ensuring Thorough Management of Quality and Safety

Figure 2: The web interface.

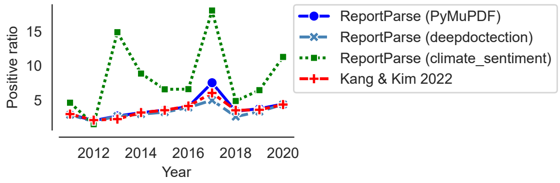
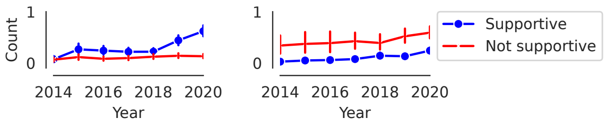


Figure 3: Reproducing experiments of Kang and Kim, 2022 using REPORTPARSE with different readers and annotator.



(a) Stances for IPCC or UN climate policy.

(b) Stances for energy transition policies.

Figure 4: Average label frequency analysis for companies from energy and basic materials sectors during 2014–2020.

Respondent	Usefulness of the tool	The num. of correct outputs by 'climate_policy'	Comments

NPO analyst	4
	(1=min,
	5=max)
		Correct=1
		Partly correct=2
		Incorrect=2
			- This tool would cut down on the time it takes to read through every page, allowing me to do more work and track more evidence of engagement.

Table 2: Survey result of an NPO analyst

System Verification. By replicating an existing study with REPORTPARSE, we verify that the tool works properly. We replicate the work by [Kang and Kim, 2022], where we investigate trends in the sentiment ratio (i.e., num. of positive sentences / num. of negative sentences) from reports for a firm. To follow the setting of the original paper as much as possible, we use PyMuPDF for the reader and DistilBERT-SST2 for the annotator. Additionally, we explored variants by using deepdoctection as the reader or using deepdoctection as the reader and 'climate_sentiment' [Bingler *et al.*, 2024] as the annotator. Figure 3 shows that REPORTPARSE, with PyMuPDF as the reader, reasonably replicates the original study's trends, including the notable 2017's peak, which corresponds to the firm's ESG crises [Kang and Kim, 2022]. It indicates our tool works properly. However, we observe different trends when using deepdoctection as the reader or 'climate_sentiment' as the annotator, suggesting that the choice of reader or annotator methods can significantly impact the analysis. 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 Kacperczyk, 2021]. REPORTPARSE can assess the consistency of a firm's communications on a specific topic within its sustainability reports over time. However, challenges associated with data quality, bias, and 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

[Abid *et al.*, 2019] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ML models in the wild. *arXiv preprint arXiv:1906.02569*, 2019.

[Bingler *et al.*, 2024] Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191, 2024.

[Bolton and Kacperczyk, 2021] Patrick Bolton and Marcin Kacperczyk. Do investors care about carbon risk? *Journal of Financial Economics*, 142(2):517–549, 2021.

[Bosi *et al.*, 2022] Mathew Kevin Bosi, Nelson Lajuni, Avnner Chardles Wellfren, and Thien Sang Lim. Sustainability reporting through environmental, social, and governance: A bibliometric review. *Sustainability*, 14(19), 2022.

[deepdoctection, 2024] deepdoctection. deepdoctection Github repository. https://github.com/deepdoctection/deepdoctection, 2024. Accessed: 2024-05-29.

[Deng *et al.*, 2023] Ming Deng, Markus Leippold, Alexander F Wagner, and Qian Wang. War and policy: Investor expectations on the net-zero transition. In *Swiss Finance Institute Research Paper Series*, number 22–29, 2023.

[DistilBERT-SST2, 2022] DistilBERT-SST2. distilbert-base-uncased-finetuned-sst-2-english (revision bfdd146). https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english, 2022.

[Gutierrez-Bustamante and Espinosa-Leal, 2022] Marcelo Gutierrez-Bustamante and Leonardo Espinosa-Leal. Natural language processing methods for scoring sustainability reports: A study of Nordic listed companies. *Sustainability*, 14(15), 2022.

[Honnibal *et al.*, 2020] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in python, 2020.

[In *et al.*, 2019a] Soh Young In, Ki Young Park, and Ashby Monk. Is 'being green' rewarded in the market?: An empirical investigation of decarbonization and stock returns. *Stanford Global Projects Center Working Paper*, page Available at SSRN 3020304, 2019.

[In *et al.*, 2019b] Soh Young In, Dane Rook, and Ashby Monk. Integrating alternative data (also known as ESG data) in investment decision making. *Global Economic Review*, 48(3):237–260, 2019.

[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.

[Kang and Kim, 2022] Hyewon Kang and Jinho Kim. Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods. *Applied Sciences*, 12(11), 2022.

[Li *et al.*, 2022] Mei Li, Gregory Trencher, and Jusen Asuka. The clean energy claims of BP, Chevron, ExxonMobil and Shell: A mismatch between discourse, actions and investments. *PLOS ONE*, 17(2):1–27, 02 2022.

[Lo *et al.*, 2023] Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore, December 2023. Association for Computational Linguistics.

[Morio and Manning, 2023] Gaku Morio and Christopher D Manning. An NLP benchmark dataset for assessing corporate climate policy engagement. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2023.

[Mukherjee, 2020] Mukut Mukherjee. ESG-BERT: NLP meets sustainable investing. Towards Data Science, 2020. https://towardsdatascience.com/nlp-meets-sustainable-investing-d0542b3c264b.

[Ni *et al.*, 2023] Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore, December 2023. Association for Computational Linguistics.

[Polignano *et al.*, 2022] Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni 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*, pages 1–8, Marseille, France, June 2022. European Language Resources Association.

[PyMuPDF, 2024] PyMuPDF. PyMuPDF Github repository. https://github.com/pymupdf/PyMuPDF, 2024. Accessed: 2024-05-29.

[Rouen *et al.*, 2022] Ethan Rouen, Kunal Sachdeva, and Aaron Yoon. The evolution of ESG reports and the role of voluntary standards. In *Harvard Business School Working Paper*, number 23–024, 2022.

[Sautner *et al.*, 2023] Zacharias Sautner, Laurence Van Lent, Grigory Vilkov, and Ruishen Zhang. Firm-level climate change exposure. *The Journal of Finance*, 78(3):1449–1498, 2023.

[Schimanski *et al.*, 2023] Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. ClimateBERT-NetZero: Detecting and assessing net zero and reduction targets. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15745–15756, Singapore, December 2023. Association for Computational Linguistics.

[Stammbach *et al.*, 2023] Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.