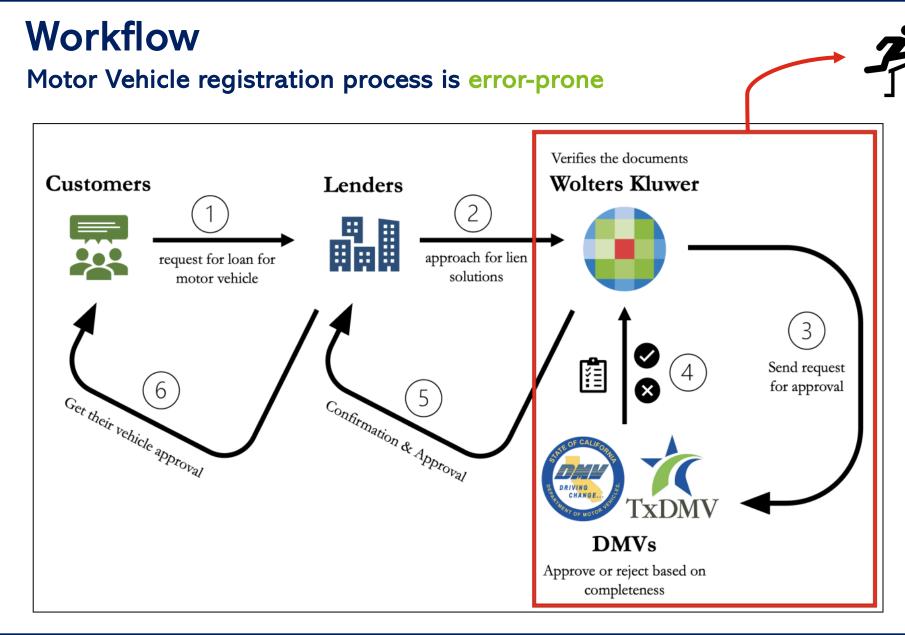


Document Classification Capability

Revving up manual paperwork with Computer Vision & NLP

Faculty Advisor: Dr. Ilya Jackson Wolters Kluwer Supervisors: Pooja Srivastava, Varun Dixit





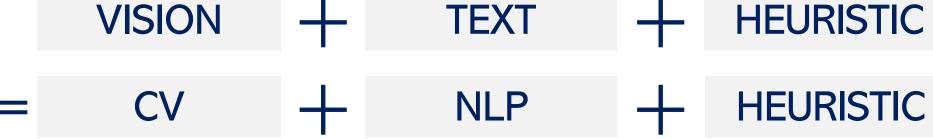
Challenges Manual processing → bottleneck **Huge Volume** 100k+ pages* come in per day Multiple rejections 10% rejections High processing time 10 mins per request (~20 pages)

Goal Need for more than just rule-based systems

Build an automated, generalized document classification capability to make historically manual logistics paperwork easier to execute and more accurate

Solution

Multi-modal approach for document types inside & outside the training set



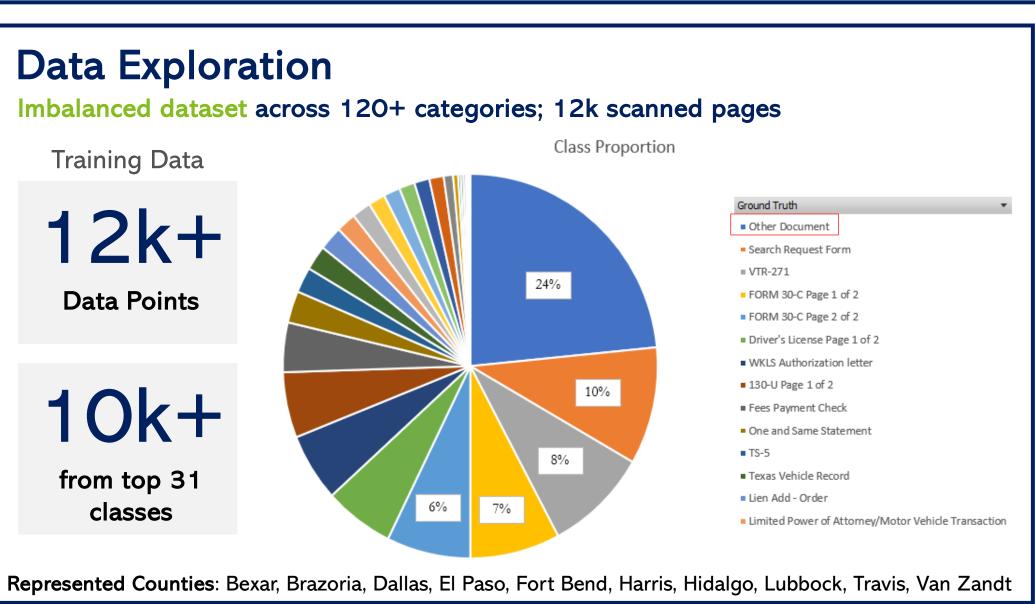
(1)

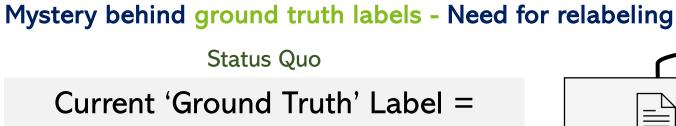
saving in time for

manual labelling

Computer Vision Natural Language Processing

Title Extraction





Output of Champion model Wrong Labels → Poor Models

Solution?

Data Processing

'Smarter' Manual Labelling [Unsupervised Clustering on Deep Embeddings]

- Image embeddings from trained vision model
- Merge clusters on common categories & create sub-clusters
- Create sub-clusters within the clusters Manually label documents, but smartly!

Cluster 5 Fig: Clustering based on embedding similarity

88% 10.6%

Mis-classified labels identified

Results

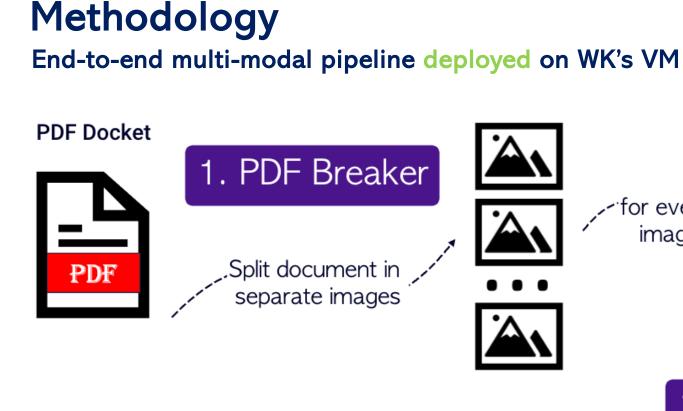
8% F1 Score jump!

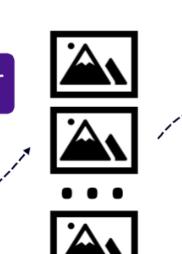
Cluster

Cluster 2

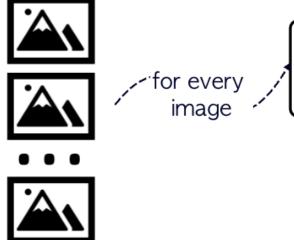
Cluster 3

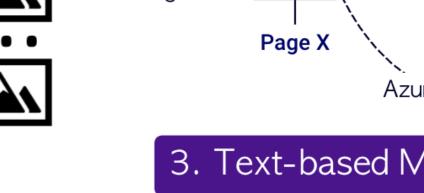
Cluster 4

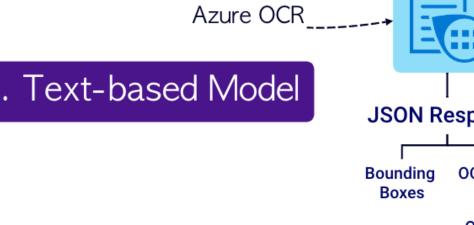


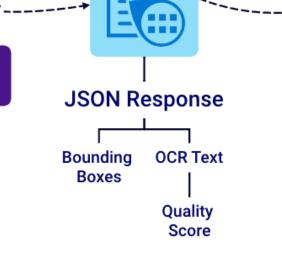








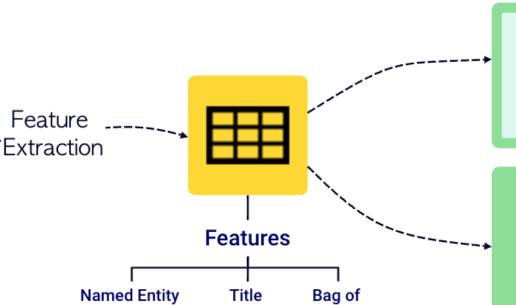




Extracted Information

Feature

2. Image-based Model



Words

Dissimilarity

Trained CNN Model

Confidence Score

-Confidence Score

_abel A

Label B

-Label C Confidence Score

Vision-based Model

PDF Breaker API

Extracting each page as an

image from PDF document

CNN model fine-tuned on 12k scanned pages **Text-based Model**

BERT + XGBoost running on text extracted from AzureOCR

Highlights: Image padding, Image and text processing, Transfer Learning, Hyper-parameter tuning, Data Augmentation, Feature Engineering

Heuristic-based Model

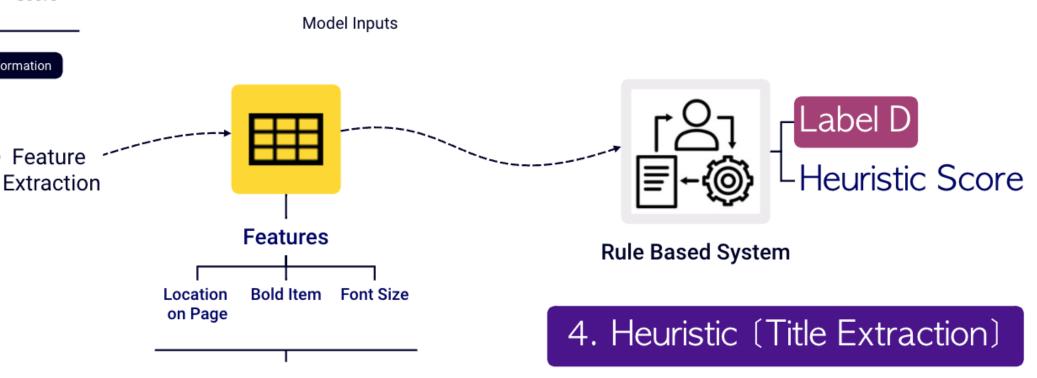
Potential title extraction from any document type

Used bounding boxes from Azure OC, narrowed search space and engineered features

Ensemble Labels

Selecting 'supreme' model; if clash, choose heuristic

The final model is chosen based on the F1 score during training & rules based on heuristic title



| Case | Same Label | Vision Supermacy? | Text Supremacy? | Final Model |
|--------|------------|----------------------|-----------------|---------------|
| Case 1 | 1 | 1 | 0 | Same |
| Case 2 | 0 | 1 | 1 | Manual Review |
| Case 3 | 0 | 1 | 0 | Vision |
| Case 4 | 0 | 0 | 1 | Text |
| Case 5 | 1 | 0 | 0 | Heuristic |

Table: Ensemble technique rules

Results + Deliverables + Impact

Final model predictions with added visibility

Fully functional Classification API deployed on Wolters Kluwer's virtual machine to be productionized into their current capability, giving added flexibility & scalability!

| | | | | Prediction | Document | |
|----------|--------|-------------|------------------------------|------------|----------|---------------|
| Docket # | Page # | Final Model | Final Prediction | Confidence | Quality | Notes |
| 616823_1 | 1 | Matched | Lien Add - Order | 1.00 | 0.699 | |
| 616823_1 | 2 | Matched | FORM 30-C Page 1 of 2 | 1.00 | 0.766 | |
| 616823_1 | 3 | Text | MV-50 | 0.49 | 0.640 | |
| 616823_1 | 4 | Vision | Driver's License Page 1 of 2 | 0.88 | 0.292 | |
| 616823_1 | 5 | Heuristic | Binder of Insurance | 0.68 | 0.612 | Manual Review |

Table: Results not only have the final prediction, but chosen model, confidence score & added interpretability

More ACCURATE Labelling

0.86 F1

High → Better

Flexible PDF BREAKER

Unrestricted

of PDFs broken

Result INTERPRETABILITY Insightful

results with predictions

AUTOMATED pipeline

1 API running everything

FEWER rejections

Challenger >

Champion (status-quo)

END-TO-END pipeline

Streamlined workflow

EASY-TO-INTEGERATE capability

LOW processing time 10X saving

in processing time

Ready to PRODUCTIONZE

Deployed

over WK's VM

Scalable and reproducible