EMjoyers

Julia Jakubowska, Rafał Kaczmarek, Agata Kulesza, Łukasz Niedźwiedzki, Weronika Skibicka

Task

Document type classification
based on scans
(or OCR-generated words dataset
for scan)

Problems

- translational invariance
- space-specific placements of sections of documents
- not standarised documents layout within classes
- OCR generated words are incomplete and seem inacurate
- poor labeling of visual data

Idea

Easily-accessible open-source models.

ImageNet pretrained model, to extract **feature-embeddings** from.

Vision Transformer (ViT)

Positional retain the dependence in between the patches

Passes the knowledge of specific placement to higher-level features

OCR DATA

Not accurate

Inconsistent in relation to data from scan

- Not all files have any words assigned
- Different number of words assigned to a document

Does not include Polish words

We bet on image scan database

VISUAL DATA

Large data repository

One of the few advantages of data

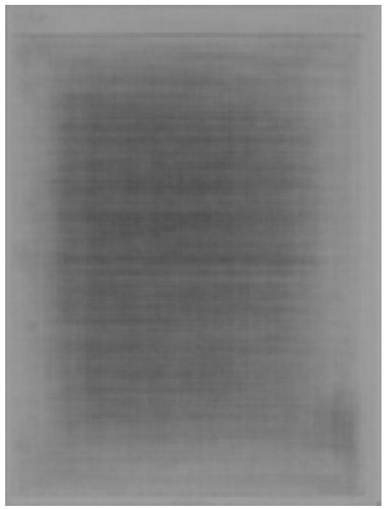
Overrepresentation of classes in the context of possible business demands

Advantage in data quality over text data from OCR

VISUAL DATA

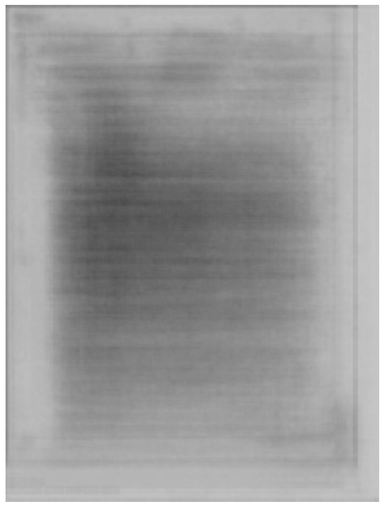
High variation among class examples

Examples of averaged image from the classes



scientific_report

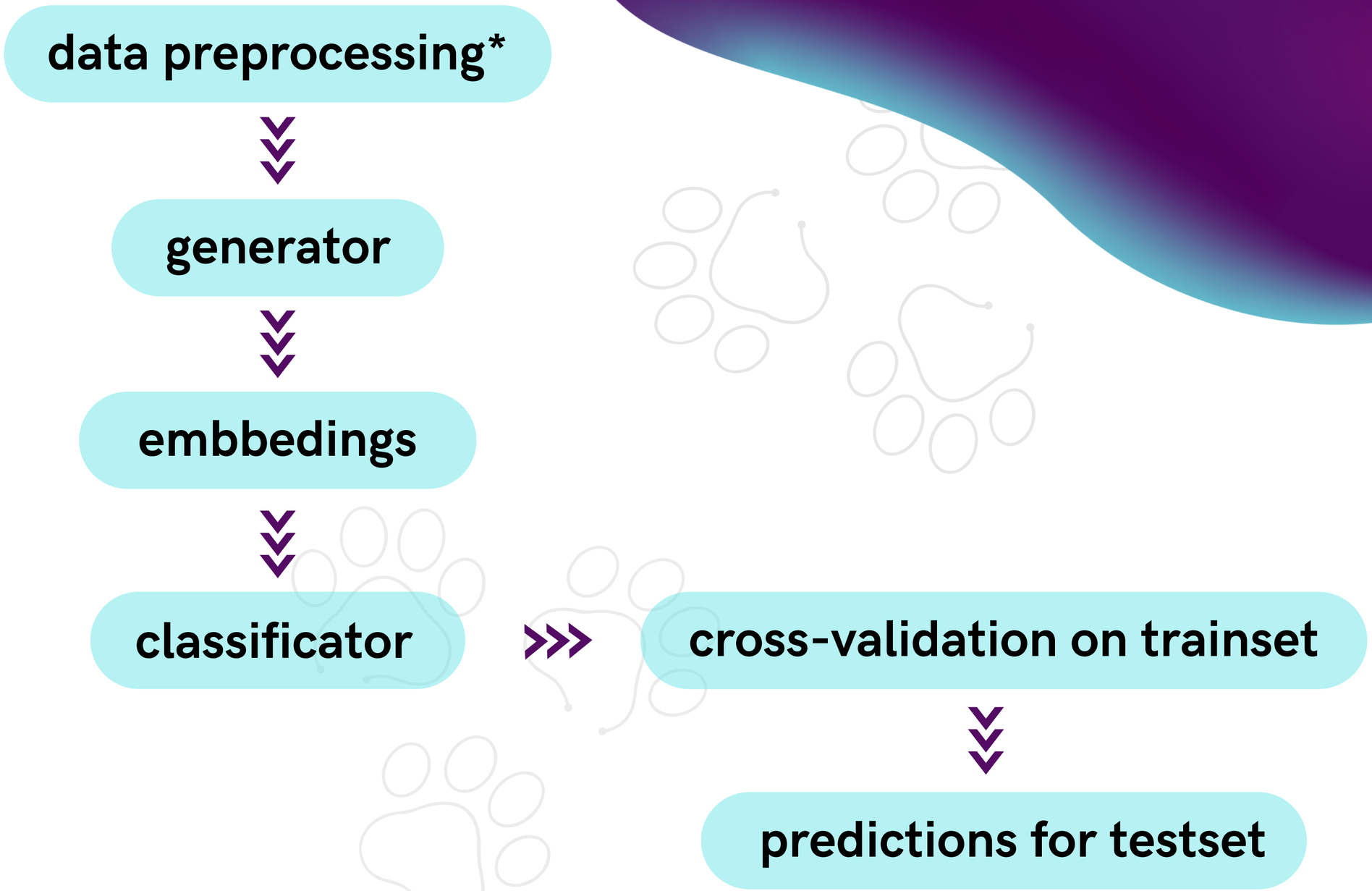
PIT-37



memo

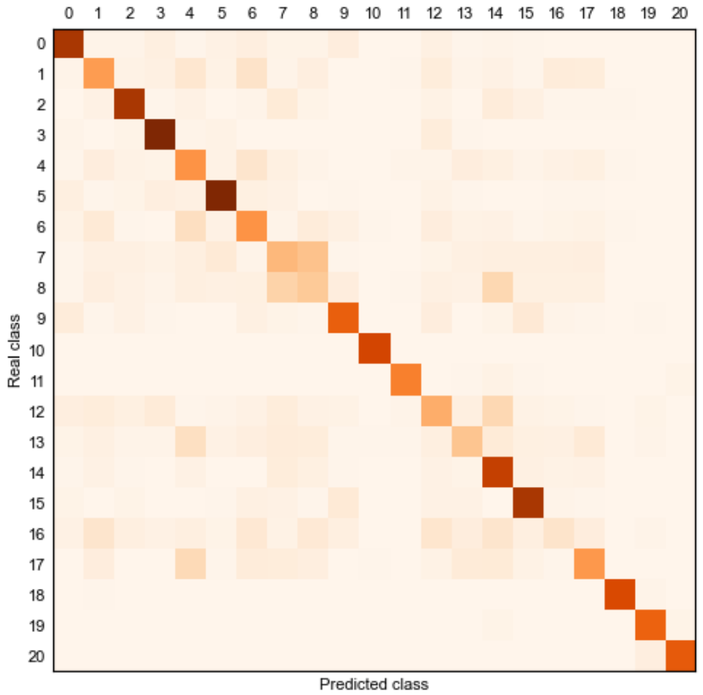
02

Document recognition POC

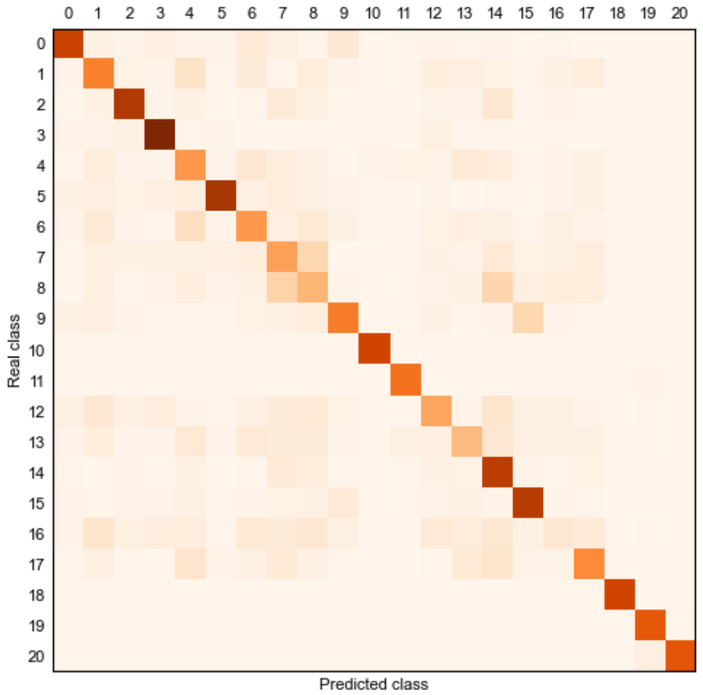

data preprocessing*
generator
embbedings
classificator
cross-validation on trainset
predictions for testset

Do feature-embeddings have any sense?

Quick test...



logistic regression

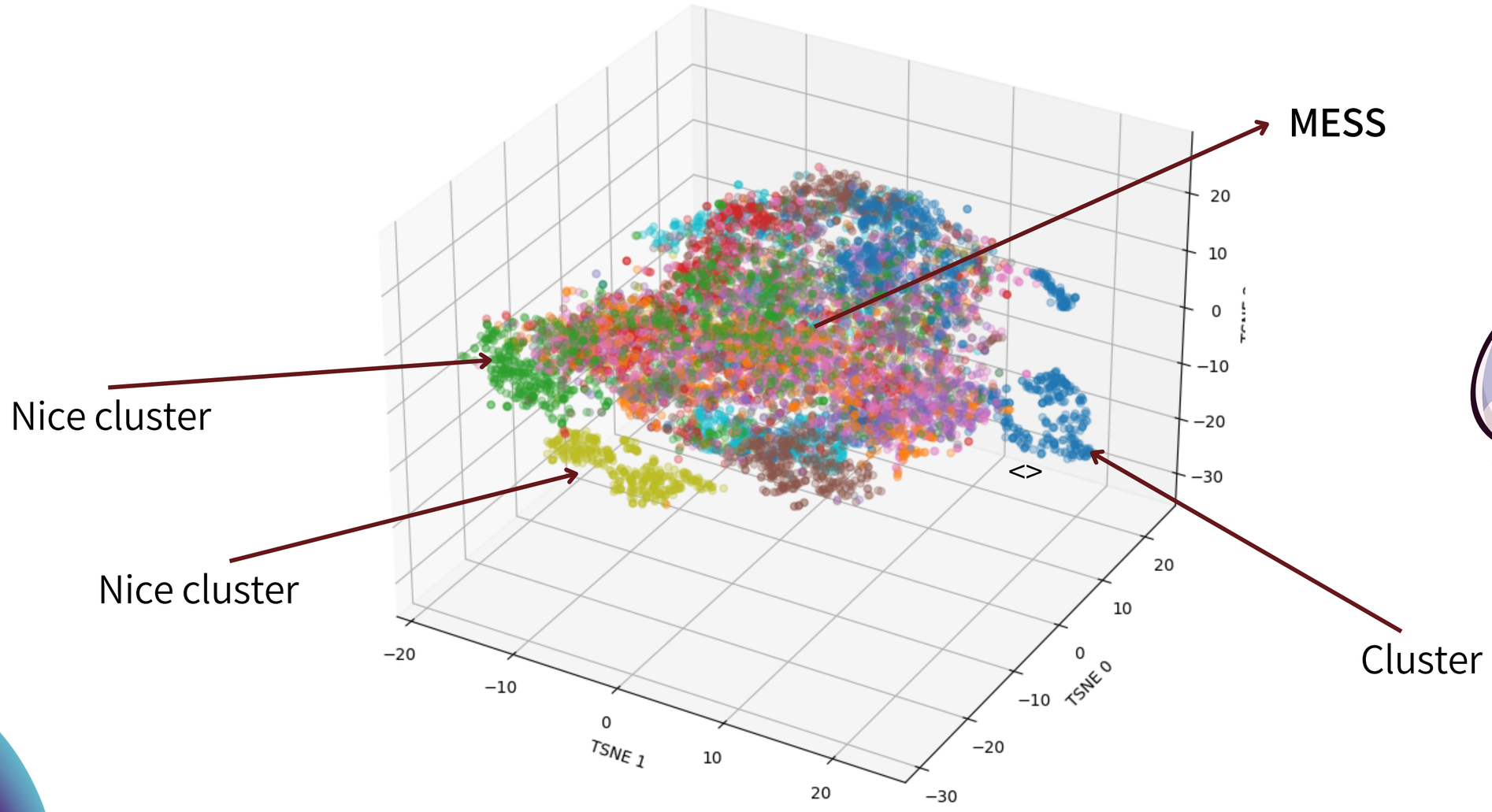


k-nearest neighbors

YES!
First succes

TSNE

T-distributed Stochastic Neighbor Embedding



Finally, we decided to bet on…

Multi-layer Perceptron classifier
2 dense layers

Results

Cross validation	10% validation, 90% database from training set
Mean accuracy	0.6648

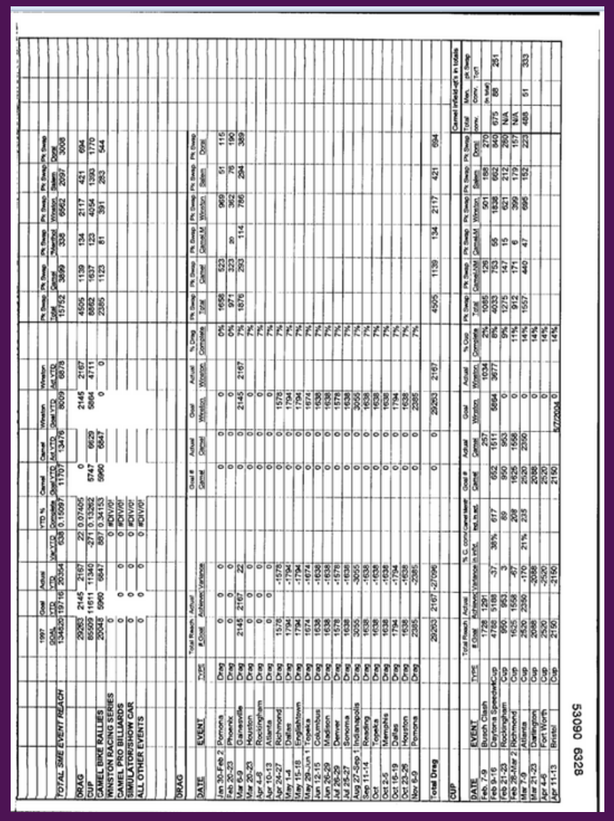
03

Areas to improve

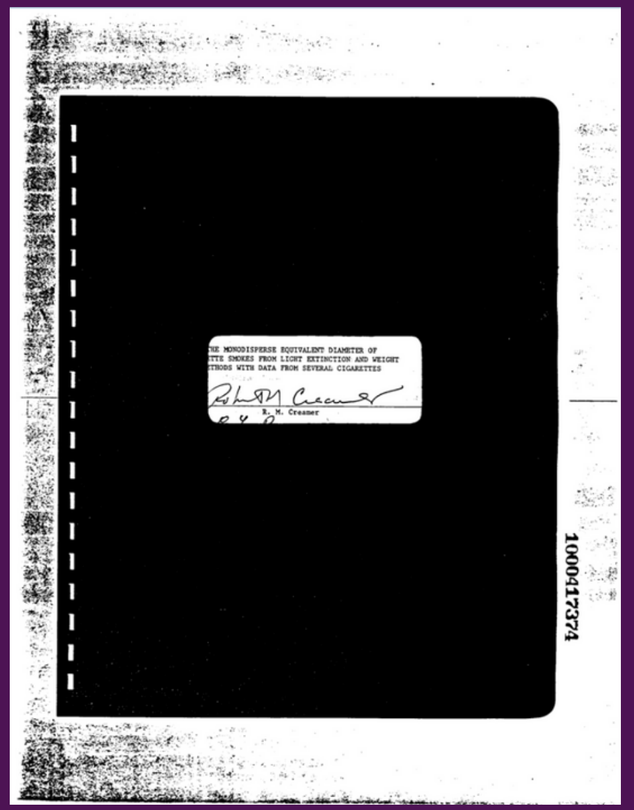
We believe a key to a better classification is a more quality dataset

We observed a significant case of poor labeling across most of classes in original trening set

Original label: questionnare



Original label: email

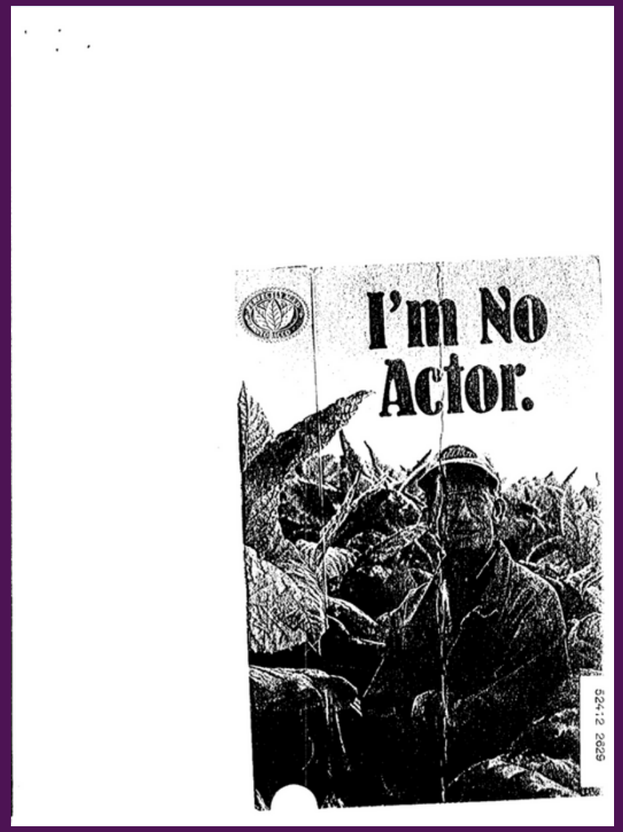


Original label: scientific research

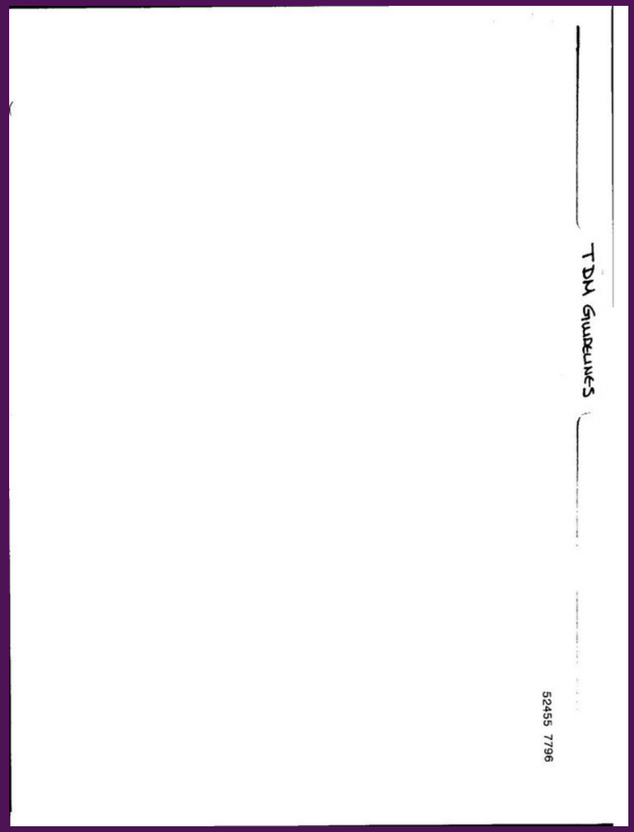
We believe a key to a better classification is a more quality dataset

We observed a significant case of poor labeling across most of classes in original trening set

Original label: questionnare



Original label: handwritten



Original label: form

Results on filtered data

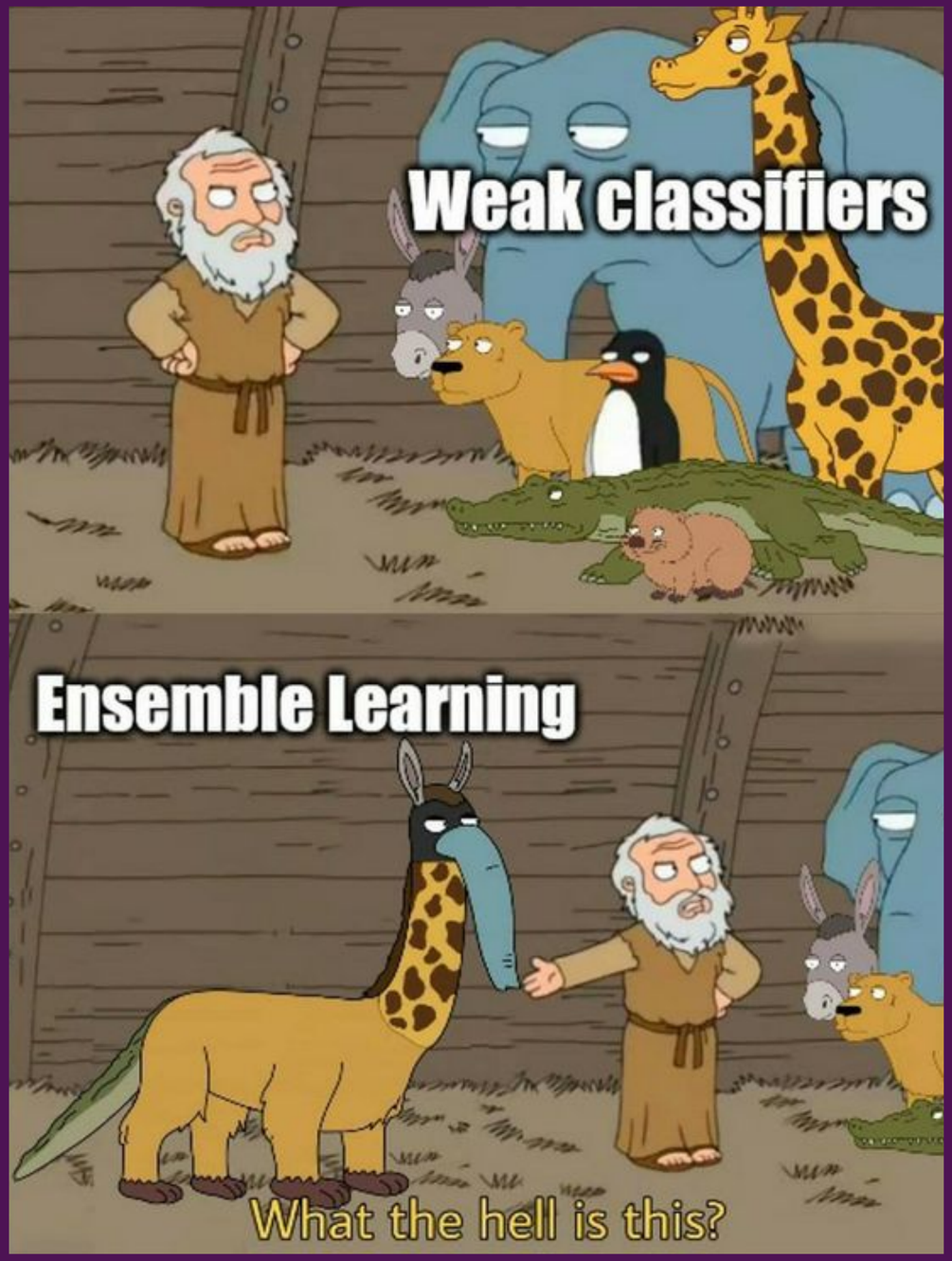
Cross validation	10% validation, 90% database from training set
Mean accuracy	0.6966

Check out our code!

https://github.com/LLynd/HackING23

See you next year!

https://github.com/LLynd/HackING23


Weak classifiers
Ensemble Learning
What the hell is this?