# IMPERIAL

# Bridging Data Gaps: Utilizing NLP to Extract Tables from Sustainability Reports

Ziyi Tang,
Supervisors: Dr Simone Cenci
and Dr Matteo Burato

## Introduction

### Backgrounds

Sustainability reporting communicates a company's environmental, social, and governance (ESG) goals, enhancing transparency and attracting investment. Key data, often presented in tables, is crucial for stakeholders' decision-making and a company's sustainable growth. However, automating table extraction is challenging due to diverse formats, such as multi-column tables and footnotes, which can disrupt content recognition and lead to incomplete results.

### Objectives

This project aims to develop a robust algorithm capable of accurately identifying and extracting tables from sustainability reports, overcoming challenges posed by footnotes, and additional page elements that typically disrupt content recognition to ensure the extracted data is structured and ready for analysis.

## Method

To extract key information, the sustainability reports collected from companies' websites were parsed using different libraries. Key text elements were tokenized to facilitate the analysis of potential tables.

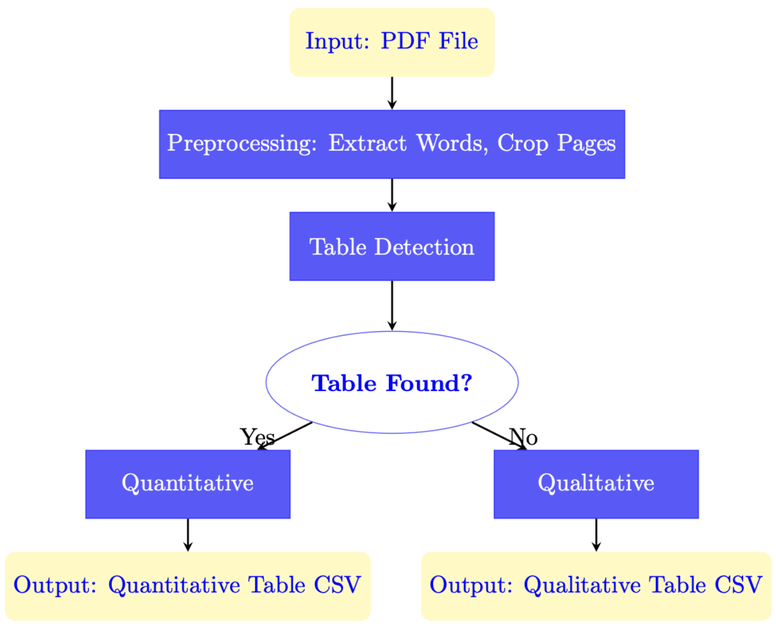


Figure1: The data processing pipeline

The next step involved developing an algorithm to detect tables based on specific criteria. An important part of the process was font size estimation, which helped detect titles, headers, and footnotes.

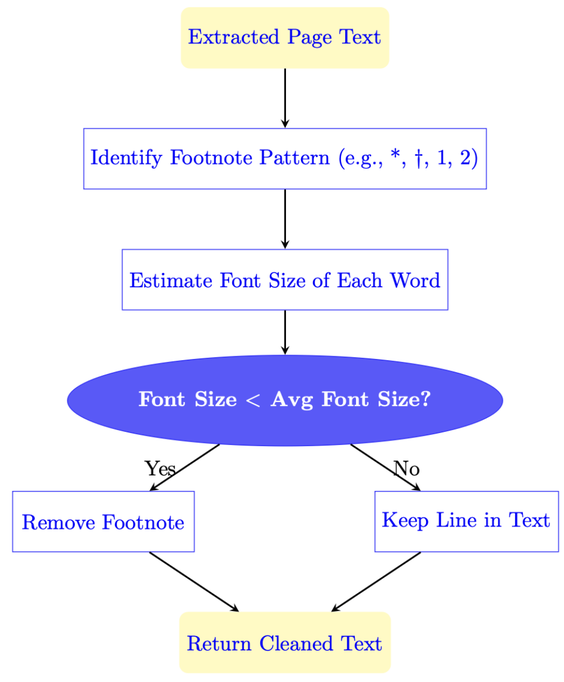


Figure2: The procedure of removing footnotes

Additionally, a footnote removal function was implemented. Footnotes were identified using patterns, such as symbols and numbers, and were filtered out based on their smaller font sizes, ensuring that only the main content was retained. The page was cropped for each table after cleaning unnecessary elements.

Once tables were detected, they were classified as either quantitative or qualitative. Finally, the detected and classified tables were exported.

## Results

The table detection algorithm demonstrates strong overall performance.

**Scope 3 emissions** (in Metric Tons $CO_2e$) (Continued)

| | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|
| **Category 9: Downstream Transportation and Distribution** [5] | - | - | 5,000 | 10,000 | 37 | 16 |
| Of Total (in %) | - | - | <1% | <1% | <1% | <1% |
| **Category 11: Use of Sold Products** [5] | - | - | 5,000 | 390,000 | 106,232 | 62,306 |
| Of Total (in %) | - | - | <1% | 8% | 2% | <1% |
| **Category 12: End-of-Life Treatment of Sold Products** [5] | - | - | <500 | <500 | 1,267 | 3,775 |
| Of Total (in %) | - | - | <1% | <1% | <1% | <1% |

1. Prior to 2021, values were rounded and totals were calculated before rounding throughout this report.

5. In the 2022 reporting year, several updates to reporting were applied to the 2021 and later inventories.
(a) Data from life cycle assessments for our hardware and sold products were used to calculate our Scope 3 emissions.
(b) 2021 Category 1, 2, 8, & 11 emissions were recalculated with higher quality data inputs to improve accuracy.

Figure3: Tables with footnotes from Meta sus reports

While it's still challenging to remove footnotes embedded between tables, the algorithm is capable of detecting the footnote's content and eliminating them from some tables such as tables in Figure 3 (Table 1 is the extracted result).

In terms of detecting the tables, the algorithm achieves an accuracy of 72% and a high recall of 93%. The high recall indicates the algorithm's effectiveness in correctly identifying most

**Scope 3 emissions (in Metric Tons CO2e) (Continued)**

| | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|
| Category 9: Downstream Transportation and Distribution 5 | - | - | 5,000 | 10,000 | 37 | 16 |
| **Of Total (in %)** | - | - | <1% | <1% | <1% | <1% |
| Category 11: Use of Sold Products 5 | - | - | 5,000 | 390,000 | 106,232 | 62,306 |
| **Of Total (in %)** | - | - | <1% | 8% | 2% | <1% |
| Category 12: End-of-Life Treatment of Sold Products 5 | - | - | <500 | <500 | 1,267 | 3,775 |
| **Of Total (in %)** | - | - | <1% | <1% | <1% | <1% |

Table 1: Extracted tables from Meta sus reports

tables in the sustainability reports, which is critical to minimizing missed tables. By ensuring most tables are detected, the algorithm significantly reduces the risk of incomplete data extraction, which could otherwise compromise the quality of analysis.

However, the precision rate of 76% suggests there is room for improvement in minimizing

| | Actual_Positive | Actual_Negative |
|---|---|---|
| **Predicted_Positive** | 272 | 20 |
| **Predicted_Negative** | 86 | 0 |

Table2: Confusion matrix

false positives. The confusion matrix highlights 86 false positives out of 379 cases, meaning that 86 pages were incorrectly flagged as containing a table when, in fact, they did not. Although these errors increase the post-processing workload, they are generally less critical than missing a table, as missed tables directly impact data completeness.

## Conclusion

This project focuses on addressing the challenges posed by diverse table formats and the presence of footnotes to ensure a more efficient analysis of sustainability data. The algorithm's performance was tested on a relatively small sample of 7 companies' sustainability reports. Expanding the dataset to include more reports will provide a better assessment of the generalization. Future work will focus on improving the recognition of various structures and understanding text elements in tables and their relationships to semantic clues inside and outside the table.

## References

Burdick, D. et al. (2020). VLDB Endow.
Qin, J et al. (2024). (eds) Artificial Intelligence in China.
Desai, H. et al. (2021). (eds) Document Analysis and Recognition.
Amran, A. et al. (2014). Strategic Direction.
Doktoralina, C. et al. (2018). Jurnal Akuntansi/Volume XXII