International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 07 | July 2022 www.irjet.net p-ISSN: 2395-0072

Development of Information Extraction for Data Analysis using NLP

Geetha K S[1], Yashwanth G[2], Tanisha Jain[3]

[1]Professor and Vice Principal, RV College of Engineering
[1]Student, Dept. of Electronics and Communication Engineering, RV College of Engineering
[3]Student, Dept. of Electronics and Electrical Engineering, RV College of Engineering

---

Abstract - *Information Extraction from PDFs for analysis is a common sight in the corporate world. The manual work done by the analysts consumes time depending on the size of the annual reports they are referring to. It also hinders the scalability of the process. Therefore, automation of data analysis for the analysis of PDFs is a necessity today. Hence this paper provides an algorithm by which information can be extracted from the PDFs and mapped to various categories of interest. The categories of interest can be varied, depending on the requirements by the user. The text extraction can be done using simple modules like PDFMiner. However, the dictionary creation has to be done for the sentences to be mapped to particular topics. Using rule-based filters will help extract the required sentences without much consumption of memory and can be understood very easily compared to complex procedures in the algorithm. The proposed algorithm simplifies the entire process of information extraction by providing a broad framework inside the algorithm that can be further modified based on the interests of the user.*

Key Words: **Data Analysis, NLP, Data Embedding, Text extraction, Table extraction**

1. INTRODUCTION

Information Extraction (IE) is the method of parsing through the unstructured data and deducing required information into editable and readable formats of the data. We usually search for some required data when the context is in digital format or manually check the same. IE tools make this possible to pull the required information present in text documents, database, websites or multiple sources. Using IE in Natural Language Processing (NLP) algorithms, we can automate the extraction of data with all required information such as tables, company growth metrics and other financial details from various kinds of documents, vis-à-vis PDFs, Docs, Images, and so on. Convolutional Neural Network (CNN) are already common in computer vision models to process and derive the relations in multi-dimensional data. Therefore, NLP models have already been combined with computer vision models in the past, to benefit from positional information and to improve performance of these key information extraction models.

A document contains information in various forms and the useful information can be present in any of the forms. Hence, the tool built to extract the information from all the various forms. The information in the document present in the form of text and is represented in a presentable format successfully using NLP as well as word-embedding. Therefore, the steps involved in the project include Keyword analysis, Information extraction from text and tables and UI Development with feedback mechanism. The main objectives of the project include to enable intelligent keyword search for data present in text format using pos tagging and word embedding, to extract data from the text and tables by building NLP algorithms and finally combining all of the data extracted and presenting in the form of a table.

2. LITERATURE REVIEW

T. Hassan and R. Baumgartner [1] provide a unique approach for the text extraction by combining the top-down approach as well as the existing bottom-up approach by segmenting a page in a PDF and later converting the text into Hyper Text Markup Language (HTML) and presenting the extracted data to the user. This would also mean that structured data inside the PDF into semi-structured formats. An automatic PDF extractor is proposed by Reza M. Parizi et al. [2] to extract health parameters in the report present in a PDF. It features language compatibility, batch processing, ease of use and an open-source tool as parameters for efficient text extraction in the required format. Ying Liu et al. [3] describe an algorithm to extract metadata from a table that would help in the extractions of tabular data from a file. Metadata extracted in the algorithm includes page number, position, column number and number of rows. It is capable of extracting texts, numbers, symbols and images.

Xiaonan Lu et al. [4] proposes an algorithm to extract data from 2-dimensional plots for the line graphs. It uses the concepts of line segmentation, denoising, PCC coding at pixel level. The identification of curves is necessary for connectivity between two segments. The intersection between two segments is identified based on whether the intersection is M-type, L-type or R-type. The squared mean error is the mathematical parameter used in the extraction process. The method limits the identification of graphs that are not line graphs. Another limitation is that squared mean may not be the suitable mathematical parameter that can give accurate prediction of presence of the line. Karina Weichork and Andrea Charao [5] use the methods of PDFMiner and CyberPDF for the extraction of texts and

later use other methods for looking into interest regions. The PDF is first extracted into XML format and then a script is written to extract the XML files to the interest regions. The various literature studies thus suggest that a good approach to extract information is to develop an algorithm that is rule-based algorithm with feedback loops in the system.

3. Design Methodology

The design methodology followed in the algorithm proposed is as shown in Figure 1.

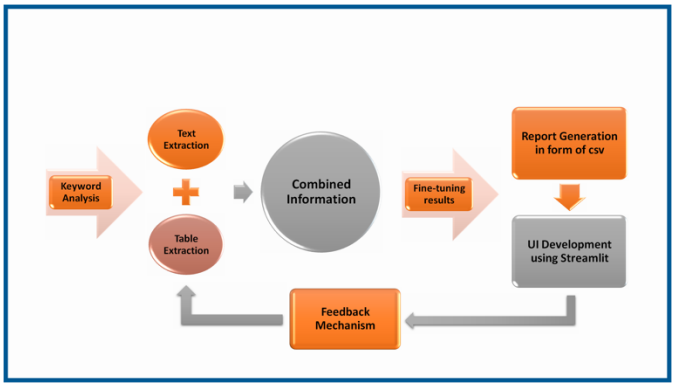


Fig -1: Design Methodology

In initial stages, the thorough understanding of data is done from analyst point which followed by keyword analysis. In the next stage in-depth literature survey is carried out to understand the existing works carried out with regards to information extraction tool for a specific criterion. Feasibility of the intended work is also brainstormed to ensure completion of project in prescribed period. This is followed by problem statement definition based on previous work and present need. Using NLP techniques, algorithms are been built to extract the information from text and table that is useful to the analysts. Based on analysts review the algorithms are been refined and final design cycle is initiated. With outlines of design of the tool, the development of tool takes place alongside simultaneous testing of the algorithm. In view to carry forward the project, a feedback mechanism is built which captures the user inputs to match the complete process efficient and automated in the near future.

4. Implementation

The implementation of the design methodology mentioned earlier is implemented in the following stages:

4.1. Keyword Analysis

Keyword analysis is the first step in the design of the project. It is a manual process in which the keywords and synonyms of each category are identified as per the user requirement. The keywords here, mean the words/data that if present contains some relevant information in and around them in the paragraph or sentences. It is a process of creating a dictionary.

For example, Number of employees FTEs: Number of employees FTEs, Number of employees, Employees, FTEs, workforce, total workforce, intergeneration

Gender Diversity: Gender Diversity, Gender Distribution, share of women, share of men, gender-female, female/total, women in management position, share of women in management, women in leadership, women, men, female, male, etc.

4.2. Design for Text Extraction

The flowchart represented in Figure 2 describes the design algorithm for text extraction.

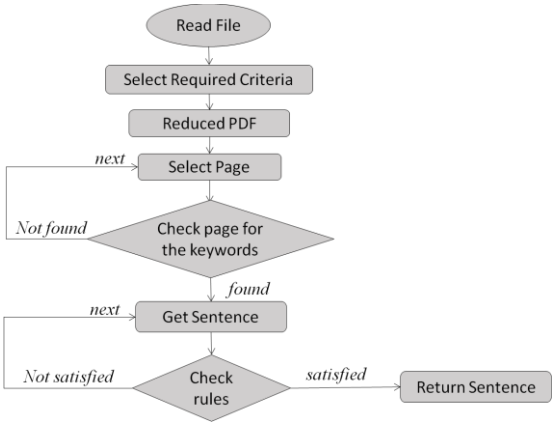


Fig -2: Algorithm for Text Extraction

The text extraction is a logic that this algorithm uses is largely rule-based systems designed by the analysts. Some of the rules followed will be as follows:

- The extracted information must have a numerical value in it for analysis related to the keyword or category of information we are looking for.
- The numerical value can be represented as numbers, digits or in words.
- The information extracted should be present or future tense.
- The past and future related information should be represented separately.
- The extraction of only a number without context is of no use and is bound to be discarded.

4.3. Design for Table Extraction

The extraction of information present as a paragraph and the information present inside a table is different.

When the data is read from a file it considers the text and tables differently. Figure 3 shows the steps for the extraction of relevant data from the table. The first step involved in any extraction is reading the file, so is the case here. All the tables are identified in the file and are numbered. Then filtering is performed on these tables based on the category selected by the user. All the information from the tables related to the category are taken and according to the rules stated above, the information is extracted.

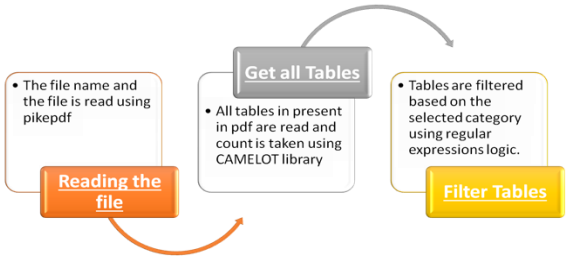


Fig -3: Algorithm for Table Extraction

4.4. Generation of the Report

A report is generated containing all the combined information that was extracted from both the textual and tabular formats in the document. The information is combined together using pandas library in python. Finally, only one csv file is generated and presented as report of that document for the category selected by the user. The conversion of a data frame to csv is carried out using python programming language.

4.5. Development of Feedback Mechanism

In the case of information extraction, feedback mechanism refers to the inputs given by the users or actions performed by them on the extraction tool. This contains 2 steps:

- Highlighting the document and
- Annotate the information

The annotation mechanism is built by representing the extracted information from text as well as tables in a consolidated format. Each record or sentence generated after the extraction process is given an annotation. These annotations are displayed to the users with the information they can select. In this way, the information relevant to them is captured and stored in the backend. This stored information will then be used for feedback loop and adjusting mechanism.

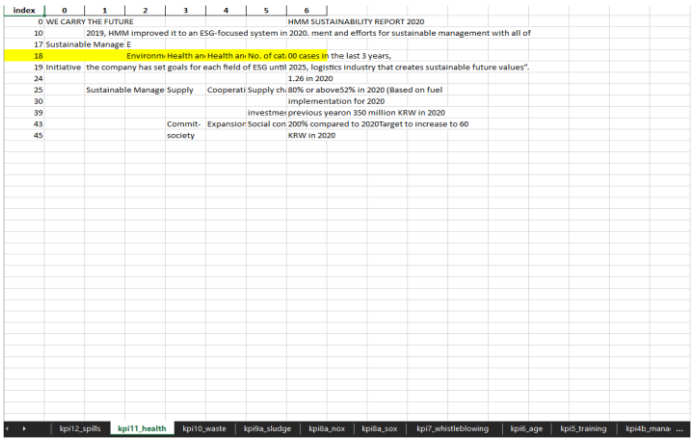
5. Results

The text extracted from the PDFs is represented in a format in Excel file that is as shown below in Figure 4.

sentences filtered by keywords	tense	filtered by numbers
We're also a part of the RE100 business initiative, and aim for all of our entities' electricity needs to be met from renewable sources by 2	{'future': 0, 'present': 2, 'past': 1}	TRUE
By 2025 we aim to reduce emissions for selected investment classes in our portfolio of clients' money by 25 percent compared to 2019.	{'future': 0, 'present': 1, 'past': 2}	TRUE
To deliver on our strategic objectives, we set clear annual targets:2021 target Compound Annual Growth Rate (CAGR) 2019–21 5 %> 13 %	{'future': 0, 'present': 3, 'past': 0}	TRUE
WhyWe secure your futureWhat Outperform Transform RebalanceHow Renewal Agenda01Introduction01.1 Message from the CEO01.2	{'future': 0, 'present': 13, 'past': 12}	TRUE
This target was announced in early 2021.	{'future': 0, 'present': 0, 'past': 2}	TRUE
Our long-term commitment of achieving net-zero GHG emissions in our proprietary investment portfolio by 2050 is now being deliverec	{'future': 0, 'present': 4, 'past': 2}	TRUE
All members1 have now carried out and disclosed portfolio baseline assessments and have defined an intermediary five-year target for	{'future': 0, 'present': 2, 'past': 3}	TRUE
Sustainability Report 2020●03.2 Sustainability in proprietary investmentsIn line with these commitments, we have announced our first	{'future': 0, 'present': 4, 'past': 1}	TRUE
In addition to traditional investment criteria, equities and corporate bonds will be reviewed for their 1.5-degree pathway compatibility	{'future': 1, 'present': 0, 'past': 1}	TRUE
By 2025, we aim for at least 50 percent of our AuM in Oil and Gas Industry to have set net-zero 2050 targets.	{'future': 0, 'present': 1, 'past': 1}	TRUE
Furthermore, we aim to increase our bilateral engagement activities by at least 100 percent by 2025.	{'future': 0, 'present': 1, 'past': 0}	TRUE
Allianz Real Estate is working to reduce the GHG emissions of our portfolio to net-zero by 2050 by embedding ESG criteria and collaborat	{'future': 0, 'present': 4, 'past': 0}	TRUE
This will significantly enhance our overall sustainable investment, offering: • Substantial conversion of traditional strategies into sustai	{'future': 1, 'present': 2, 'past': 4}	TRUE
This specific issuer announced in September that they are revising their climate goals and now aim to reduce by 60 percent (versus 50 pe	{'future': 0, 'present': 4, 'past': 1}	TRUE
Our target is to reach an IMIX score of 73 percent by 2021.	{'future': 0, 'present': 1, 'past': 0}	TRUE
The aim is to increase to 80 percent by the end of 2021 and to address the identified learning gaps through our Learning & Development	{'future': 0, 'present': 1, 'past': 0}	TRUE
01Introduction02Sustainability strategy and governance03Sustainability in our core business activities Diversity and InclusionTraining ar	{'future': 0, 'present': 5, 'past': 1}	TRUE
This surpasses our WWI+ ambition for 2021, which has been set to 66 percent in 2019, indicating that our employees rate the company's	{'future': 0, 'present': 3, 'past': 2}	TRUE
Our target is for over 75 percent of Allianz Group business segments to score above market or at Loyalty Leader position and 50 percent	{'future': 0, 'present': 1, 'past': 0}	TRUE
01Introduction02Sustainability strategy and governance03Sustainability in our core business activitiesTraining and developing our peopl	{'future': 0, 'present': 2, 'past': 5}	TRUE
Our new GHG emission target is a 30 percent GHG reduction per employee by 2025 (baseline year 2019).	{'future': 0, 'present': 1, 'past': 0}	TRUE
We have set a number of targets to achieve this commitment (baseline year 2019, target year 2025 unless stated otherwise):30 % GHG e	{'future': 0, 'present': 4, 'past': 4}	TRUE
It consists of the world's largest pension funds and insurers who are committed to reduce GHG emissions of their investment portfolios	{'future': 0, 'present': 2, 'past': 1}	TRUE
Members of the AOA commit to reduce GHG emissions of their proprietary investment portfolios to net-zero by 2050.	{'future': 0, 'present': 0, 'past': 0}	TRUE
As part of the long-term commitment to have net-zero greenhouse gas emissions by 2050, we have set our first intermediate target: We	{'future': 1, 'present': 1, 'past': 3}	TRUE
Having over-achieved our previous targets by 12 percentage points (target: 50 percent reduction in emissions by 2020 vs. 2010) we have	{'future': 0, 'present': 3, 'past': 1}	TRUE
in operationalizing our commitment through the RE100 initiative to source 100 percent renewable power for our group-wide operations	{'future': 0, 'present': 2, 'past': 1}	TRUE
01Introduction02Sustainability strategy and governance03Sustainability in our core business activities04Sustainability in our organizatio	{'future': 0, 'present': 0, 'past': 3}	TRUE
We are also advocating for:• embedding of 'net-zero by 2050' in short- to long-term governmental NDCs, climate strategies and emissio	{'future': 1, 'present': 5, 'past': 0}	TRUE
01Introduction02Sustainability strategy and governance03Sustainability in our core business activities04Sustainability in our organizatio	{'future': 0, 'present': 1, 'past': 4}	TRUE
In 2021, we aim to develop comprehensive quantitative scenario analysis on physical, transition and litigation aspects of climate change,	{'future': 0, 'present': 2, 'past': 0}	TRUE
01Introduction02Sustainability strategy and governance03Sustainability in our core business activities04Sustainability in our organizatio	{'future': 0, 'present': 6, 'past': 13}	TRUE

Fig -4: Results of text extraction

The extraction of tables carried out is extracted onto an Excel file in CSV format as shown in Figure 5.



index	0	1	2	3	4	5	6
0	WE CARRY THE FUTURE						HMM SUSTAINABILITY REPORT 2020
10	2019, HMM improved it to an ESG-focused system in 2020. ment and efforts for sustainable management with all of						
17	Sustainable Manage E						
18			Environm	Health an	Health an	No. of cat	00 cases in the last 3 years,
19	Initiative	the company has set goals for each field of ESG until 2025, logistics industry that creates sustainable future values".					
24							1.26 in 2020
25		Sustainable Manage	Supply	Cooperati	Supply ch	80% or above	52% in 2020 (Based on fuel
30							implementation for 2020
39					investmer	previous yearon 350 million KRW in 2020	
43				Commit-	Expansior	Social con	200% compared to 2020Target to increase to 60
45				society			KRW in 2020

Fig -5: Results of table extraction

The information is highlighted and represented according to the implementation design explained earlier. The highlighted parts of the text is as shown in Figure 6.

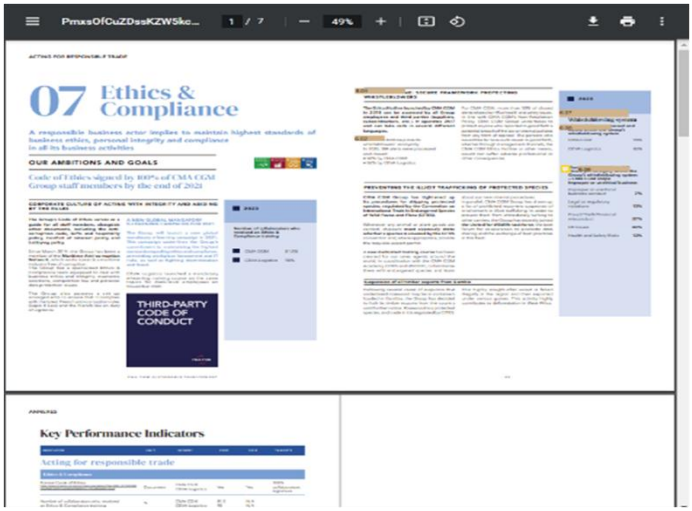


Fig -6: The highlighted parts of text extracted in the PDF

6. CONCLUSION

In this project carried out, the needs of the analysts' specific to their purpose at the bank were taken into account. The algorithm for extraction of textual and tabular format data was built separately and on satisfactory results from them they were combined to make them run in parallel saving computational power and time for processing. The extraction algorithm was built using the combination of NLP, pos tagging and word embedding techniques with a set of predefined rules that the data extracted should satisfy. The implementation of this tool saved weeks of time required by the document analysis team to go through each and every document and make a report ready for the analysts to use. The tool was able to the task in few minutes thus saving a lot more time and making the work faster and more efficient.

REFERENCES

[1] T. Hassan and R. Baumgartner, "Intelligent text extraction from pdf documents," in International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, and Internet Commerce (CIMCA-IAWTIC'06), vol. 2, 2005, pp. 2–6. doi: 10.1109/CIMCA.2005.1631436

[2] R. M. Parizi, L. Guo, Y. Bian, A. Azmoodeh, A. Dehghantanha, and K.-K. R. Choo, "Cyberpdf: Smart and secure coordinate-based automated health pdf data batch extraction," in 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2018, pp.106–111. doi:10.1145/3278576.3281274

[3] K. Bai, P. Mitra, C. L. Giles, and Y. Liu, "Automatic extraction of table metadata from digital documents," in Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06), 2006, pp. 339–340. doi: 10.1145/1141753.1141835.

[4] X. Lu, J. Wang, P. Mitra, and C. Giles, "Automatic extraction of data from 2-d plots in documents," in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 1, 2007, pp. 188–192. doi:10.1109/ICDAR.2007.4378701.

[5] K. Wiechork and A. Charao, "Automated data extraction from pdf documents: Application to large sets of educational tests," May 2021, pp. 01–04. doi: 10.5220/0010524503590366.

[6] G. D. F. Duy Duc An Bui and S. Jonnalagadda, "Pdf text classification to leverage information extraction from publication reports," Journal of Biomedical Informatics, vol. 61, pp. 141–148, 2016, issn: 1532-0464. doi: 10.1016/j.jbi.2016.03.026.

[7] P. S. Dominika Tkaczyk and M. Fedoryszak, "Automatic extraction of structured metadata from scientific literature," International Journal on Document Analysis and Recognition (IJDAR), vol. 18, pp. 317–335, Dec. 2015. doi: 10.1007/s10032-015-0249-8.51

[8] M. Hansen, A. Pomp, K. Erki, and T. Meisen, "Data-driven recognition and extraction of pdf document elements," Technologies, vol. 7, p. 65, Sep. 2019. doi: 10.3390/technologies7030065.

[9] M. Tedre, H. Vartiainen, J. Kahila, T. Toivonen, I. Jormanainen, and T. Valtonen, "Machine learning introduces new perspectives to data agency in k—12 computing education," in 2020 IEEE Frontiers in Education Conference (FIE), 2020, pp. 1–8. doi: 10.1109/FIE44824.2020.9274138.

[10] A. Ehrhardt and M. T. Nguyen, "Automated esg report analysis by joint entity and relation extraction," Springer International Publishing, 2021, pp. 325–340, isbn: 978-3-030-93733-1. doi: 10.1007/978-3-030-93733-1_23.

[11] V. Armenise, "Continuous delivery with jenkins: Jenkins solutions to implement continuous delivery," in 2015 IEEE/ACM 3rd International Workshop on Release Engineering, 2015, pp. 24–27. doi: 10.1109/RELENG.2015.19.

[12] S. Haines, Modern Data Engineering with Apache Spark. Apress Berkeley, CA, 2022, isbn: 978-1-4842-7451-4. doi: 10.1007/978-1-4842-7452-1.