Data Extraction and Visualization of Form-like Structured Documents

Vishal Salgond - 1814107
Darshan Satra - 1814109
Nikhil Sharma - 1814114
Param Shendekar - 1814115

Group 5 | Mentor: Mrs. Dipti Pawade

Agenda

01 Problem Statement

02 Literature Survey

03 Motivation

04 Scope

05 Tech Stack

06 Dataset Description

07 Platform Overview

08 Flow Chart

09 References

PROBLEM STATEMENT

Create a platform that takes templatic form-like documents as input and using the Text Extraction technique and Sentimental Analysis can give valuable business insights about the desired and valid fields from the form. Analyzed insights will be then expressed in the form of visualizations which will help users/customers understand and improvise upon their work.

Literature Survey

Key takeaway:

[1] Text Retrieval from Scanned Forms Using Optical Character Recognition

- Recognition of uppercase text from manually filled form.

[2] Segmentation of scanned documents using deep-learning approach

- Important element detection in documents such as signatures, logo, printed text blocks, and tables.

[3] Form Field Frame Boundary Removal for Form Processing System in Gurmukhi Script(2009)

- True boundary identification for overlapping text with field boundary.

[4] Image Processing Based Scene-Text Detection and Recognition with Tesseract

- Usage of Tesseract and preprocessing necessary for OCR and the accuracy attainable.

Need / Reason / Motivation

Where we are

Lack of Analysis

Manual Work & Labour Cost

Errors & Mistakes

Where we want to be

Visual Analysis

Automated Workflow

Precise & Accurate output

Scope

- This product will take a templatic form as an input and then using segmentation and OCR, it will recognise the key-value pairs of the form. This meta data will be stored in the database.
- Sentiment analysis will be done on relevant fields chosen by the user
- Data Analysis will be done on relevant fields and will be shown to the user
- The submitted forms will have the same templated submitted at the initial stage.
- Form media type will be png or jpeg.
- Limited form types will be supported. And forms should have clear text boxes for the input.
- If there is any handwritten text then the characters should be clearly separated and In uppercase format.

Functional Requirements

1. User should be able to upload their own form.
2. Feature to choose which fields upon which analysis is required.
3. Extract all fields and its values from the uploaded bulk amount of forms.
4. Personalized dashboard to view visual insights of data.
5. Recognizable and limited font styles of print media will be supported.
6. Form types supported will be limited.
7. Sentiment Analysis will be performed on fields specified by the user, and are compatible.
8. Visual Analysis will be different for different types of fields.
9. System administrator will have the highest view as well as edit access.

Non Functional Requirements

a. The system should have the similar accuracy for the blank form and the filled form.
b. Availability will be high.
c. User authentication and security should be provided.
d. UI / UX will be aesthetic and user friendly.

Technology Stack

Frontend Technology

React JS

To build user interface for uploading and visualizing the forms

Backend Technology

Django

Develop Backend that will store the data and run the model

MongoDB

MongoDB database for persistent data storage

Standard ML Libraries of Python

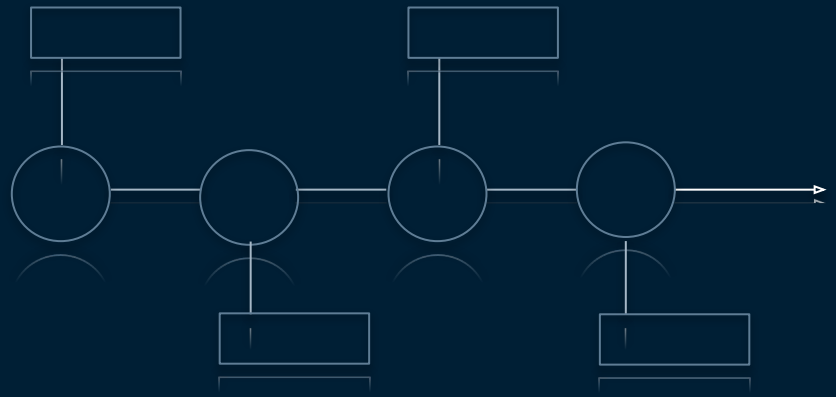
Tesseract and Tensorflow

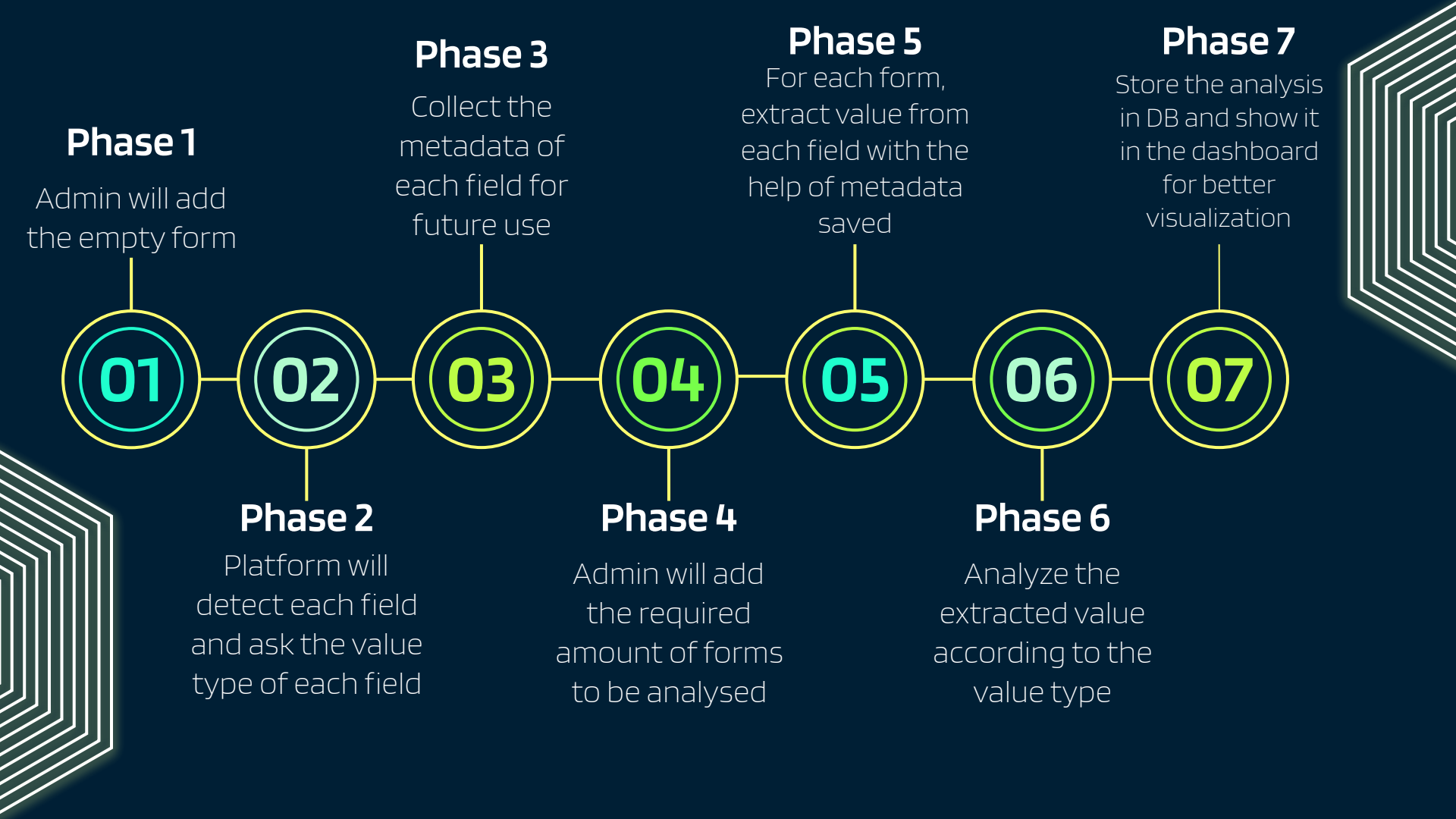
Dataset Description

Most of the datasets contain Images (png, jpg) of various documents which include forms, surveys, prescriptions, etc. We will leverage these to train and build our models for Boundary Detection & Text Extraction

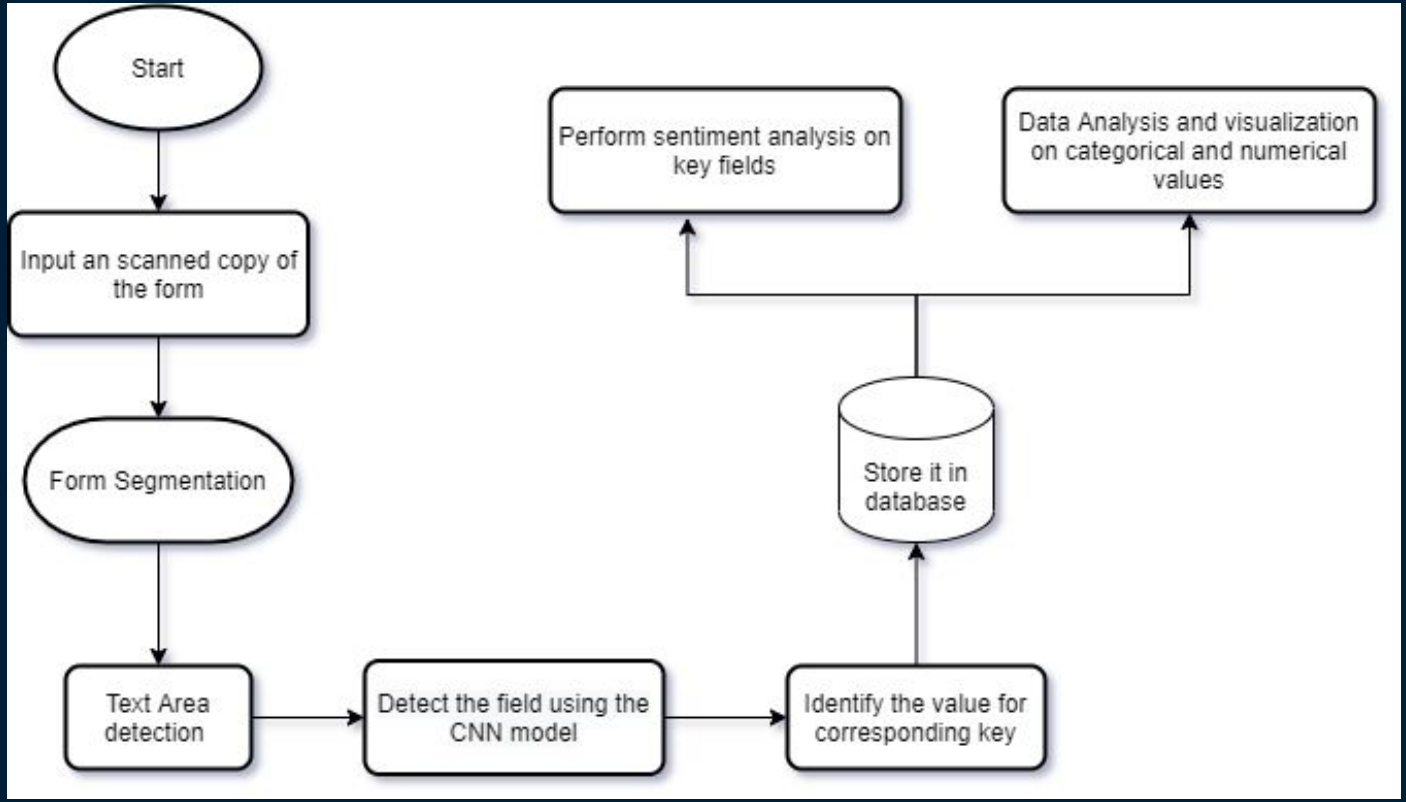
- https://guillaumejaume.github.io/FUNSD/
- https://paperswithcode.com/dataset/funsd
- https://www.kaggle.com/shaz13/real-world-documents-collections
- https://www.kaggle.com/patrickaudriaz/tobacco3482jpg
- https://www.kaggle.com/whatsappbackup/forms-images

PLATFORM OVERVIEW




Phase 1
Admin will add the empty form
01
Phase 2
Platform will detect each field and ask the value type of each field
02
Phase 3
Collect the metadata of each field for future use
03
Phase 4
Admin will add the required amount of forms to be analysed
04
Phase 5
For each form, extract value from each field with the help of metadata saved
05
Phase 6
Analyze the extracted value according to the value type
06
Phase 7
Store the analysis in DB and show it in the dashboard for better visualization
07

Flow Chart


Start
Input an scanned copy of the form
Form Segmentation
Text Area detection
Detect the field using the CNN model
Identify the value for corresponding key
Store it in database
Perform sentiment analysis on key fields
Data Analysis and visualization on categorical and numerical values

Deliverables in Semester 7

October First Fortnight	Create & Submit SRS, SPMP, SDD, STD
October Second Fortnight	Achieve Basic Text Extraction From Single Form Type
November First Fortnight	Create Product Flow, Sentiment Analysis on Extracted Data, Synopsis.

We will be working on different aspects of the project like Text Extraction, Boundary Detection, Segmentation, Character Classification, etc which then will be merged along with Product website, which will also be in creation side by side.

References

1. Thomas Hegghammer. *OCR with Tesseract, Amazon Textract, and Google Document AI: A Benchmarking Experiment.* June, 2021.

2. Forczmański P., Smoliński A., Nowosielski A., Małecki K. (2020) *Segmentation of Scanned Documents Using Deep-Learning Approach.* In: Burduk R., Kurzynski M., Wozniak M. (eds) Progress in Computer Recognition Systems.

3. Dharam Sharma, Gurpreet Lehal "*Form Field Frame Boundary Removal for Form Processing System in Gurmukhi Script*" 2009 10th Intl. Conference on Document Analysis and Recognition.

4. Zacharias et al. "*Image Processing Based Scene-Text Detection and Recognition with Tesseract*". Apr, 2020.

5. S.N. Srihari, Y. C. Shin, V. Ramanaprasad, D. S. Lee, "*A System to Read Names and Addresses on Tax Forms*", Proc. of the IEEE, 1996, vol. 84, issue. 7, pp. 1038-1049.

6. B. P Majumder et. al. "*Representation Learning for Information Extraction from Form-like Documents*"

7. V. Aggarwal et al. "*Text Retrieval from Scanned Forms Using Optical Character Recognition*" 2018

Thank You!