CAPSTONE PROJECT

Document Classification Capability
Revving up manual paperwork with Computer Vision & NLP

MIT x Wolters Kluwer

Faculty Advisor: Dr. Ilya Jackson
WK Advisors: Pooja Srivastava & Varun Dixit

18th August 2023

Rachit Jain
Chloe Wu
Candidates of Master of Business Analytics, MIT
Agenda

Introduction
- Overview
- Challenges
- Motivation

The Process
- Solution
- Methodology
  (Re-labelling, Modelling)

Results
- Demo
- Deliverables
- Impact & Business Value
Overview | Motor Vehicle registration process is error-prone

Challenges At Scale

Huge Volume
50k+ pages\(^1\) per day

Multiple rejections
10% rejection rate

High processing time
10 mins per request (~20 pages)

\(^1\) 80 requests per day only for Lien-Add document type (1 on 8) in Texas state (1 on 50) with 20 pages per request
Challenges | Manual processing is a **bottleneck**

<table>
<thead>
<tr>
<th>Challenges At Scale</th>
<th>50k+ pages(^1) per day</th>
<th>10% rejection rate</th>
<th>10 mins per request (~20 pages)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huge Volume</td>
<td>Multiple rejections</td>
<td>High processing time</td>
<td></td>
</tr>
</tbody>
</table>

**Final Goal**

Build an automated, generalized document classification **capability** to make historically manual logistics paperwork easier to execute and more accurate.
Motivation | Need for more than rule-based systems

Similar formats, similar text, but different titles

1. Return Unprocessed
2. Search Request Form
3. Identification Card
4. Binder of Insurance
Our Solution

Computer Vision + Natural Language Processing + Title Extraction

Document Classification Capability

Capstone of the Year? 😐
Imbalanced dataset across 120+ categories; 12k scanned pages

Training Data

12k+ Data Points

10k+ from top 31 classes

Represented Counties:
Bexar, Brazoria, Dallas, El Paso, Fort Bend, Harris, Hidalgo, Lubbock, Travis, Van Zandt
Solution | End-to-End Workflow takes PDF input and gives multiple labels for each page, along with confidence scores.
Step 0: Need for relabeling - Mystery behind ground truth labels

Status Quo

Current ‘Ground Truth’ Label = Output of Champion model
Wrong Labels → Poor Models

Solution?

‘Smarter’ Manual Labelling [Unsupervised Clustering on Deep Embeddings]

- Cluster similar image embeddings from trained vision model
- Merge clusters on common categories & create sub-clusters
- Smarter Manual Label

Clustering over Image Embeddings

Fig: Embeddings for each image clustered based on similarity

Results

88% saving in time for manual labelling
10.6% Mis-classified labels identified
8% F1 Score jump!
Methodology

INPUT

PDF Docket

PDF Breaker API

PDF Breaker on a Streamlit dashboard

Each page is extracted as an image and saved in folders

OUPUT

Highlights: Image padding, Image processing, Hyper-parameter tuning, Data Augmentation, Feature Engineering…

1. PDF Breaker API
2. Vision-based Model
3. Text-based Model

PDF Breaker API

Image-based Model

Extracted Text

Features

Text-based Model

CNN model fine-tuned on 12k scanned documents

BERT + XGBoost running on text extracted from AzureOCR

Data Augmentation, Feature Engineering…

Methodology

INPUT

PDF Docket

PDF Breaker API

PDF Breaker on a Streamlit dashboard

Each page is extracted as an image and saved in folders

OUPUT

Highlights: Image padding, Image processing, Hyper-parameter tuning, Data Augmentation, Feature Engineering…

1. PDF Breaker API
2. Vision-based Model
3. Text-based Model

PDF Breaker API

Image-based Model

Extracted Text

Features

Text-based Model

CNN model fine-tuned on 12k scanned documents

BERT + XGBoost running on text extracted from AzureOCR

Data Augmentation, Feature Engineering…

Methodology

INPUT

PDF Docket

PDF Breaker API

PDF Breaker on a Streamlit dashboard

Each page is extracted as an image and saved in folders

OUPUT

Highlights: Image padding, Image processing, Hyper-parameter tuning, Data Augmentation, Feature Engineering…

1. PDF Breaker API
2. Vision-based Model
3. Text-based Model

PDF Breaker API

Image-based Model

Extracted Text

Features

Text-based Model

CNN model fine-tuned on 12k scanned documents

BERT + XGBoost running on text extracted from AzureOCR

Data Augmentation, Feature Engineering…

Methodology

INPUT

PDF Docket

PDF Breaker API

PDF Breaker on a Streamlit dashboard

Each page is extracted as an image and saved in folders

OUPUT

Highlights: Image padding, Image processing, Hyper-parameter tuning, Data Augmentation, Feature Engineering…

1. PDF Breaker API
2. Vision-based Model
3. Text-based Model

PDF Breaker API

Image-based Model

Extracted Text

Features

Text-based Model

CNN model fine-tuned on 12k scanned documents

BERT + XGBoost running on text extracted from AzureOCR

Data Augmentation, Feature Engineering…
Methodology

1. PDF Breaker API

PDF Breaker on a Streamlit dashboard

Each page is extracted as an image and saved in folders

2. Vision-based Model

CNN model fine-tuned on 12k scanned documents

3. Text-based Model

BERT + XGBoost running on text extracted from AzureOCR

4. Heuristic-based Model

Potential title extraction from any document type

5. Ensembling

Selecting ‘supreme’ model; if clash, choose heuristic
The labels need to be ensembled into one single prediction

<table>
<thead>
<tr>
<th>Category</th>
<th>Vision Supremacy</th>
<th>Text Supremacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Driving License</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Lien Add – Order</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Document</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Binary Model Strength Look-Up Table (based on F1 Score)

Fig: Logic workflow to combine the labels

Our Implementation

Model Supremacy for Vision and Text based models

Training F1 score used to assign ‘supreme’ model for each category

<table>
<thead>
<tr>
<th>Case</th>
<th>Same Label</th>
<th>Vision Supremacy?</th>
<th>Text Supremacy?</th>
<th>Final Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Same</td>
</tr>
<tr>
<td>Case 2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>Manual Review</td>
</tr>
<tr>
<td>Case 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>Vision</td>
</tr>
<tr>
<td>Case 4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>Text</td>
</tr>
<tr>
<td>Case 5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Heuristic</td>
</tr>
</tbody>
</table>

Ensemble Technique
Result | **End-to-End multi-modal architecture deployed**

Video: Demo implemented over FastAPI and **deployed** over WK's Virtual Machine
Deliverables | Scalable model pipeline deployed over WK Cloud

**Result Summary**

- **0.86** F1 Score
  - Over 31 document types over 2.1k highly noisy test dataset

**Challenger > Champion**

- Our implementation beats the status quo with 5% higher F1 score

**Our Deliverables to Wolters Kluwer**

1. **OCR API**
   - Page as Image → Azure OCR Response
2. **CAPABILITY API**
   - Page as Image → Final Label + Insights
3. **Documentation & Knowledge Transfer**
   - for smooth integration with current system

**Learnings**

- **Industry Practices**
  - Dashboarding
  - Capability Design

- **Tech Stack**
  - Azure
  - FastAPI

- **Business Skills**
## Business Value

<table>
<thead>
<tr>
<th>Shareholders</th>
<th>Customers</th>
<th>Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10X growth in business</strong></td>
<td><strong>3X lower turn-around time</strong></td>
<td><strong>10X faster processing</strong></td>
</tr>
<tr>
<td>Generalized capability → Scalable business model</td>
<td>Less headache → Happy customers</td>
<td>Free from mundane work → higher productivity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Market Differentiation</th>
<th></th>
<th>Future-Ready by leveraging AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leveraging AI → journey to be market-leaders</td>
<td><strong>70% lower rejection rate</strong></td>
<td>Machine aided human experts → Upskilling</td>
</tr>
<tr>
<td>Quick and efficient → drives customer experience</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
High learning + Solid deliverables + Strong impact =

More ACCURATE Labelling

0.86 F1
High → Better

AUTOMATED pipeline

1 API
running everything

LOW processing time

10X saving
in processing time

Flexible PDF BREAKER

Unrestricted
# of PDFs broken

FEWER rejections

Challenger >
Champion (status-quo)

PRODUCTIONIZED pipeline

Deployed
over WK’s VM

Result INTERPRETABILITY

User Insights
behind predictions

END-TO-END pipeline

Streamlined
workflow

EASY-TO-INTEGRATE capability

Scalable
and reproducible

Deployed
over WK’s VM

User Insights
behind predictions

Result INTERPRETABILITY

Streamlined
workflow

EASY-TO-INTEGRATE capability

Scalable
and reproducible

Deployed
over WK’s VM

User Insights
behind predictions

Result INTERPRETABILITY

Streamlined
workflow

EASY-TO-INTEGRATE capability

Scalable
and reproducible
Thank You!

“Only those who will risk going too far can possibly find out how far one can go!”
~ T.S. Eliot

Rachit Jain
Candidates of Master of Business Analytics, MIT

Chloe Wu