Information Technology
Thesis
Prof. Dr. Josef Fink

Exposé Master Thesis

Design, Implementation, and Evaluation of an Information Extraction Framework for ESG Business Reports

Abbad Zafar

1360066

Supervisors:

Prof. Dr. Josef Fink
Fabian Winkel, MSc.

at

Frankfurt University of Applied Sciences

Faculty 2: Computer Science and Engineering

Information Technology

Table of Contents

1. Motivation 3
2. Research Scope 4
3. Approach 6
    1) Data Collection 6
    2) Data Extraction 7
4. Evaluation 8
5. Experimental Setup 10
6. Literature 11
7. References 11

1. Motivation

In today's world, extracting information from ESG business reports is critical because it allows consumers to align their financial decisions with their morals, discover firms that prioritize sustainability and ethical practices, and manage risks related to environmental, social, and governance aspects. Extracting this information encourages openness, accountability, and stakeholder involvement while also identifying possible economic possibilities and assuring regulatory compliance. We can make educated decisions, generate good change, and contribute to a more sustainable and responsible global economy by leveraging the insights from ESG reports.

The widespread use of portable document format (PDF) documents has sparked interest in research on automated data extraction from these files [1]. With each successful extraction, hidden patterns, significant statistics, and critical trends are revealed. Our drive to perfecting the art of data extraction from PDFs enables us to overcome complicated obstacles, allowing us to innovate and be more efficient. The retrieved data are relevant in a wide range of sectors, including academic research and student and teacher support [1].

Extracting information from ESG business reports for text processing and modelling offers a lot of possibilities and a reward. We may acquire significant insights and quantitative indicators of a company's environmental, social, and governance adheres to by using natural language processing and machine learning approaches to ESG reports. This allows us to effectively analyze vast amounts of unstructured ESG data and identify significant patterns,
and trends analysis. Extracted data may be used to build prediction models, evaluate sustainability performance, detect new risks or opportunities, and guide decision-making processes for investors, regulators, and other stakeholders. Using text processing and modelling on ESG reports allows us to make use of the

richness of information included in these reports, supporting data-driven ways to improve business sustainability, transparency, and responsible decision-making.

2. Research Scope

For this thesis, the research scope aims to develop an innovative and versatile data extraction technique capable of extracting valuable data from various types of reports, such as product reports, business reports, and more. The primary objective is to design a method that can effectively identify and extract relevant information from these reports, thereby facilitating data analysis and decision-making processes.

Our goal is to provide a framework for automated information extraction from ESG (Environmental, Social, and Governance) business reports. It acknowledges the difficulties in manually extracting meaningful information from ESG reports and suggests a methodical way to address this issue. The system will use natural language processing, machine learning, and information retrieval approaches to automatically extract critical information from ESG reports, such as data on environmental impact, social activities, and corporate governance. Designing the framework, defining its aims and scope, implementing, and assessing its performance using relevant metrics, and researching practical uses of the retrieved ESG data will all be part of the research. The study will help to streamline ESG reporting procedures, enable more efficient monitoring of ESG performance, and aid organizations and stakeholders in making informed decisions.

To accomplish this, the research will involve exploring existing data extraction techniques and methodologies commonly used in the field. A comprehensive literature review will be conducted to gain insights into the different approaches employed for data extraction in various domains. This will help in identifying the strengths and limitations of current methods and formulating a novel technique that can overcome existing challenges.

The proposed data extraction technique will be designed to be adaptable and scalable, allowing it to be applied to a wide range of report formats and structures. It will incorporate advanced algorithms and machine learning models to automatically detect and extract relevant data points, such as numerical figures, textual information, and categorical data.

Furthermore, the research will emphasize the development of robust data validation and error-checking mechanisms to ensure the accuracy and reliability of the extracted data. This will involve implementing quality control measures and statistical analyses to verify the integrity of the extracted information.

Comparative studies will be done to evaluate the effectiveness of the proposed method, comparing the performance of the created approach with existing methodologies. Real-world report datasets from various domains will be used to evaluate the technique's capacity to handle various sorts of reports and accurately extract meaningful information.

Our research scope includes:

1- Metadata Extraction, which covers tools for extracting document metadata such as titles, authors, abstracts, and so on.
2- Reference Extraction, which includes tools for accessing and parsing bibliographic reference strings into fields such as author names, publication titles, and venue.
3- Table Extraction refers to technologies that allow you to retrieve both the structure and data of tables.
4- General Extraction, which includes tools for extracting elements such as paragraphs, sections, illustrations, captions, equations, lists, and footers [6].
5- Document Analysis, which includes document layout analysis and document parsing as well as text preparation of the reports.

One noteworthy paper, "A Benchmark of PDF Information Extraction Tools Using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents," by Norman Meuschke, Apurva Jagdale, Timo Spinde, Jelena Mitrovi, and Bela Gipp [6], provides a benchmark on ten freely available tools for extracting document metadata, bibliographic references, tables, and other content elements from academic PDF documents. This study found that GROBID produces the best metadata and reference extraction results. Adobe Extract beats competitors for table extraction, whereas other tools struggle to extract lists, footers, and equations.

In the discipline of document processing and natural language comprehension, document layout analysis, parsing, and text preparation are critical phases. The automatic extraction and comprehension of structural components inside a document, including as paragraphs, headers, tables, and graphics, is referred to as document layout analysis. To analyze the spatial layout and visual properties of the document, techniques from computer vision and image processing are frequently used in this procedure. Once the structure has been decided, the textual information is split into meaningful units, such as phrases and words, through parsing. Parsing is the process of detecting the grammatical structure and links between words in order to extract the intended meaning through syntactic and semantic analysis. At last, text preparation focuses on cleaning and normalizing the extracted text, eliminating noise, fixing mistakes, and putting it into a standardized format for further analysis or processing. These processes are required for a variety of applications, including information retrieval, text summarization, machine translation, and others.

The ultimate goal of this thesis research is to contribute to the field of data extraction by introducing a flexible and efficient technique that can be utilized across various industries and domains. By enabling automated data extraction from reports, organizations will be able to streamline their data analysis processes, make informed decisions, and enhance overall operational efficiency.

3. Approach

1) Data Collection

To develop an effective data extraction technique, obtaining sample reports from diverse companies or products is crucial. These reports will serve as the basis for understanding the variations in report formats, structures, and data types. By analyzing and extracting data from these samples, we can design a technique that is capable of handling the complexities and nuances present in real-world reports, ensuring its applicability and accuracy across different contexts.

Also we would also need for testing and evaluation a pre label dataset which includes the data we need to extract from ESG reports for comparison and evaluation purposes.

2) Data Extraction

- Once we are able to collect reports and datasets, we will start working on different data extraction techniques. We will have to apply different techniques to extract different parts of data in the ESG report pdf files.
- For extracting author, title etc. we would need to extract metadata and get the title, author detail from there.
- For extract table we would need to specify the pdf page we need to extract data from so far, I have used “tabula” library for extraction of tables from pdf.
- For extraction of images data, first we would need to download images and then we can apply OCR techniques to extract text from pdf.
- Next task would be to extraction of header, footers, and paragraphs.

Text extraction accuracy from OCR (Optical Character Recognition) may be considerably enhanced by leveraging supporting techniques such as Tesseract and adding new algorithms such as a custom block recognition algorithm.

Tesseract is a popular open-source OCR engine with exceptional text recognition capabilities. It extracts text from photographs or scanned documents using several approaches such as character segmentation, feature extraction, and language modelling. Tesseract's strength is its ability to work with a wide range of typefaces, languages, and complicated layouts. Accuracy may be enhanced in instances where text extraction is difficult owing to various typefaces or languages by employing Tesseract's excellent OCR capabilities.

In addition to employing OCR engines, implementing unique algorithms can improve accuracy even more. Implementing a unique block detection algorithm is one such method. The layout of a document might vary greatly, with distinct sections, columns, or headings. A bespoke block identification method may analyze the visual structure of the document, identify discrete blocks or areas, and then apply OCR to each block individually. This method enables targeted OCR processing on specified regions, reducing mistakes caused by overlapping or complicated layouts. Text extraction accuracy may be considerably enhanced by segmenting the document into relevant parts and performing OCR independently to each block.

4. Evaluation

After implementing the data extraction techniques, it is crucial to evaluate the accuracy and correctness of the extracted data. This evaluation ensures that the chosen extraction method is reliable and effectively captures the desired information. By comparing the extracted data with ground truth or manually validated data, we can assess the precision, recall, and overall performance of the extraction technique.

The process of evaluation begins with a simple ESG report file in pdf format, on which we use extraction techniques based on our needs, such as extracting references, table data, and so on. In addition, we have a labelled dataset of reports/papers. Depending on our goal, we can use datasets offered by DocBank [7] or the OSDG [8] Community. After extracting information with our extraction approach, we compare the results with the datasets and assess our extraction technique/tool using evaluation metrics such as Precision, Recall, Accuracy, and F1 Score. Based on these data, we can establish whether or not the information was successfully extracted.

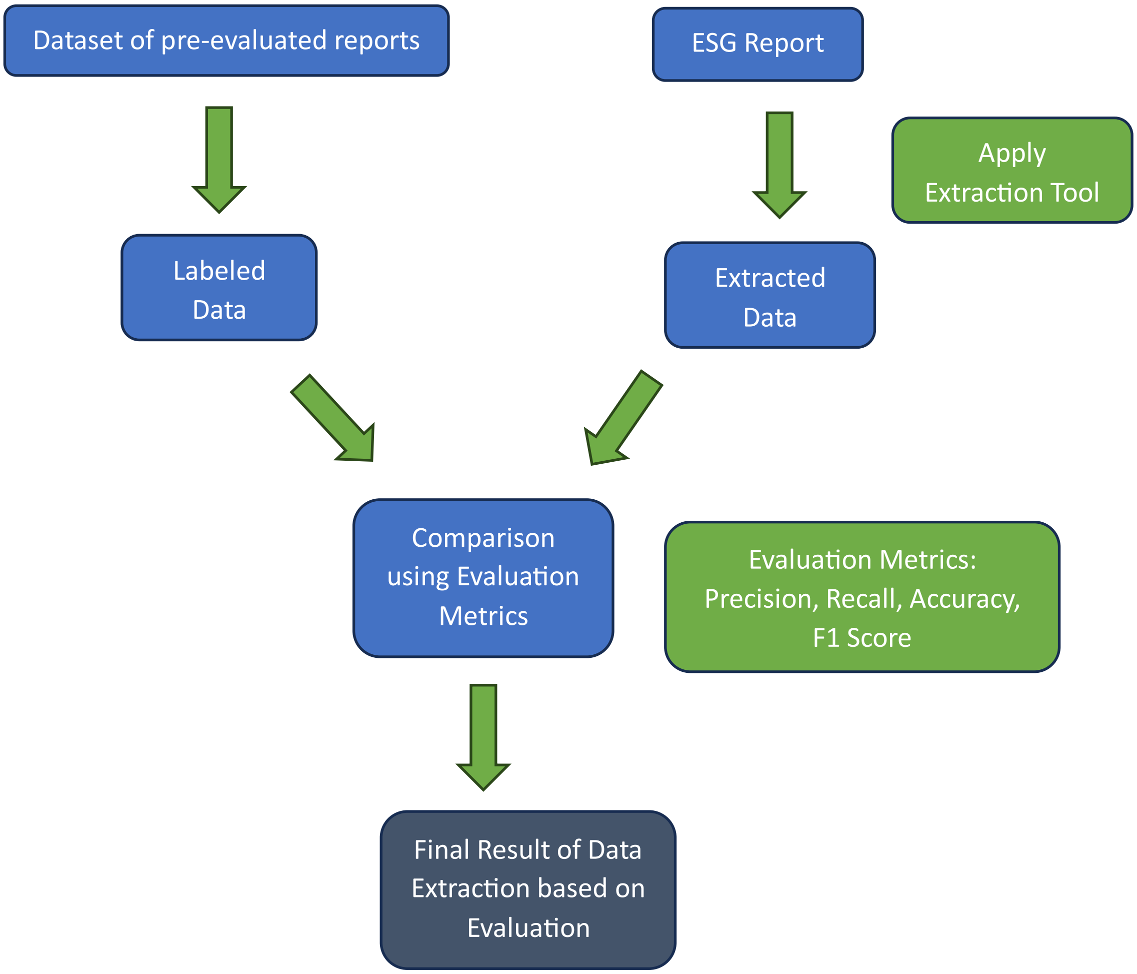
Dataset of pre-evaluated reports
ESG Report
Apply Extraction Tool
Labeled Data
Extracted Data
Comparison using Evaluation Metrics
Evaluation Metrics: Precision, Recall, Accuracy, F1 Score
Final Result of Data Extraction based on Evaluation

Fig. 1: Flowchart of Evaluation Process

Once the correctness of the extracted data is confirmed, it can be utilized for further analysis or predictions. Machine Learning algorithms can be applied to extracted data to uncover patterns, relationships, and insights. These algorithms can perform tasks such as classification, regression, clustering, or predictive modeling, depending on the specific objectives of the research or application.

5. Experimental Setup

In this section, I will be discussing the project setup frameworks and library used:

1- All code will be implemented using Python classes. The aim is to provide a module-like utility that can be invoked via a terminal to extract and pre-process the content and metadata stored in ESG business reports.

2- In this research we will be evaluating various (e.g., Python) libraries, such as:

- GROBID
- PyPDF2
- Adobe Extract
- Apache Tika
- Textract
- PyMuPDF
- PDFtotext (Command Line based)
- Tabula
- Camelot
- CERMINE
- RefExtract
- Pdfminer.six
- optional: Cloud based pdf extraction tool.

3- ElasticSearch must preserve all data (including all experimental stages and results). Elasticsearch is an open-source search and analytics engine that is free to use. It is built on the Apache Lucene library. It may be used to search for any type of data. It offers a scalable search solution, near real-time search, and multi-tenancy support. Elasticsearch collects unstructured data from many sources, stores and indexes it using user-specified mapping (which may potentially be derived automatically from data), and makes it searchable.

6. Literature

Data extraction from PDF reports has been extensively studied in the literature. Notable research papers include "A Study on Information Extraction from PDF Files " by Yuan and Fang [3] which proposes techniques for extracting information from PDF reports. Another significant paper is " A Benchmark and Evaluation for Text Extraction from PDF" by Bast, Hannah & Korzen, [5] providing a comprehensive survey of different techniques. Additionally, "PDFX: Fully-automated PDF-to-XML Conversion of Scientific Literature" by Alexandru Constantin, Steve Pettifer, and Andrei Voronkov [4] presents an automated approach for converting PDFs to XML.

7. References

[1] Wiechork, Karina & Charão, Andrea. (2021). Automated Data Extraction from PDF Documents: Application to Large Sets of Educational Tests. 10.5220/0010524503590366.

[2] Bast, H. and Korzen, C. (2017). A Benchmark andEvaluation for Text Extraction from PDF. In 2017ACM/IEEE Joint Conference on Digital Libraries(JCDL), pages 1–10.

[3] Yuan, Fang et al. “A Study on Information Extraction from PDF Files.” International Conference on Machine Learning and Computing (2005). 10.1109/ACCESS.2021.3107975.

[4] Alexandru Constantin, Steve Pettifer, and Andrei Voronkov. 2013. PDFX: fully-automated PDF-to-XML conversion of scientific literature. In Proceedings of the 2013 ACM symposium on Document engineering (DocEng '13). Association for Computing Machinery, New York, NY, USA, 177–180. https://doi.org/10.1145/2494266.2494271

[5] Bast, Hannah & Korzen, Claudius. (2017). A Benchmark and Evaluation for Text Extraction from PDF. 1-10. 10.1109/JCDL.2017.7991564.

[6] Meuschke, N., Jagdale, A., Spinde, T., Mitrović, J., Gipp, B. (2023). A Benchmark of PDF Information Extraction Tools Using a Multi-task and Multi-domain Evaluation Framework for Academic Documents. In: , et al. Information for a Better World: Normality, Virtuality, Physicality, Inclusivity. iConference 2023. Lecture Notes in Computer Science, vol 13972. Springer, Cham. https://doi.org/10.1007/978-3-031-28032-0_31

[7] Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., & Zhou, M. (2020). DocBank: A benchmark dataset for document layout analysis. arXiv preprint arXiv:2006.01038.

[8] https://osdg.ai/