

# **PROJECT REPORT**

Project title

## **Contextual Text Extraction from PDF files using Machine Learning Practices**

**Industrial Project Based Learning**

**Capstone Project**

By

**TEAM 7**

### **Team Members:**

CHINTA SAI RANGANATH	22R15A6703
SIDRAMYNA RATHNAKAR	22R15A6707
MARIPALLI FARIDUDDIN	21R11A6730
LUNAVATH THIRUPATHI	22R15A6705
NANDIGAMA LIKITH CHOWDARY	21R11A6739
KOTA AMITH	21R11A6724

**Geethanjali College of Engineering and Technology**

**Cheeryal, Keesara, 500085, Telangana**

**APRIL 2024**

## ABSTRACT

Sometimes, we need to take information out of PDFs for things like making them easier to read or using them in legal or research work. Using Machine Learning (ML) can be a good way to do this because it's good at understanding and taking out text from PDFs, even if they're complicated. But it can still be hard to get specific information from PDFs, like finding the names of directors in company documents. This project tries to solve that problem by using ML to automatically find and take out important details like directors' names, whether they're independent or executive, and their ID numbers. This will make it easier and faster to get the right information from PDFs without needing to do it all manually.

## TABLE OF CONTENTS

S.NO	TOPICS	PAGE NUMBERS
i	Abstract	i
1	Introduction	2
2	Literature survey	3
3	Problem statement	3
4	Objectives	3
5	Methodology	4
5.1	Data Collection	4
5.2	Importing PDF's and required packages	4
5.3	Text Pre-Processing	4
5.3.1	Extracting Text	4
5.3.2	Forming Sentences	5
5.3.3	Finding Director Sentence	5
5.3.4	Finding Director Names	5
5.3.5	Blacklist	5
5.3.6	Finding DIN and Director Type	5
6	Implementation	6
6.1	Building web Application	6
7	Conclusion	7
8	Future Scope	7
9	References	7

## LIST OF FIGURES

Fig No	Text	Page No
6.1	Result page 1	6
6.2	Result page 2	6

# 1. INTRODUCTION

Extracting contextual information from PDFs can be a challenging task, but it's something many jobs need to do. Using modern technology tools like called Machine Learning (ML) can help make it easier. ML can understand how PDFs are made and pick out the important bits of text. In this project, we're using ML to figure out who the directors of companies are from their PDF documents. This saves time and stops mistakes compared to doing it all by hand.

Taking text out of PDFs has been a challenge, and people have tried different ways to do it. Some use strict rules, some use special tools like Optical Character Recognition (OCR), and some use ML. ML is quite better because it can learn from lots of different documents and get better at finding the right information. It's been used successfully to find things like money details in bank statements or important info in legal papers. But there are still problems, like Sthe problem of finding directors' details in PDF.

## 2. LITERATURE SURVEY

- **Extracting text from PDFs has been a challenge addressed through various methods.**

There are various methods to extract the text from pdf files. But the most efficient way to extract the pdf files by using python libraries like pdfplumber and PyPDF2. By using these libraries, we can read the pdf files and extract the text from it.

- **Using various method to find the names of the person.**

There are various ways to find the person's name by using regular expressions and NLP. We used Natural Language Processing which is most suitable for this project.

## 3. PROBLEM STATEMENT

The challenge we're facing is finding important information like the names of company directors, their roles (independent or executive), and their DIN numbers from PDF files. Doing this by hand is slow and prone to errors. PDFs can be messy, making it difficult for computers to understand them properly. Our goal is to use modern tools like Machine Learning to solve this problem, making the process faster, easier, and more accurate.

## 4. OBJECTIVES

1. PDF Data Collection: Collecting various types of pdf files.
2. Importing PDF's and Required Packages: Importing necessary Python libraries such as PyPDF2, SpaCy, NLTK, Re for importing the pdf files, reading the pdf file, extracting the text.
3. Text Pre-Processing: Filtering the text using various methods.
4. Implementation: Implementing the task from theory to practical.
5. Building Web Application: Developing a web application using HTML, CSS, and Flask framework for interaction with the user and display the results to user.

## 5. METHODOLOGY

### 5.1 Data Collection

- **Description of data**

We collected PDF's for Extracting the useful information from it.

- **Brief description of the data**

Here, first we collected finance related PDF's and we extracted the text based on requirements by using various libraries and packages of python.

### 5.2 Importing PDF's and required packages

- **PyPDF2:** PyPDF2 is a Python library for importing, reading, writing, and manipulating PDF files.
- **NLTK:** NLTK (Natural Language Toolkit) is a Python library for natural language processing tasks like tokenization, stemming, tagging, parsing.
- **SpaCy:** SpaCy is a powerful open-source natural language processing library for Python, offering efficient tokenization, parsing, named entity recognition.
- **Re:** Regular expressions (regex or regex) are sequences of characters defining a search pattern, used mainly for pattern matching within strings.

### 5.3 Text Pre-processing

- **5.3.1 Extracting Text**

First, we imported necessary libraries as mentioned above. Read the pdf file by using PyPDF2 Pdf-reader. Then we tried to extract the text from the pdf. We iterated through each page and extracted text from each page excluding tables.

➤ **5.3.2 Forming Sentences**

We find the tables in the pdf and we extracted the text cell by cell as a sentence. Then we processed the regular text into sentences. And concatenated both the texts.

➤ **5.3.3 Finding Director Sentence**

We search the director word in each sentence of the extracted text if a sentence found containing the word “Director” we created an array called `director_sentences` and stored all the sentences which contains the word “Director”.

➤ **5.3.4 Finding Director Names**

We used Natural Language Processing to identify the names in the sentences. We selected the English as a language of our choice to find the names, and we found the names of the directors as well as other names like places etc, Overall, we found the nouns in each sentence and we extracted them.

➤ **5.3.5 Blacklist**

We found some unwanted entities and also duplicate names we used a blacklist array to remove unwanted entities and dictionary set to remove the duplicates. Now we have the Director names of the respected company.

➤ **5.3.6 Finding DIN and Director Type**

Now, we have director names by using these names we find the sentence with the director’s name. Once found the sentence, we would find the line where the director’s name is present in that sentence. Once the line is found, we check whether this line contains DIN number followed by director name. If found we add it to the result. If not, we check the next line in the sentence, if found we add it to result, if not we again check every sentence to match the director’s name.

We again check the director’s named line for the director type like Executive or Independent, if found we add it to result. If not, we check the next line in the same sentence if not found until the last line in the same sentence, we check the complete sentence once if found we add it to result, if not we continue the process up to the last sentence.

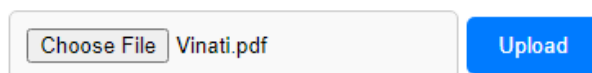
## 6. IMPLEMENTATION

### 6.1 BUILDING WEB APPLICATION

The Web application is developed using HTML (Hyper Text Markup Language) and CSS (Cascading Style sheet). We are using flask as background framework.

First, the user will upload the pdf file and then submit the pdf file. Then the pdf file is stored in local storage and the framework takes the path of the uploaded pdf file.

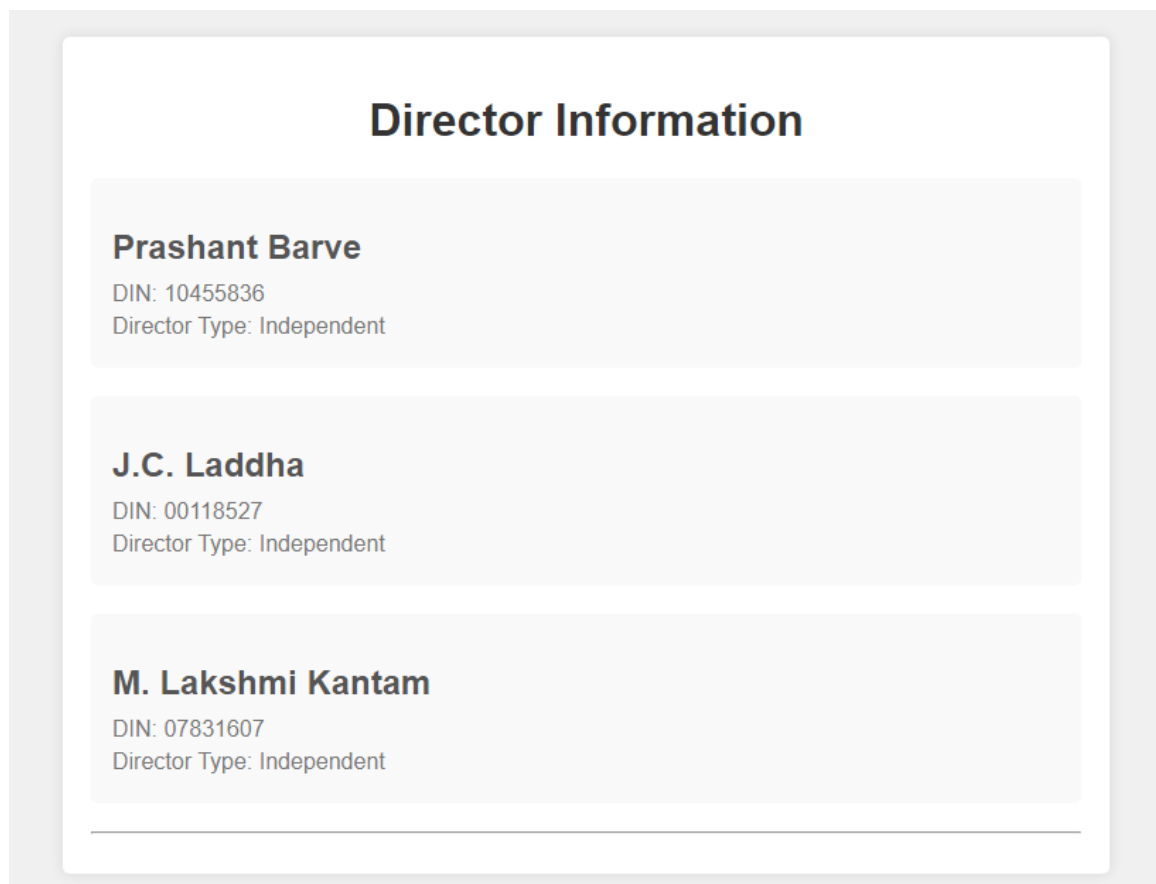
### Upload PDF File



Choose File Vinati.pdf Upload

Fig 6.1 Result page 1

The User clicks on the choose file option to upload the file, and then clicks upload. From that uploaded file the Director name, DIN, Director Type has been extracted.



### Director Information

<b>Prashant Barve</b> DIN: 10455836 Director Type: Independent
<b>J.C. Laddha</b> DIN: 00118527 Director Type: Independent
<b>M. Lakshmi Kantam</b> DIN: 07831607 Director Type: Independent

Fig 6.2 Result page 2



Once the user clicks on upload button the required text got extracted and printed as output in html page.

## 7. CONCLUSION

In conclusion, this project has been a challenging yet enlightening journey into the realm of PDF text extraction. From grappling with initial hurdles to mastering the art of extracting and printing the required results, every step has been a lesson in problem-solving and perseverance. As we wrap up, we emerge with sharpened skills and a deeper understanding of data manipulation techniques. This project has not only expanded our technical capabilities but also instilled confidence for future endeavours in similar domains.

## 8. FUTURE SCOPE

- To build a model to accurately extract text from the pdfs
- Improve Accuracy
- Adding the Search Query functionality to fetch the query from the pdf
- Image Processing
- Perform Image Processing in pdf to extract the text that the image has.

## 9. REFERENCES

- <https://trysakai.longsight.com/portal/site/520b8591-6130-4776-bf04-504e18e885f1/tool/e59fe5bb-2464-4b5c-9679-4d9af5857ede?panel=Main>
- [www.google.com](http://www.google.com)
- [www.youtube.com](http://www.youtube.com)