1

2

3

4

5

6

7

8

9

Data understanding

Stamp Data

LayoutLM data - Preparing Data (OCR & Transformation)

Preprocessing results example

Overlap and merge steps

Next steps

Advenssement

1

2

3

4

5

6

7

Data understanding

Missing Files

non existing files

```
image_files = set(os.listdir(image_folder_path))

# the expected filenames (adding the suffix '_1.png')
expected_filenames = set(data['filename'].apply(lambda x: f"{x}_1.png"))

csv_not_in_folder = expected_filenames - image_files
folder_not_in_csv = image_files - expected_filenames

print("Files listed in CSV but not found in the folder:\n", csv_not_in_folder)
print('\n')
print("Files in the folder but not listed in CSV:\n", folder_not_in_csv)
```

196 images

```
Files listed in CSV but not found in the folder:
 {'IEEE fAC NÃ°84.pdf_1.png', 'ilovepdf_merged-2(1).pdf_1.png', 'ilovepdf_merged-4(1).pdf_1.png', '2022-09-14--2022-10-12_RÃ©sumÃ©_Facture.pdf_1.png',
_merged-3(1).pdf_1.png'}
```

Empty Fields

```
Empty fields in each column:
filename            0
title_bbx          26
date_bbx            9
stamp_bbx          52
signature_bbx      65
ieee_bbx           26
total_bbx          12
totalValue_bbx     13
title              27
handwritten         0
```

As the reaction of our models to nan fields in json files, Two repositories were created. These initial files are designed to stamp/signature and Donut model

The annotations are now aligned

1

2

3

4

5

6

7

Stamp Data


```
{"filename": "files (65).pdf_1.png",
"title_bbx": NaN,
"date_bbx": "[0.8035055026194582, 0.22101015654140369, 0.10984140250389518, 0.019127320758753452]",
"stamp_bbx": NaN,
"signature_bbx": NaN,
"ieee_bbx": NaN,
"total_bbx": "[0.5312225544596436, 0.5808427928096865, 0.058788357100898025, 0.01852960124816716]",
"totalValue_bbx": "[0.8398615030090394, 0.5838314359656938, 0.08972960389250173, 0.019127320758753452]"}
```



```
{"filename": "files (65).pdf_1.png",
"date_bbx": "[0.8035055026194582, 0.22101015654140369, 0.10984140250389518, 0.019127320758753452]",
"total_bbx": "[0.5312225544596436, 0.5808427928096865, 0.058788357100898025, 0.01852960124816716]",
"totalValue_bbx": "[0.8398615030090394, 0.5838314359656938, 0.08972960389250173, 0.019127320758753452]"}
```

LayoutLM data - Preparing Data (OCR & Transformation)

1
2
3
4
5
6
7

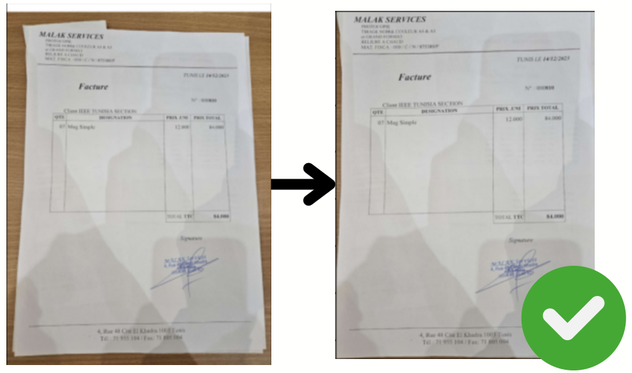
initial data

Images:
bad quality, noise,... => Preprocessing needed
Annotations:
Missing Text Values and unsuitable for Layoutlmv3

Preprocessing
Images + bboxes

- **Gray Scaling**
- **Fix Illumination (CLAHE)**
- **Denoising (NLM)**
- **Skew angle determination**
- **rotation**
- **Resizing**
- **DPI Scaling**

- ~~**Binarization**~~ **(Adaptive Thresholding, OITSU ...)**
- ~~**Dilation , Erosion**~~
- **Wrapping**

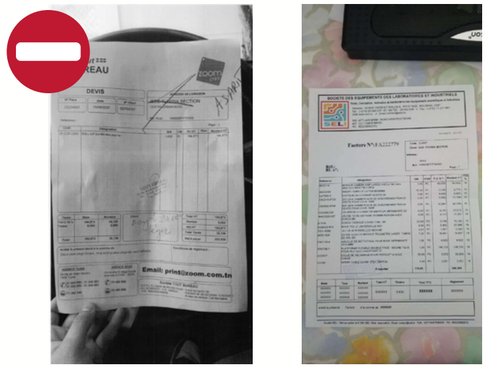


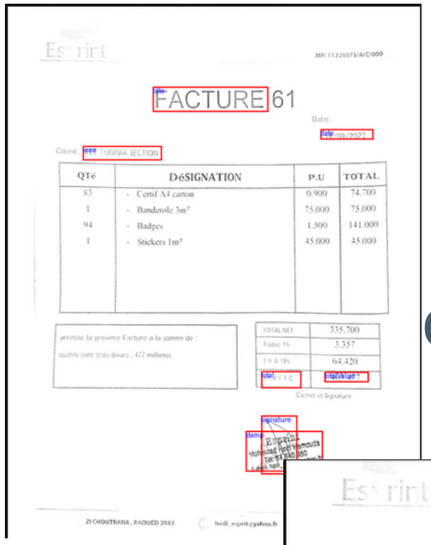
OCR Application & Transformation

- Change annotation shape => Json file
- OCR Application
- Overlapping Test with our annotation

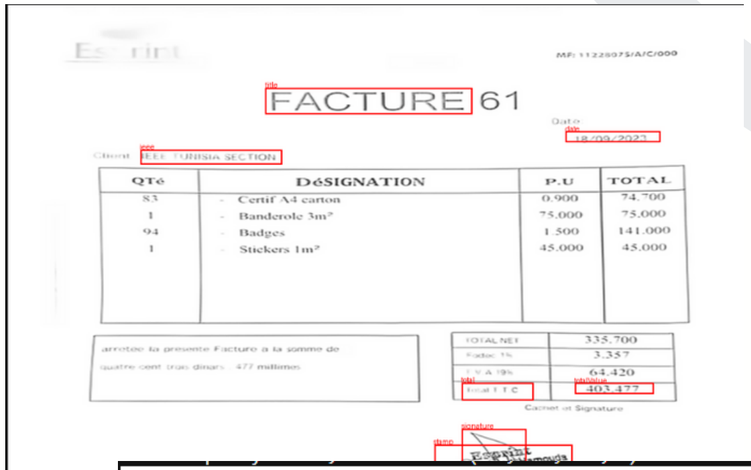
if yes => label<- annotation.label
else label = other

- **if an annotation.label doesn't exist => merge the "item"** 😔
- **Merge repeated labels**





Original image



Preprocessed image

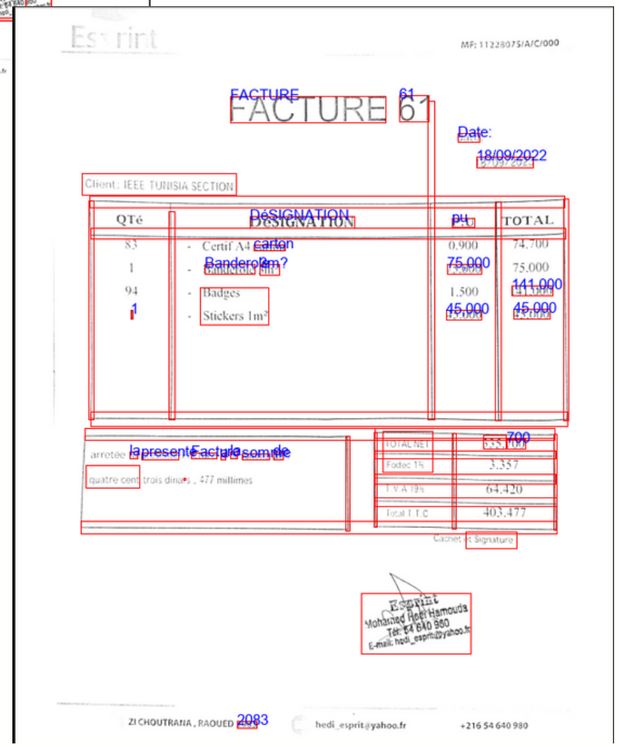
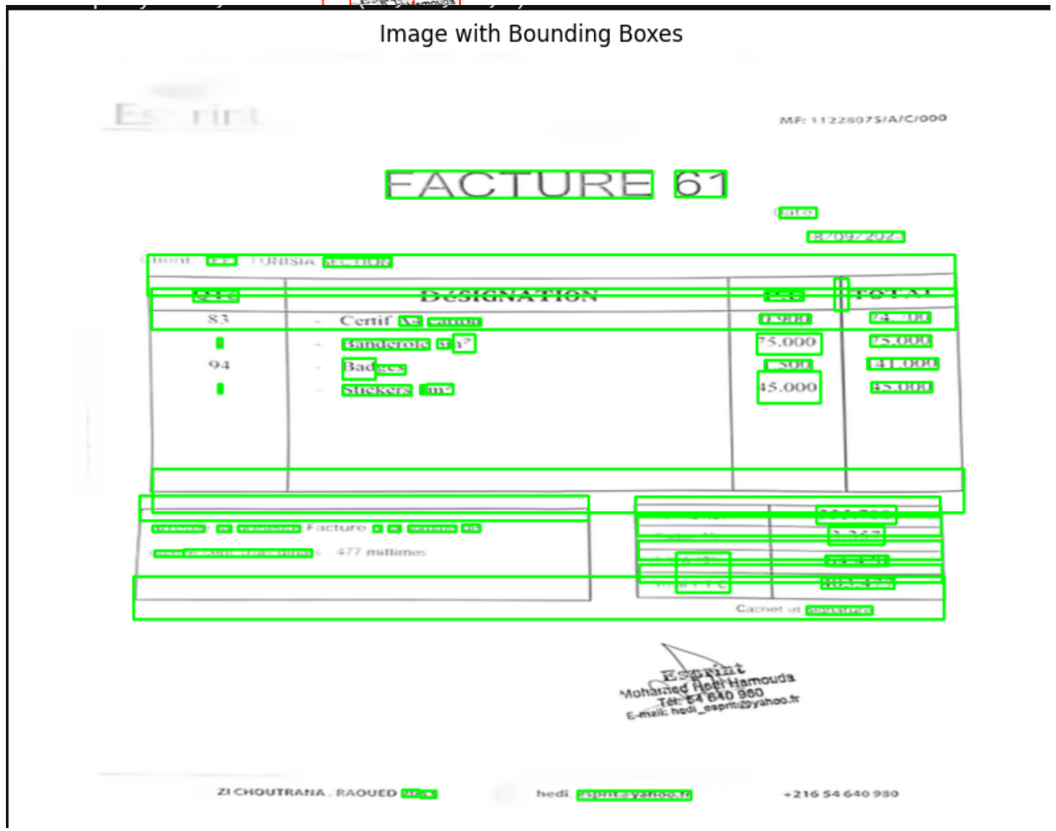
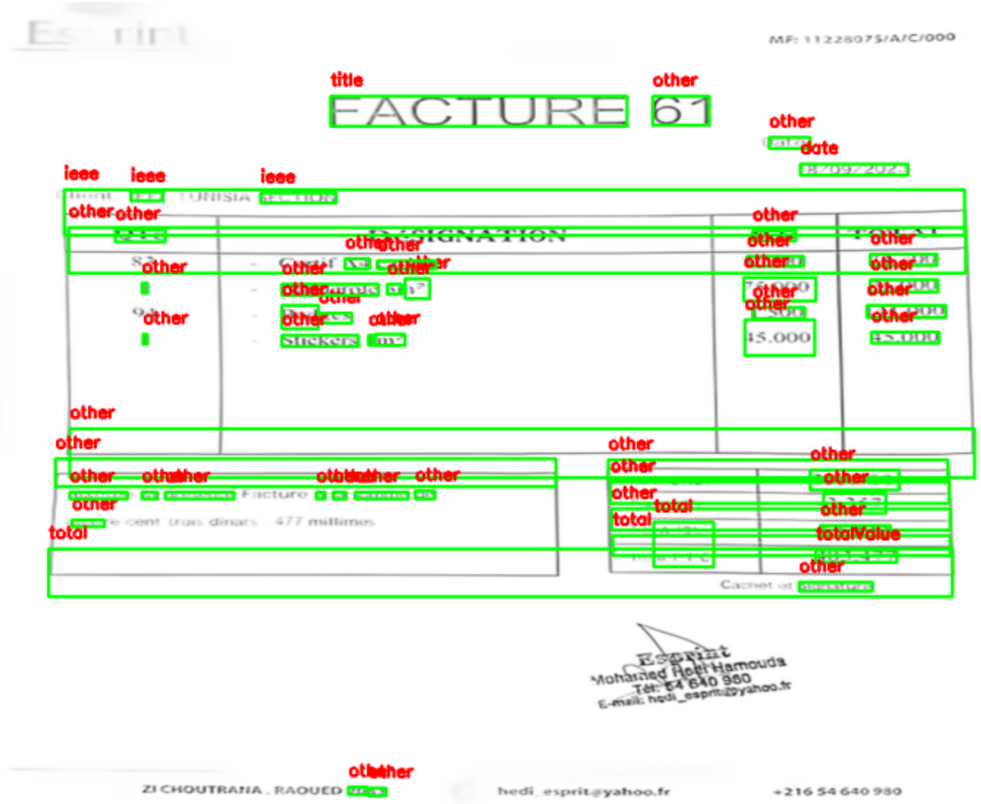


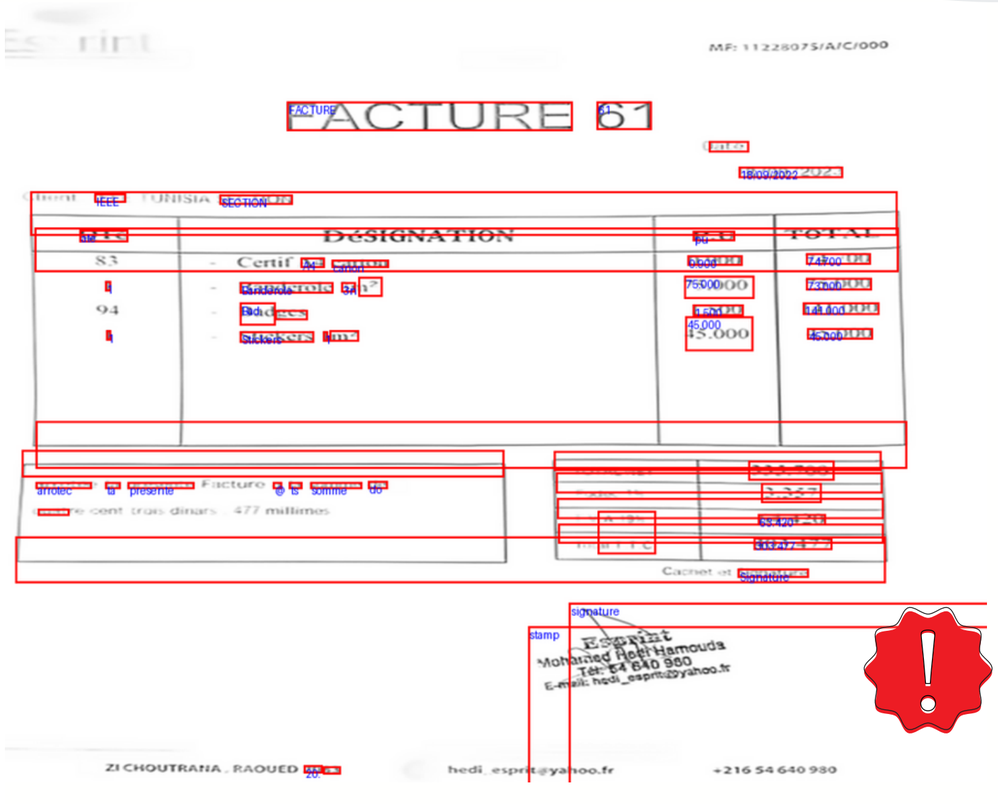
Image with Bounding Boxes



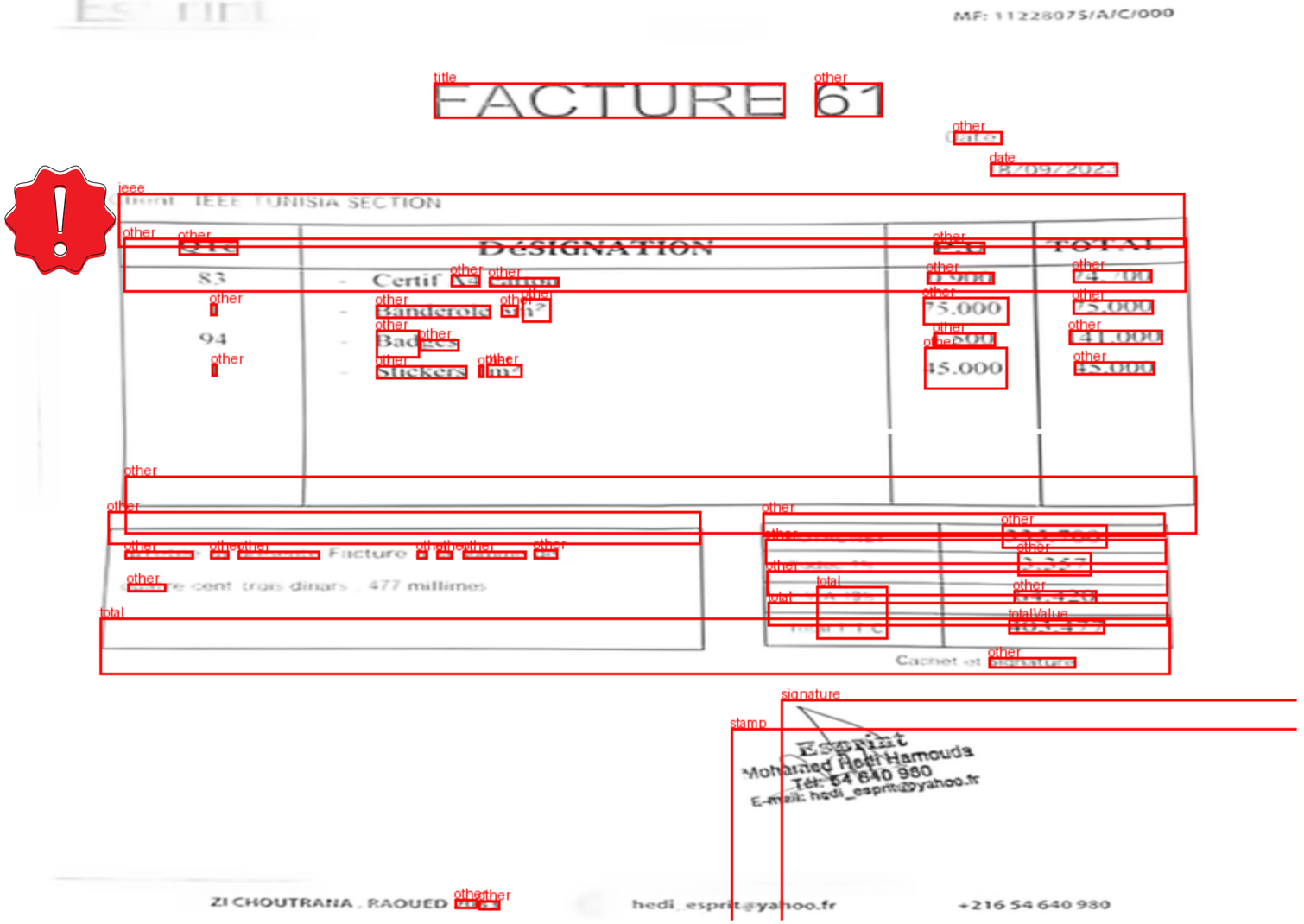
Test Overlapping



Merge with labels



Merge text with the same label



1
2
3
4
5
6
7

Next Steps

Model Fine Tuning
Test after Data Augmentation & Better OCR results