Revealing Master's theses structures using Machine Learning Methods

Susman M., Hint H., Šinkūnienė J., Leijen D. A. J.

The Bwrite Project

Academic Writing in the Baltic States: Rhetorical Structures through culture(s) and languages

Goal

- **Measure** and map the writing traditions of Estonian, Latvian and Lithuanian
- **Develop a research method** to determine which features of a text are related to genre, discipline, culture and experience
- **Provide empirical results** that allow writers and instructors of writing to better apply those text features for teaching and writing

Website: https://www.bwrite.ut.ee/

Introduction

Academic texts are expected to follow an organizational structure

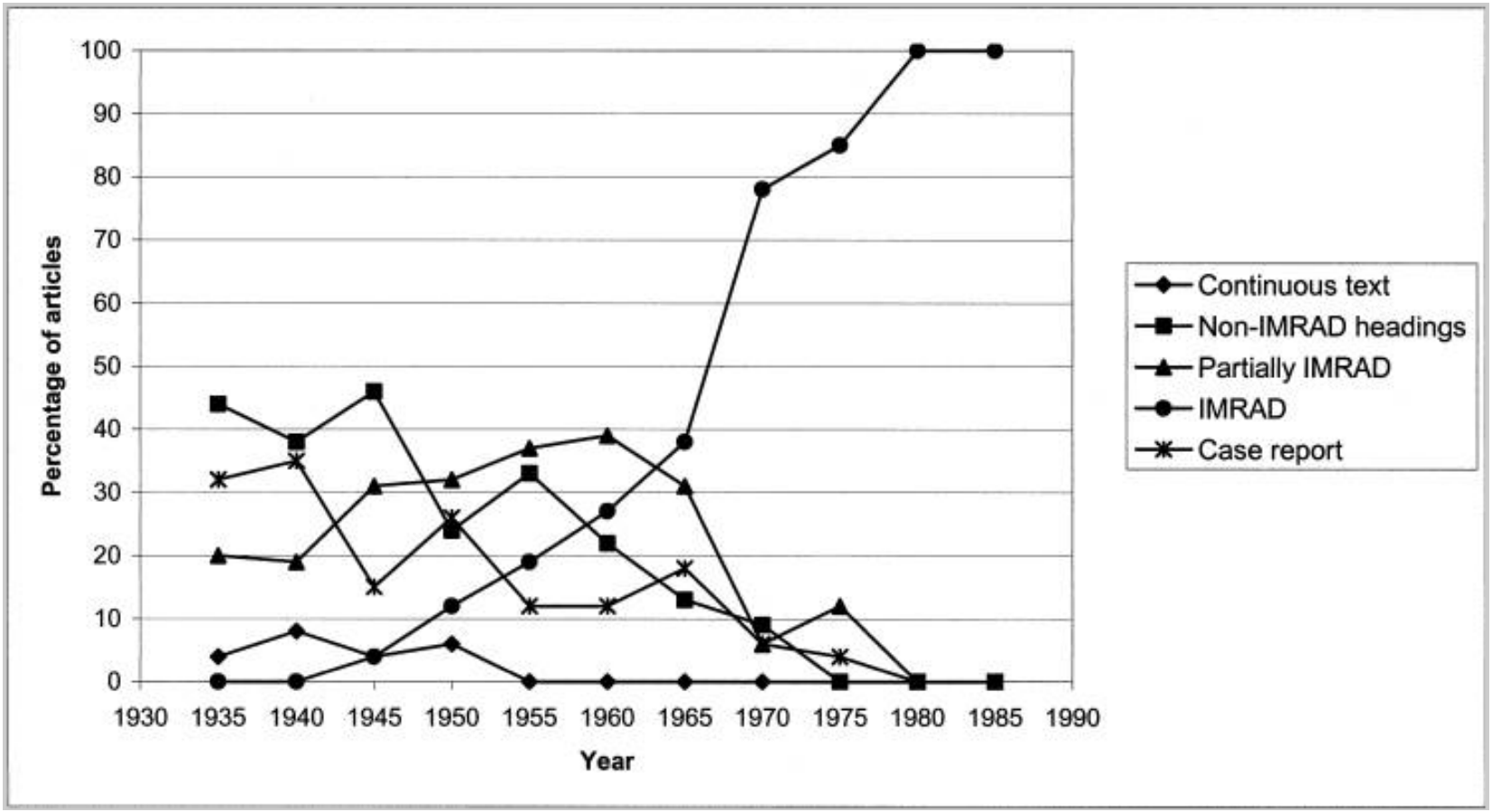
IMRaD structure as a standard in some communities (English, medicine, STEM, . . .)

Introduction, **M**ethod, **R**esults, and **D**iscussion

Research article genre, often imposed by the journal

Lin & Evans 2012

"IMRD is not an especially prevalent pattern in contemporary RA writing, so strict adherence to such a structure when conducting move-based or linguistic analyses is likely to result in incomplete or unrepresentative findings."



Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. Journal of the Medical Library Association : JMLA, 92(3), 364–367.

Questions

→ What do we observe in other, longer academic text genres, e.g. in master's theses? Can we detect IMRaD?

→ What do we observe in languages other than English?

→ What other structures the algorithm can detect?

→ (What structures or patterns do non-expert writers follow?)

Manual annotations

Our database: web-scraped academic texts (PDFs), all available genres, disciplines, journals

Random selection of 467 **theses** in Lithuanian and Estonian

Based on TOC's, manually annotated on a 4-way continuum:

- Not IMRaD (NI),
- Rather Not IMRaD (RNI),
- Rather IMRaD (RI),
- IMRaD (I)

Overview of the method

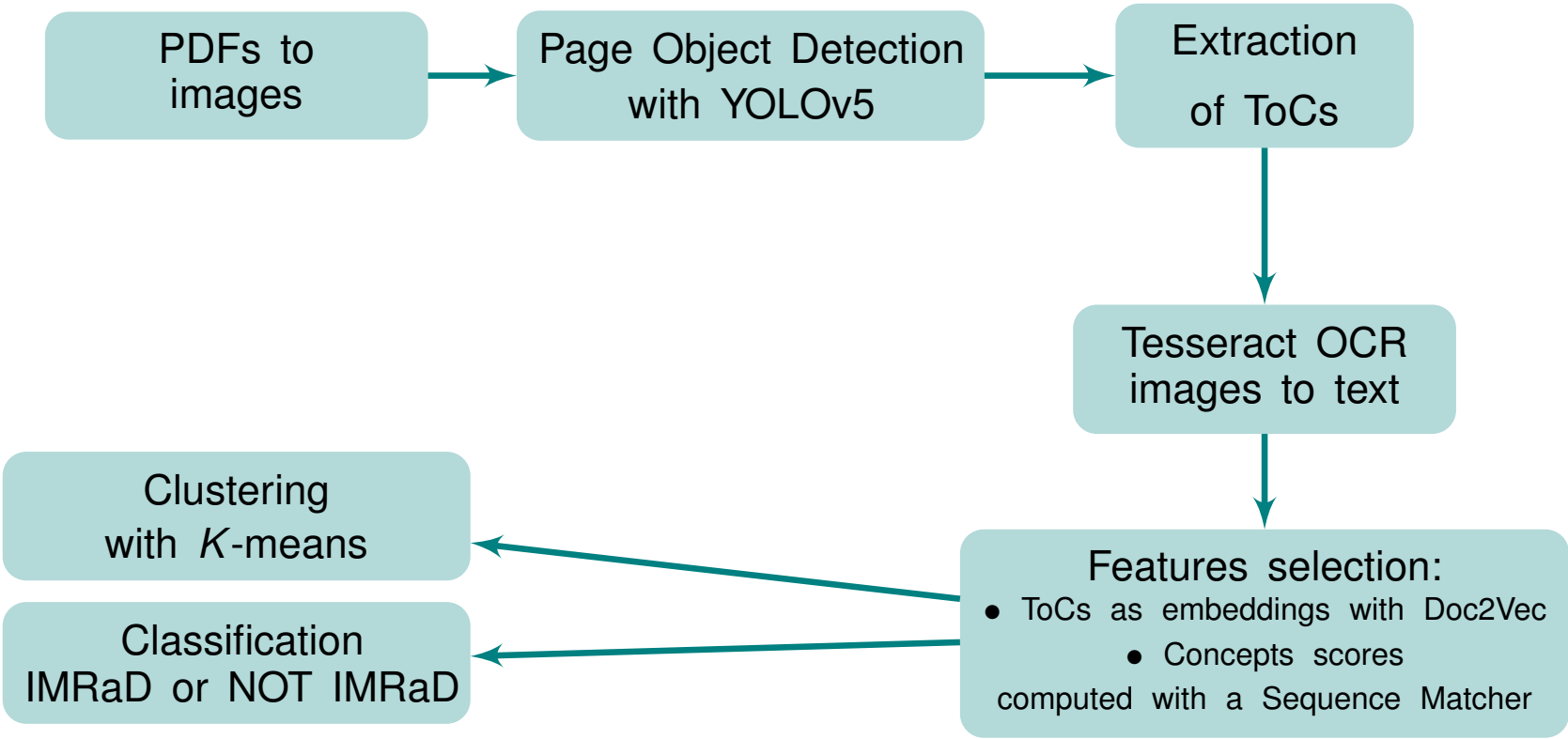
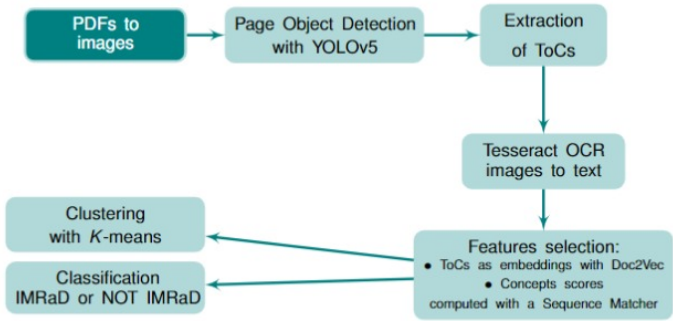
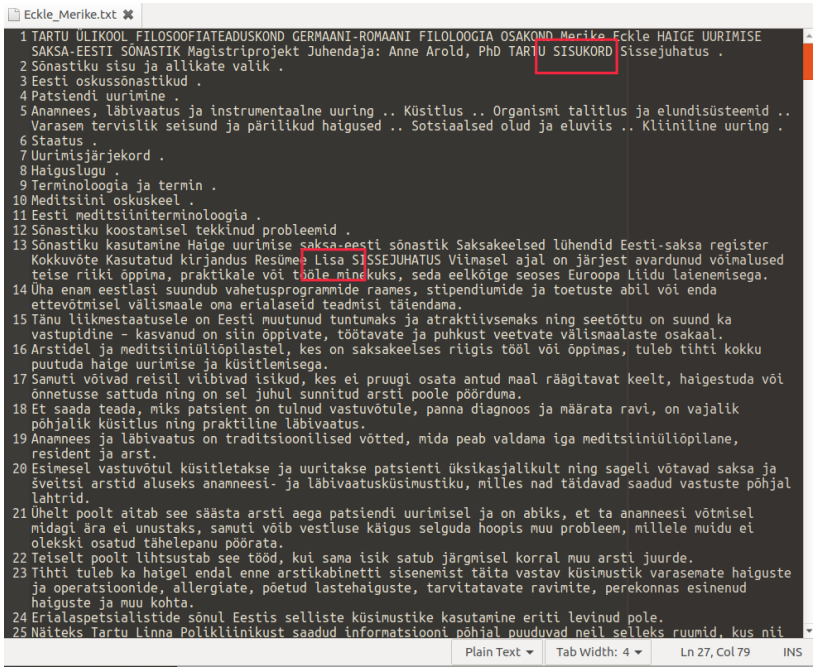


Figure: **Revealing the structure of Masters theses: the method's steps**

PDFs as images, why?

- Avoid loss of visual information
- Ease of extraction of relevant features



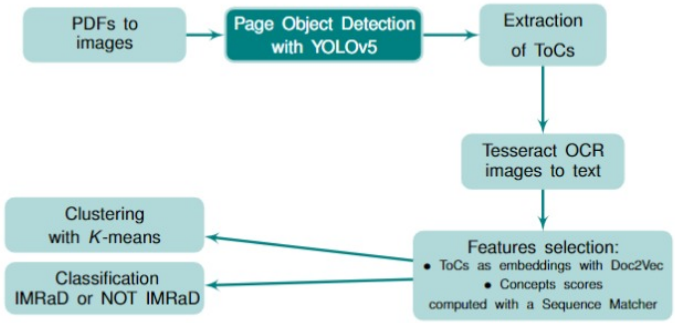

Eckle_Merike.txt

```
 1 TARTU ÜLIKOOL FILOSOOFIATEADUSKOND GERMAANI-ROMAANI FILOLOOGIA OSAKOND Merike Eckle HAIGE UURIMISE
   SAKSA-EESTI SÕNASTIK Magistriprojekt Juhendaja: Anne Arold, PhD TARTU SISUKORD Sissejuhatus .
 2 Sõnastiku sisu ja allikate valik .
 3 Eesti oskussõnastikud .
 4 Patsiendi uurimine .
 5 Anamnees, läbivaatus ja instrumentaalne uuring .. Küsitlus .. Organismi talitlus ja elundisüsteemid ..
   Varasem tervislik seisund ja pärilikud haigused .. Sotsiaalsed olud ja eluviis .. Kliiniline uuring .
 6 Staatus .
 7 Uurimisjärjekord .
 8 Haiguslugu .
 9 Terminoloogia ja termin .
10 Meditsiini oskuskeel .
11 Eesti meditsiiniterminoloogia .
12 Sõnastiku koostamisel tekkinud probleemid .
13 Sõnastiku kasutamine Haige uurimise saksa-eesti sõnastik Saksakeelsed lühendid Eesti-saksa register
   Kokkuvõte Kasutatud kirjandus Resümee Lisa SISSEJUHATUS Viimasel ajal on järjest avardunud võimalused
   teise riiki õppima, praktikale või tööle minekuks, seda eelkõige seoses Euroopa Liidu laienemisega.
14 Üha enam eestlasi suundub vahetusprogrammide raames, stipendiumide ja toetuste abil või enda
   ettevõtmisel välismaale oma erialaseid teadmisi täiendama.
15 Tänu liikmestaatusele on Eesti muutunud tuntumaks ja atraktiivsemaks ning seetõttu on suund ka
   vastupidine - kasvanud on siin õppivate, töötavate ja puhkust veetvate välismaalaste osakaal.
16 Arstidel ja meditsiiniüliõpilastel, kes on saksakeelses riigis tööl või õppimas, tuleb tihti kokku
   puutuda haige uurimise ja küsitlemisega.
17 Samuti võivad reisil viibivad isikud, kes ei pruugi osata antud maal räägitavat keelt, haigestuda või
   õnnetusse sattuda ning on sel juhul sunnitud arsti poole pöörduma.
18 Et saada teada, miks patsient on tulnud vastuvõtule, panna diagnoos ja määrata ravi, on vajalik
   põhjalik küsitlus ning praktiline läbivaatus.
19 Anamnees ja läbivaatus on traditsioonilised võtted, mida peab valdama iga meditsiiniüliõpilane,
   resident ja arst.
20 Esimesel vastuvõtul küsitletakse ja uuritakse patsienti üksikasjalikult ning sageli võtavad saksa ja
   šveitsi arstid aluseks anamneesi- ja läbivaatusküsimustiku, milles nad täidavad saadud vastuste põhjal
   lahtrid.
21 Ühelt poolt aitab see säästa arsti aega patsiendi uurimisel ja on abiks, et ta anamneesi võtmisel
   midagi ära ei unustaks, samuti võib vestluse käigus selguda hoopis muu probleem, millele muidu ei
   oleksi osatud tähelepanu pöörata.
22 Teiselt poolt lihtsustab see tööd, kui sama isik satub järgmisel korral muu arsti juurde.
23 Tihti tuleb ka haigel endal enne arstikabinetti sisenemist täita vastav küsimustik varasemate haiguste
   ja operatsioonide, allergiate, põetud lastehaiguste, tarvitatavate ravimite, perekonnas esinenud
   haiguste ja muu kohta.
24 Erialaspetsialistide sõnul Eestis selliste küsimustike kasutamine eriti levinud pole.
25 Näiteks Tartu Linna Polikliinikust saadud informatsiooni põhjal puuduvad neil selleks ruumid, kus nii
```

Plain Text | Tab Width: 4 | Ln 27, Col 79 | INS

SISUKORD

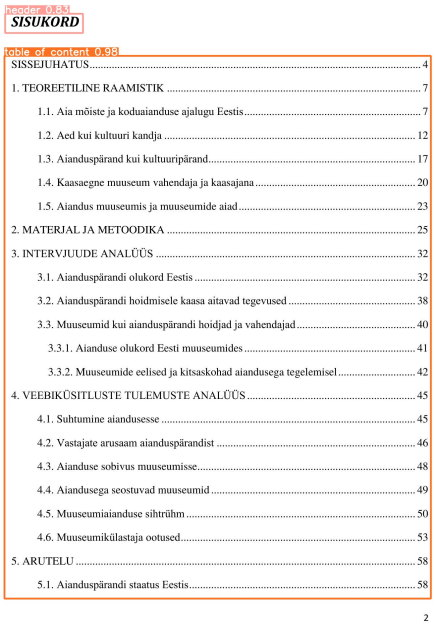
YOLO [6]



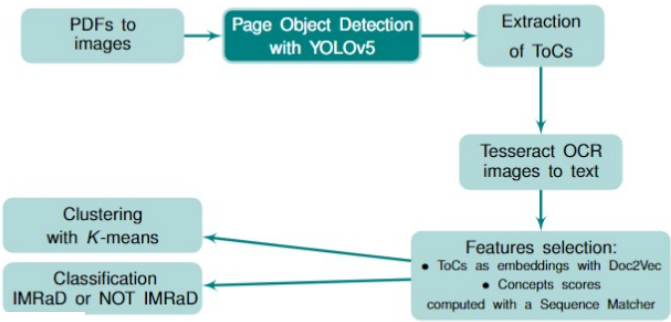
- Object detection algorithm
- Images divided into a grid. Objects detected within each grid cell
- Draws bounding boxes around regions of interest AND attribute them a label
- Used in our method: YOLOv5

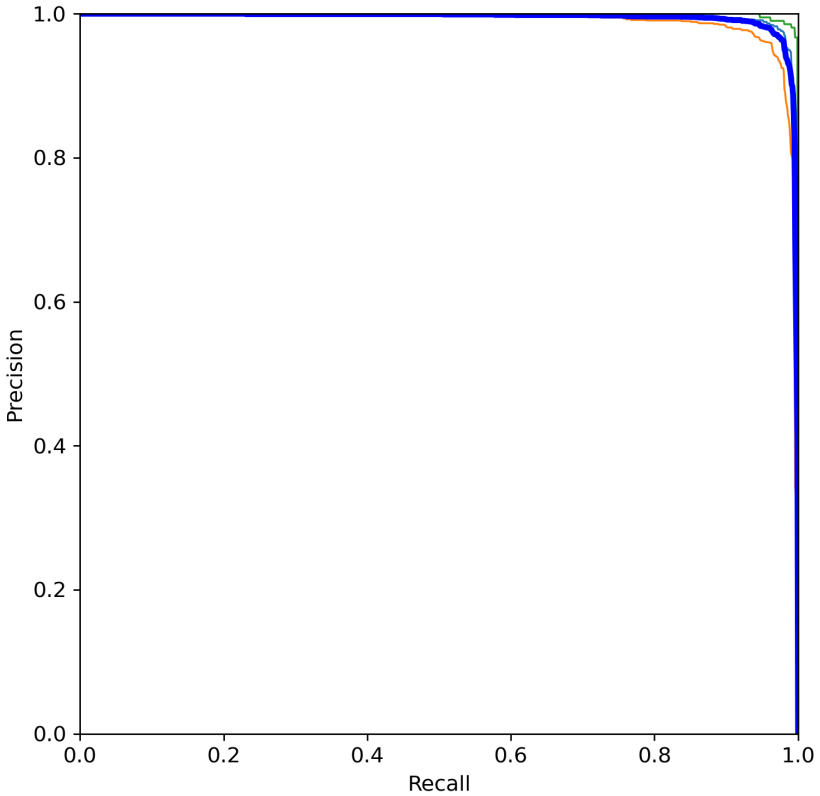
Training on our data

- 3 classes: ‘headers’, ‘ToCs’, ‘body’
- Training set: 1400 images, testing set: 757 images
- Model: YOLOv5 small [6]
- Trained **from scratch** on GPU - 150 epochs

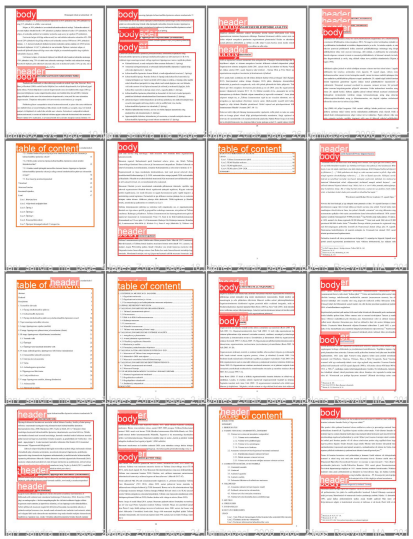

header 0.83
SISUKORD
table_of_content 0.98
SISSEJUHATUS ... 4
1. TEOREETILINE RAAMISTIK ... 7
1.1. Aia mõiste ja koduaianduse ajalugu Eestis ... 7
1.2. Aed kui kultuuri kandja ... 12
1.3. Aianduspärand kui kultuuripärand ... 17
1.4. Kaasaegne muuseum vahendaja ja kaasajana ... 20
1.5. Aiandus muuseumis ja muuseumide aiad ... 23
2. MATERJAL JA METOODIKA ... 25
3. INTERVJUUDE ANALÜÜS ... 32
3.1. Aianduspärandi olukord Eestis ... 32
3.2. Aianduspärandi hoidmisele kaasa aitavad tegevused ... 38
3.3. Muuseumid kui aianduspärandi hoidjad ja vahendajad ... 40
3.3.1. Aianduse olukord Eesti muuseumides ... 41
3.3.2. Muuseumide eelised ja kitsaskohad aiandusega tegelemisel ... 42
4. VEEBIKÜSITLUSTE TULEMUSTE ANALÜÜS ... 45
4.1. Suhtumine aiandusesse ... 45
4.2. Vastajate arusaam aianduspärandist ... 46
4.3. Aianduse sobivus muuseumisse ... 48
4.4. Aiandusega seostuvad muuseumid ... 49
4.5. Muuseumiaianduse sihtrühm ... 50
4.6. Muuseumikülastaja ootused ... 53
5. ARUTELU ... 58
5.1. Aianduspärandi staatus Eestis ... 58
2

Results YOLOv5

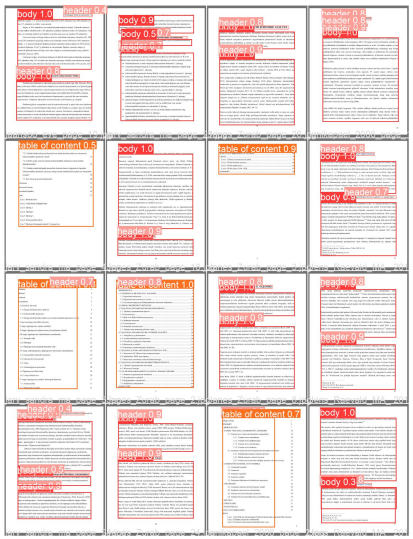




Precision-Recall curve



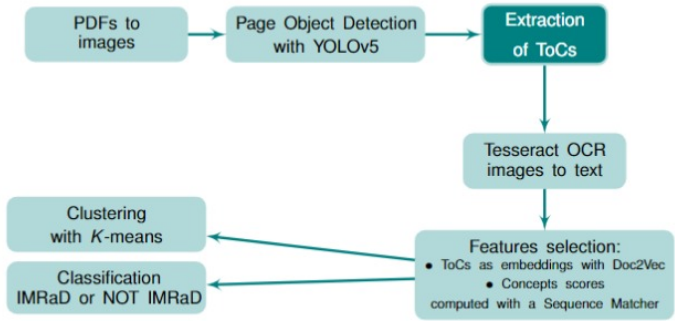
Labels

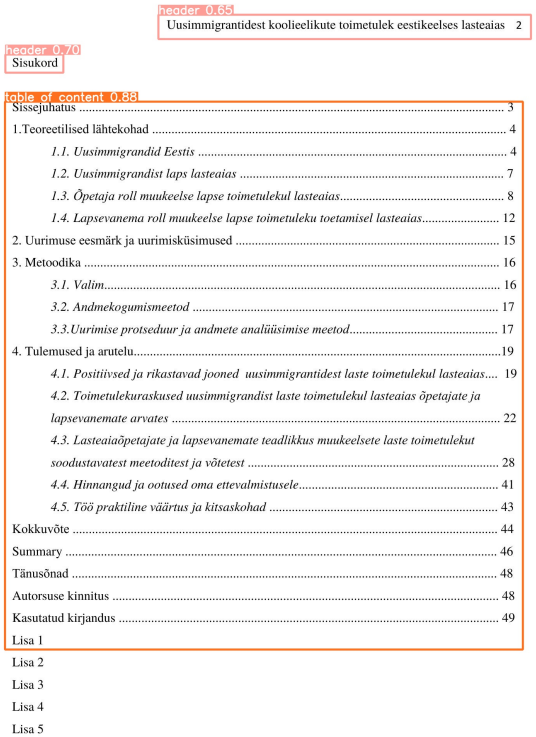


Predictions

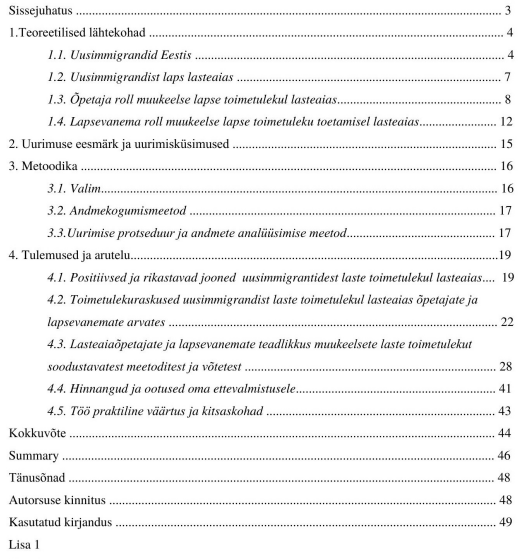
Extraction: bounding boxes

- Extract of bounding boxes with coordinates from YOLOv5

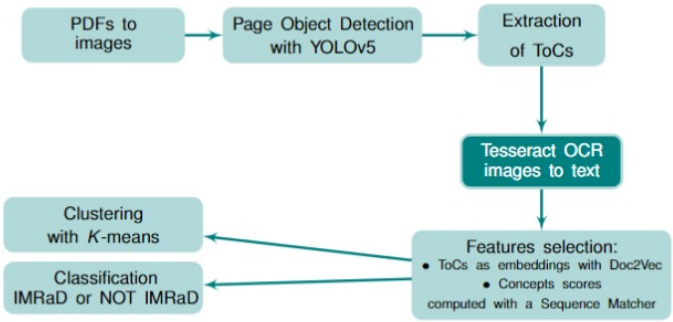




Crop



Object Character Recognition



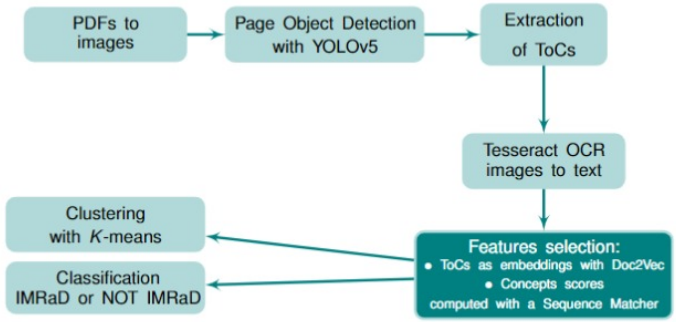
- Transforms text embedded on images to strings

From OCR output to Pre-processed text data

```
SESSBIUATUSseersccessassoreassesscecarreneaaceemtaragaasiaeaeemsiaarmaecaeaeeenneatY4|-
REORBETILINERAAMISTIKscccsssssssssssssssanrasousesnnmniasnounenumsnnnenrasanuemmennnenraue71.1.AiamGistejak
INTERVIUUDE.ANALUUSvisisnnsisinunvisnunnnniennnsacannannnnenunvineneamuneraivs32SeL.AlandusparandsOlUKOdROS
SUbtUMINEAlandUSeSSCisssxsscsssreerresemereneemeRER454.2.Vastajatearusaamaiandusparandist...........:cecee
```

sissejuhatus 0 cece 4 1 teoreetiline raamistir o 7 1.1 aia mdiste ja koduaianduse ajalugu eestis 0.0 7 12 ag kot kultit wand username nn 12 1,3 aiandusparand kui kultuuniparan mares 17 1.4 kaasaegne muuseum vahendaja ja kaasajama 20 1.5 aiandus muuseumis ja muuseumide aiad 0.0 23 2 o 25 3 interviuude-analuus wrasse 32 3.1 aiandusparandi olukord eestis 32 3.2 aiandusparandi hoidmisele kaasa aitavad tegevused ce 38 3.3 muuseumid kui aiandusparandi hoidjad ja vahendajad rece 40 3.3.1 aianduse olukord eesti muuseumides 00.0 41 3.3.2 muuseumide eelised ja kitsaskohad aiandusega tegelemisel 0 0 ce 42 4 veebikusitluste tulemuste analuus 45 4.1 suhtumine aiandusesse 4.2 vastajate arusaam aiandusparandist 0 cs 46 4:3 alanduse sobivus muuscuimni ssc 48 4.4 aiandusega seostuvad muuscumi 0 ee ee 49 4.5 muuseumiaianduse sihtrthm serves nncorneuers 50 4.6 muuseumikiilastaja o0tused 53 5 aru tel 58 5.1 aianduspdrandi staatus eestis c cc 58 5.2 aiandusparandi vaartus 60 502 12 pudkus an staap lds ences eerie er eaten sistem ere 60 51222 malgstused sexcecscspte rarest eeseerci rene a ere 62 5.2.8 toidtt kas valamgsccccnercsn ecm ne ee 63 5.3 mandusparand muuseuins ss 64 5.3.1 muuseumiaianduse sihtrti hit 64 5.3.2 muuseumisianduse voimalused renames 67 5.3.3 10ll mre 69 5.4 jareldused jasoovitused 72 ror ruy ote 75 kasutatud kirjandus ja allikad 78 hitery ud ja kusituss ance 78 tut lal cag rsa 78 kot g08 ee se et 78 taupe 1 4 rag sions corte sees etree eerie etree enema eres 86 lis ad sinterensuom sin eerie enw eerie ree 89 lisa 1 fotod eesti vabashumuuseumi aedadest ja aiandustiritustest cece 89 lisa 2 poolstruktureeritud intervjuu kava cece cess 91 lisa 3 ankeetkiisimustiku saanud muuseumide nimekiri ja ktisimustiku ktisimused 92 lisa 4 saadetud ankeetktisitluse vastused 00.0 94 lisa 5 veebiktisimustik muuseum kui aiakultuuriparandi hoidja ja vahendaja ' 98 summary envy 101

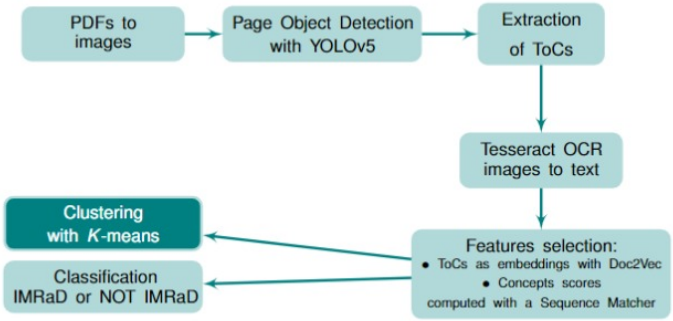
Feature selection



- ToCs $\rightarrow$ vectors of numerical values with Doc2Vec.
- Sequence Matcher [8]: compares and identifies similarities in given pairs of input strings.
- IMRaD concepts $\rightarrow$ concept scores.

IMRaD concepts	*IMRaD* words in Baltic languages
Introduction	'sissejuhatus', 'ivada', 'izanga'
Literature review	'valdkonna ulevaade', 'kirjanduse ulevaade', 'teoreetiline ulevaade', 'teoreetiline taust', 'teoreetiline raamistik', 'literaturos apzvalga'
Methods	'metoodika', 'meetod', 'metodai', 'metodologija', 'metodika', 'metodine', 'metodas'
Results	'tulemused', 'uurimistulemused', 'rezultatai', 'duomenu analize ir rezultatu apzvalga'
Discussion	'arutelu', 'diskussioon', 'jareldused', 'aptarimas', 'isvados'

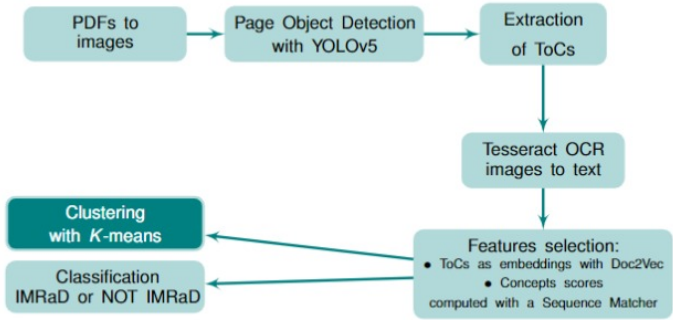
Clustering



K-means clustering [4]

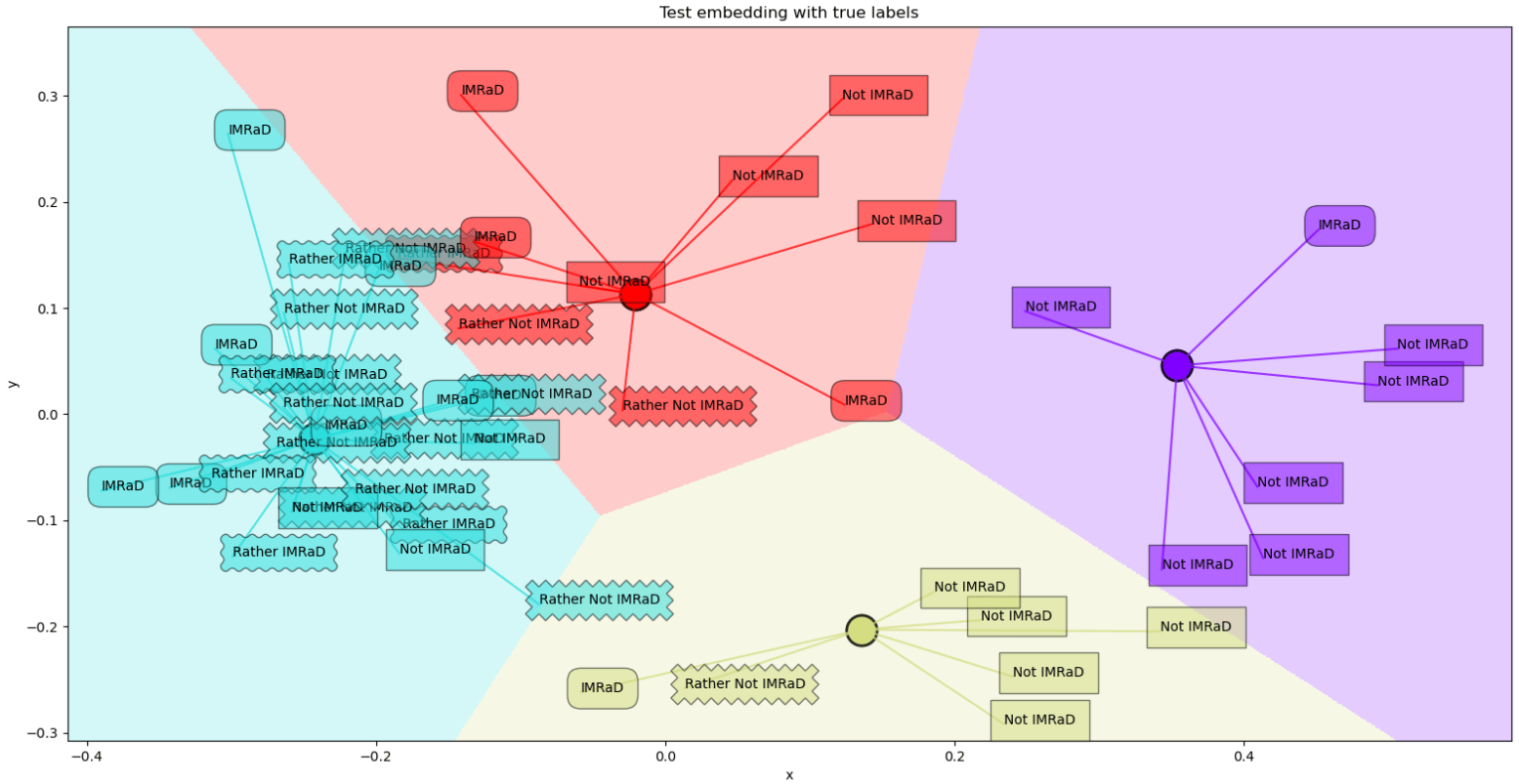
- Unsupervised method which groups similar data instances together
- Measures features selection's relevance
- 467 documents organized in four classes:
  - 98 *IMRaD* files
  - 171 *Not IMRaD* files
  - 105 *Rather IMRaD* files
  - 93 *Rather Not IMRaD* files
- Tested with $k = 4$ and $k = 2$
- Score embeddings, Text embeddings, Text and score embeddings

Clustering results

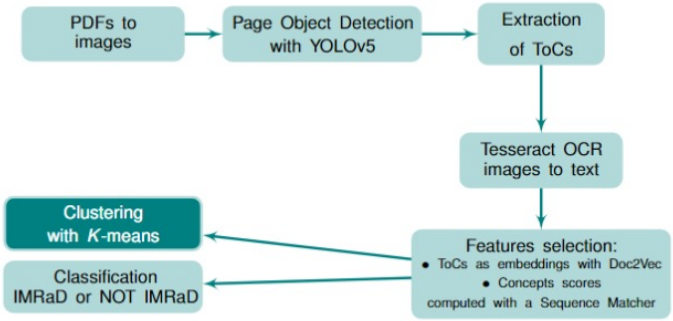


$2D$ cluster representation

$k = 4$

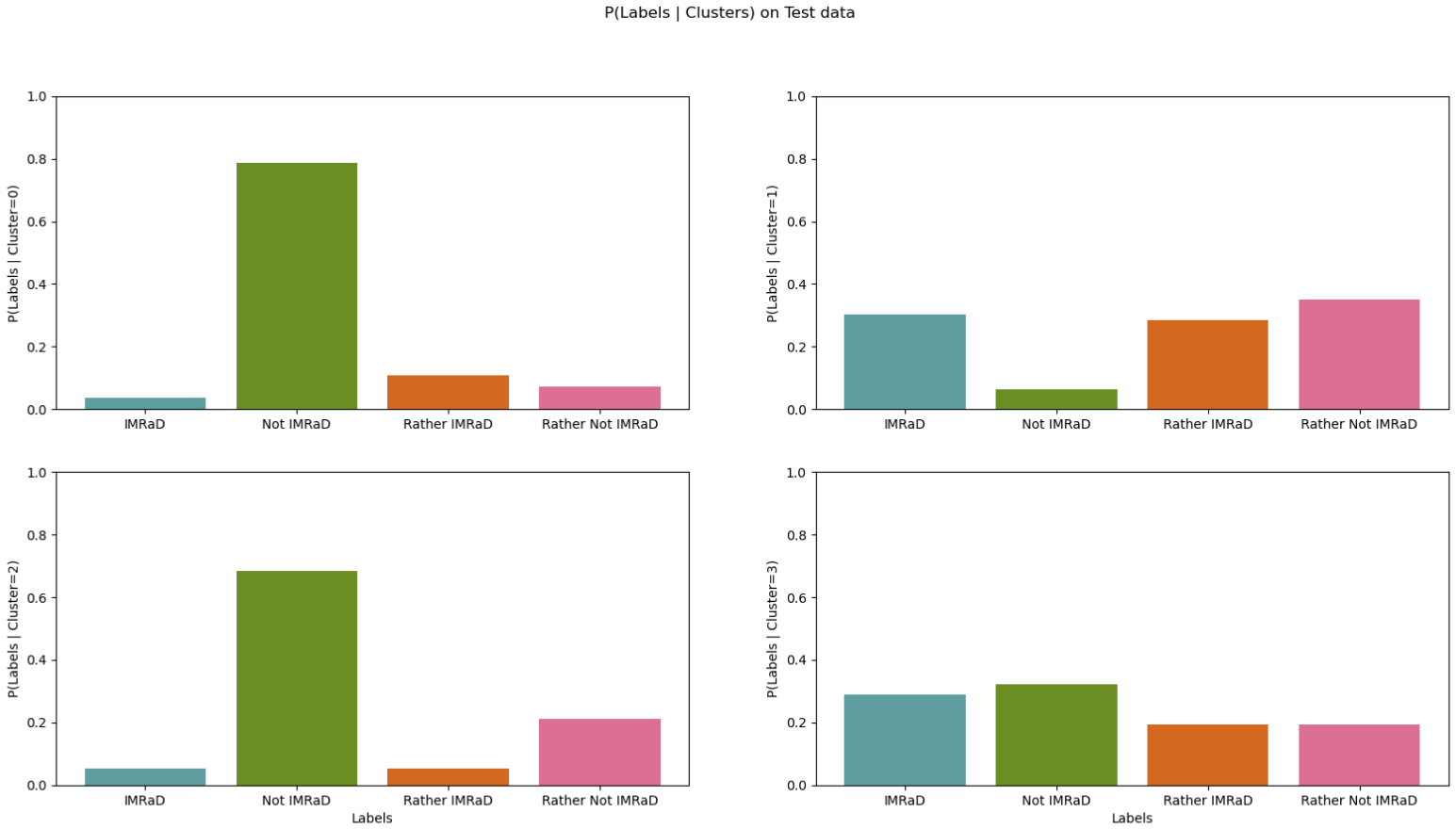


Clustering results

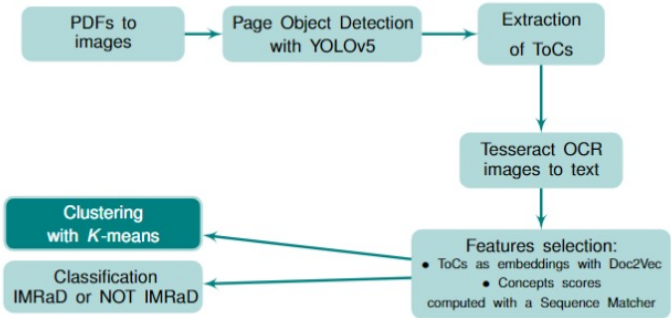


Histogram representation of distribution of classes across clusters.

$k = 4$

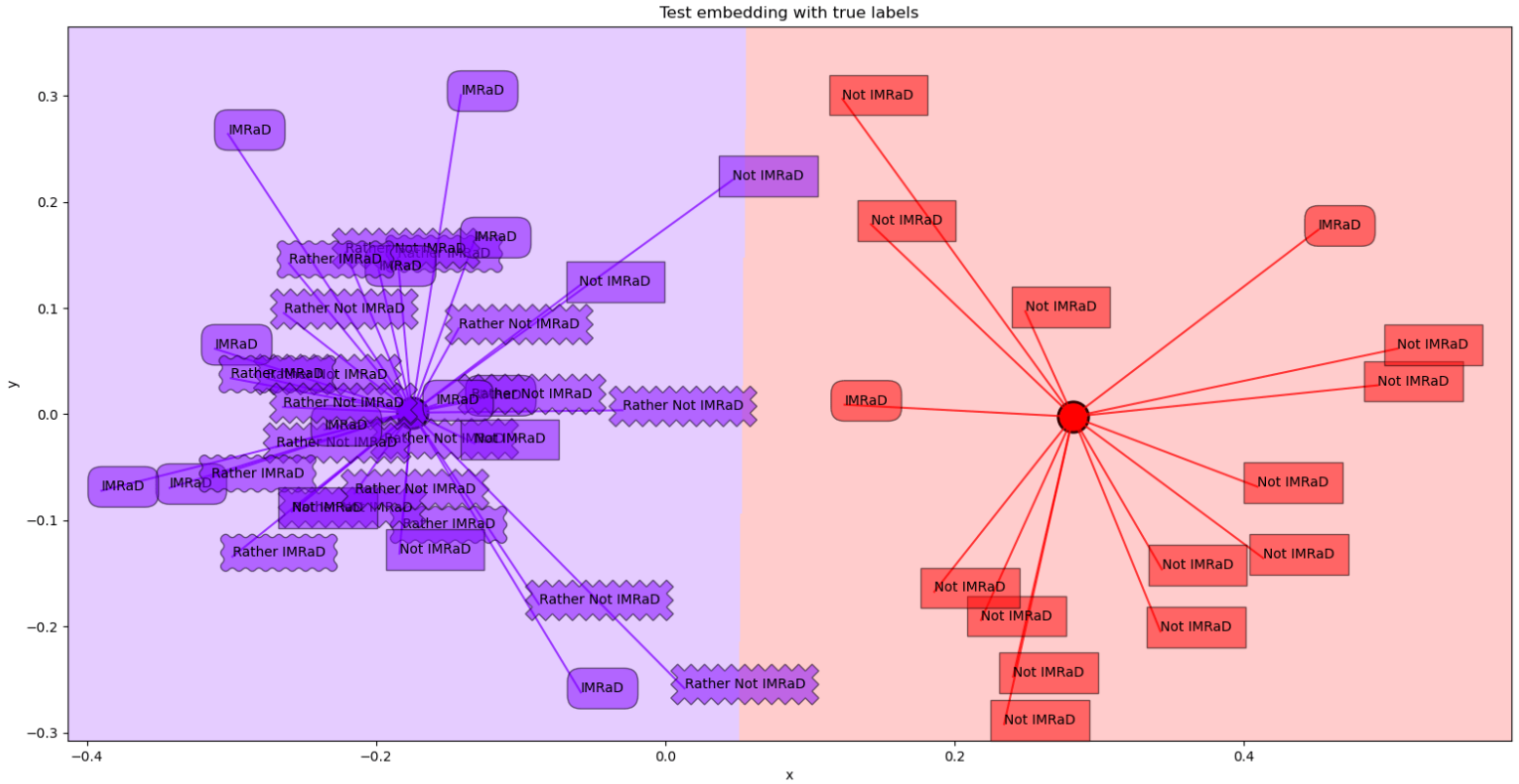


Clustering results

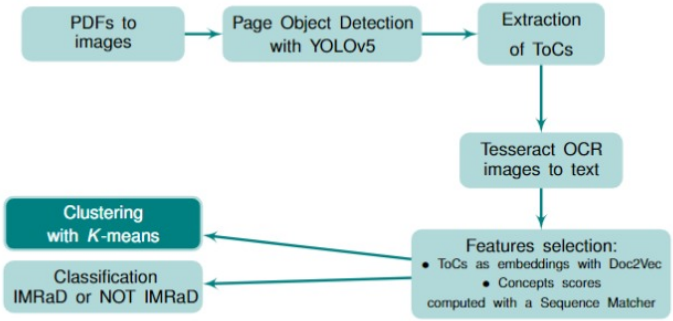


$2D$ cluster representation

$k = 2$

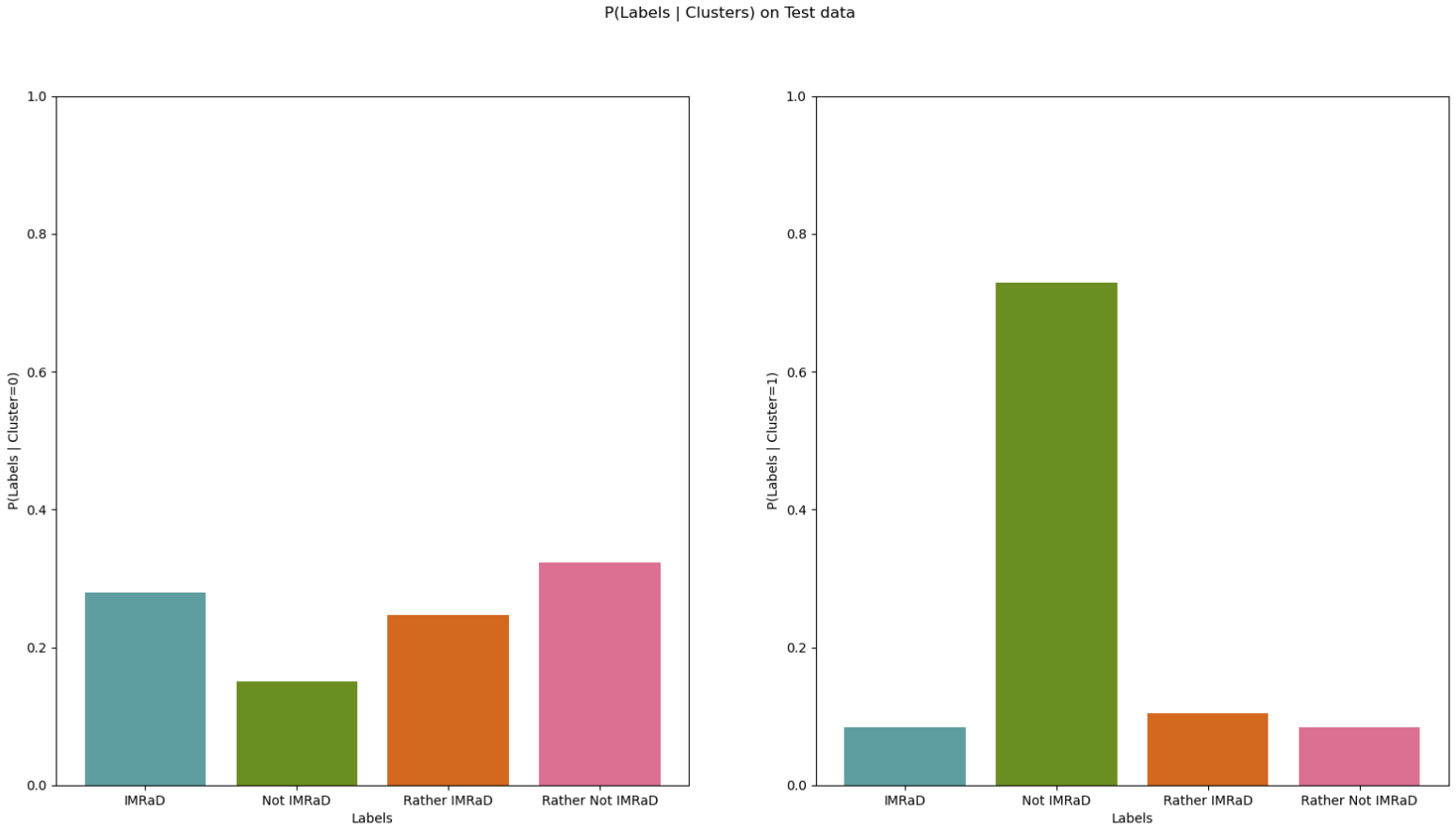


Clustering results

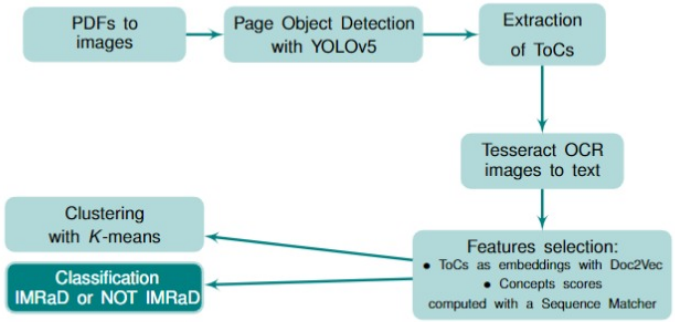


Histogram representation of distribution of classes across clusters.

$k = 2$



Classification

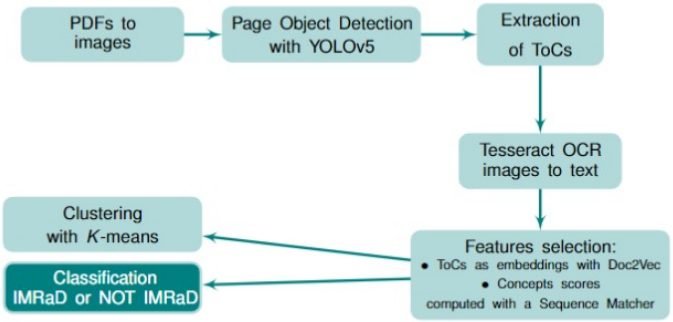


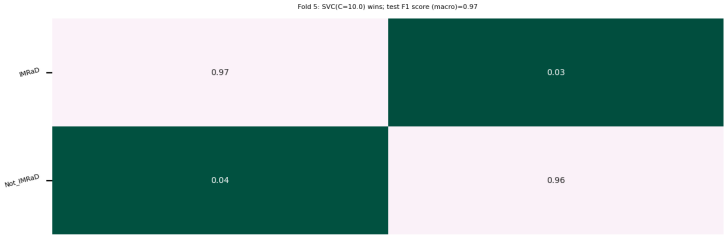
Def IMRaD			Def Not IMRaD
Def IMRaD	Rather IMRaD	Rather Not IMRaD	Def Not IMRaD
Def IMRaD	Rather IMRaD + Rather Not IMRaD		Def Not IMRaD
Def IMRaD + Rather IMRaD		Def Not IMRaD + Rather Not IMRaD	
Def IMRaD + Rather IMRaD + Rather Not IMRaD			Def Not IMRaD

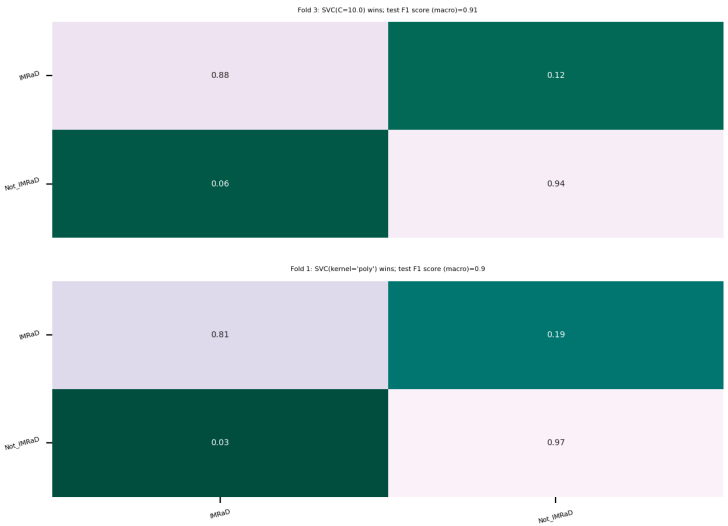
Model selection

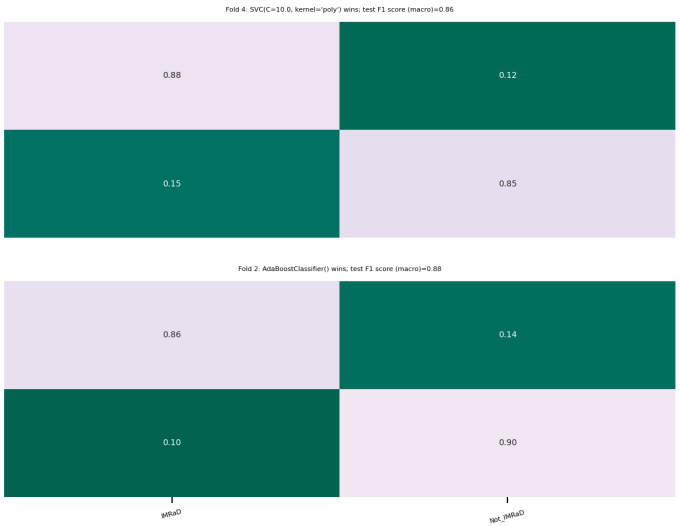
- 6 different models
- Variation in hyperparameters $\rightarrow$ total model tested: 32

Results classification

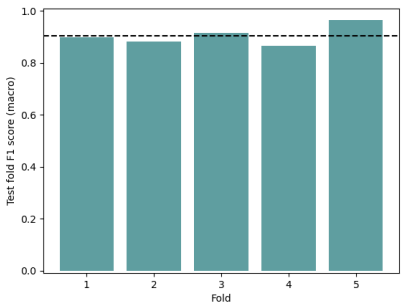








- Mean macro f1 score: 90%
- Standard deviation: 0.02



Discussion

- Pros and cons of object detection methods:
  - Usable with PDF documents without losing important graphical information and helps reduce the amount of data needing to be processed
  - Identification of headers in text remains challenging
- The possibility of detecting the organizational structure of a document independently of its language
- IMRaD
  - Not binary to a human annotator,
  - *Rather IMRaD* and *Rather Not IMRaD* documents not different enough from *IMRaD* files to a machine learning algorithm

References I

Kaplan, Robert B.
‘Cultural thought patterns in inter-cultural education’.
Language learning, 16(1-2):1–20, 1966.

Redmon, J., Divvala, R., Girshick, R., Farhadi, A.
‘You only look once: Unified, real-time object detection’.
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788, 2016.

Bochkovskiy, A. and Wang, C.-Y. and Liao, H.-Y. M.
‘Yolov4: Optimal speed and accuracy of object detection’.
arXiv preprint arXiv:2004.10934, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and Duchesnay, E.
‘Scikit-learn: Machine learning in Python.’
the Journal of machine Learning research, 12, 2825–2830, 2011.

References II

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., and Varoquaux, G.
'API design for machine learning software: experiences from the scikit-learn project.'
arXiv preprint arXiv:1309.0238., 2013.

Jocher, G., Nishimura, K., Mineeva, T. and Vilariño, R.
YOLOv5, 2020
https://github.com/ultralytics/yolov5

Hoffstaetter, S., Bochi, J., Lee, M., Kistner, L., Mitchell, R., Cecchini, E., Hagen, J., Morawiec, D., Bedada, E., and Akyüz, U.
Pytesseract
https://github.com/madmaze/pytesseract

References III

- Python Software Foundation
  Difflib - SequenceMatcher
  https://github.com/python/cpython/blob/master/Lib/difflib.py
- Cartucho, J., Ventura, R., Veloso, M.
  Robust Object Recognition Through Symbiotic Deep Learning In Mobile Robots
  In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2336–2341, 2018.