Alma Mater Studiorum – Università di Bologna

REVEALING CONTESTED MEMORY

Automatic sensitive content detection in colonial photographic archives

Digital Humanities and Digital Knowledge
Dissertation in *Semantic Digital Libraries*

Defended by Orsola Maria Borrini
Supervisor Prof. Giovanni Colavizza
Co-supervisor Prof. Charles Jeurgens (University of Amsterdam)

Session III
Academic Year 2022/2023

Project phases

01 SENSITIVE CONTENT DEFINITION

Development of a working context-specific definition and a taxonomy used as an aid to the annotation process

02 DATA AND METHODS

The data and methods used for the development of the Machine Learning pipeline

03 RESULTS

Exploration of the results and error analysis

04 DISCUSSION

Discussion on the work done and possible future avenues of research

problem definition → *data collection*
data annotation
training

01 Sensitive content definition

- **No fixed definition** of sensitive content, depends on the purview of inquiry
- In the GLAM sector, institutions are addressing the issue through cautionary statements

- Need to address context-specific features:

Colonialism

No clear definition, depends on the goals and assets of the specific case

We accept Osterhammel's definition encompassing all the fundamental aspects

Photography

Inherent problematic aspects of photography (Sontag, Crane)

Used as instrument in colonial dominions

Archival institutions

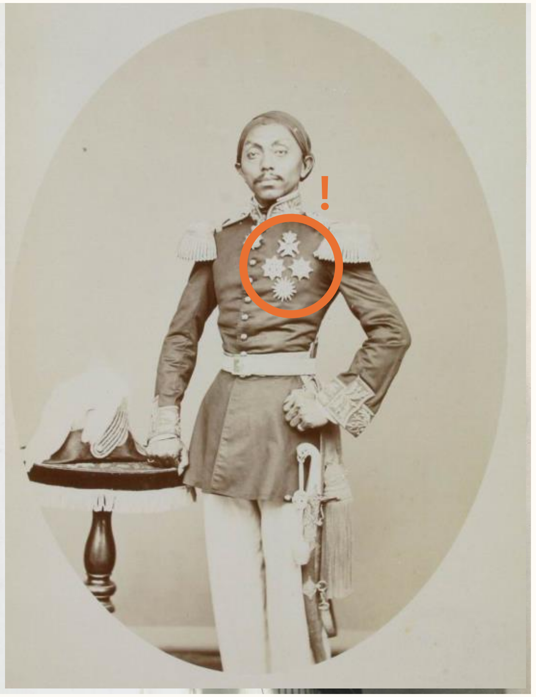
Postmodern approach: archives as active sites of contested power

Traditionally highly dominated by Western perspectives

- Premises and limitations: only visual content, no intersectionality
- Definition of three different degrees of recognisability

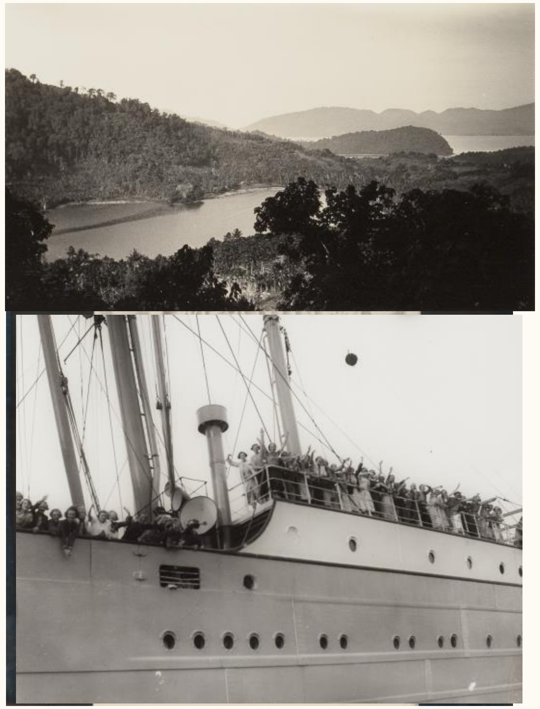
01 Sensitive content definition

1. **Sensitive content**: content which is more explicit and easily recognisable as immediately sensitive (either reiterates discriminatory beliefs, has violent graphic content or symbols and references to the colonial context)
2. **Dubious content**: unclear content which would benefit the most from the contribution of Indigenous communities and experts to the workflow (production context is ambiguous)
3. **Not-sensitive content**: content which does not display any clear or explicitly sensitive feature



sensitive

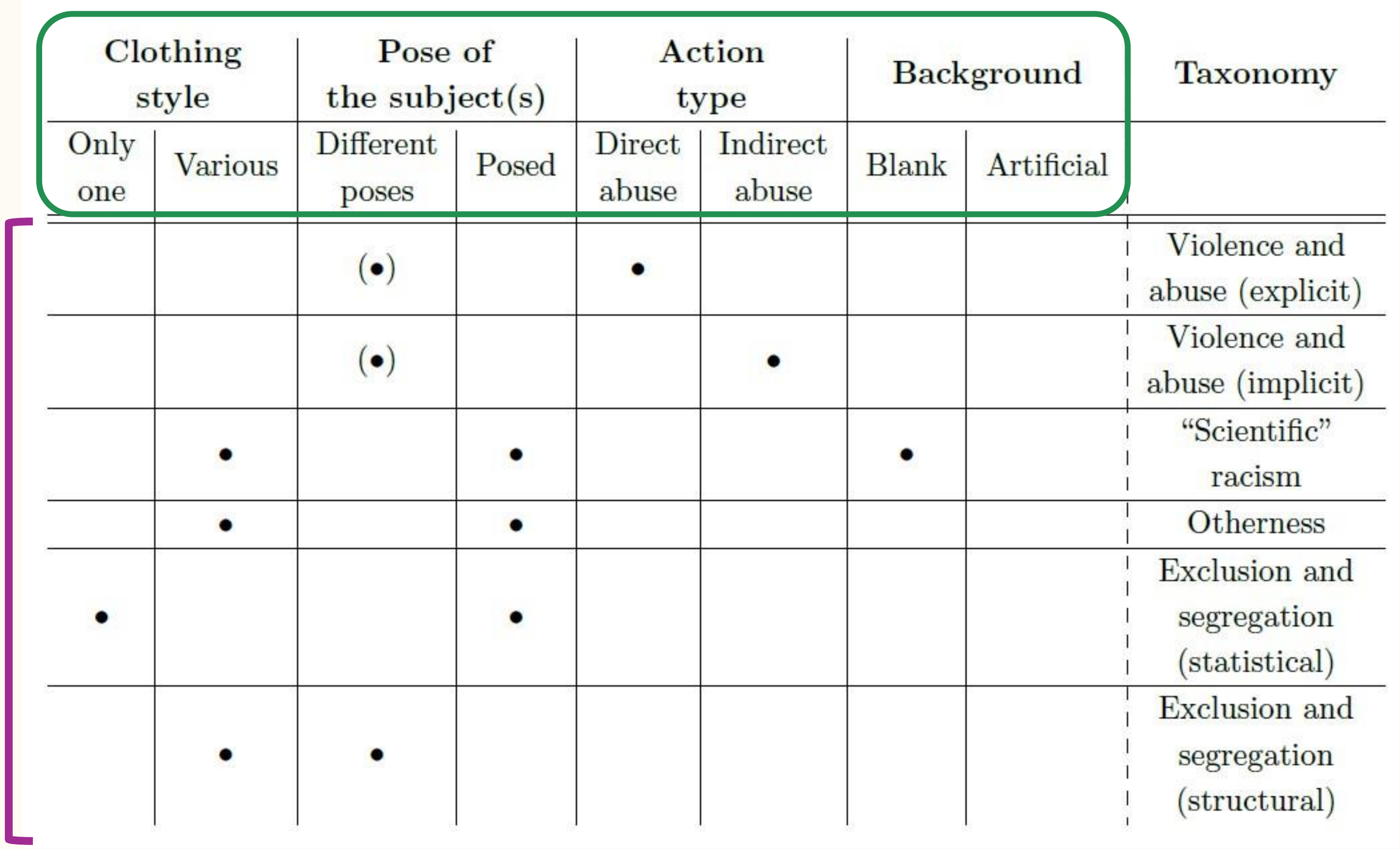
dubious



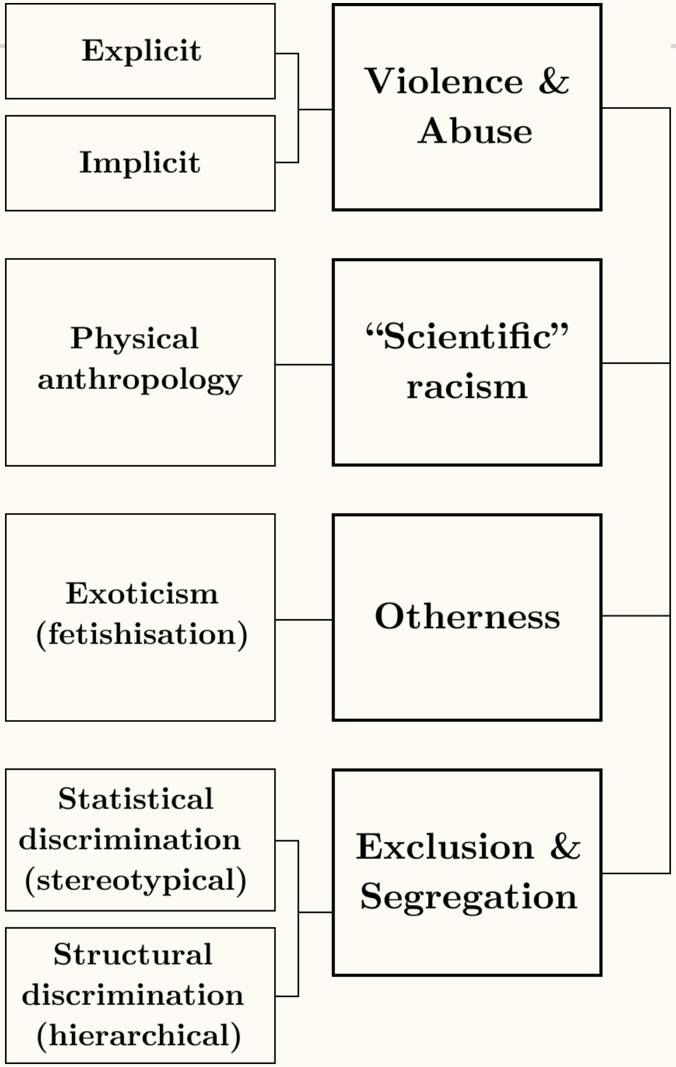
not-sensitive

01 Taxonomy development

Observation of triggering phenoptypical characteristics (**abstract categories**) and selection of the **most relevant combinations**



Clothing style		Pose of the subject(s)		Action type		Background		Taxonomy
Only one	Various	Different poses	Posed	Direct abuse	Indirect abuse	Blank	Artificial	
		(•)		•				Violence and abuse (explicit)
		(•)			•			Violence and abuse (implicit)
	•		•			•		"Scientific" racism
	•		•					Otherness
•			•					Exclusion and segregation (statistical)
	•	•						Exclusion and segregation (structural)



Explicit

Implicit

Violence & Abuse

Physical anthropology

"Scientific" racism

Exoticism (fetishisation)

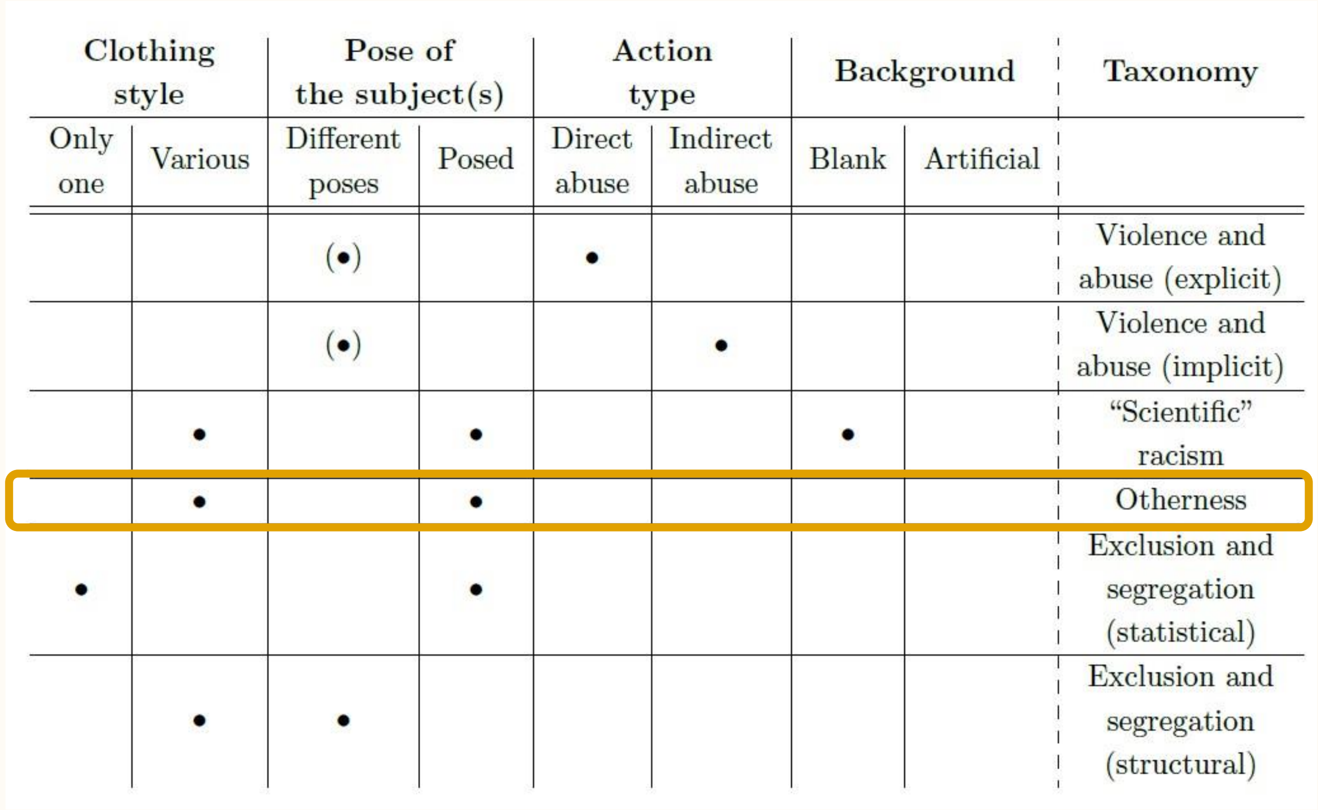
Otherness

Statistical discrimination (stereotypical)

Structural discrimination (hierarchical)

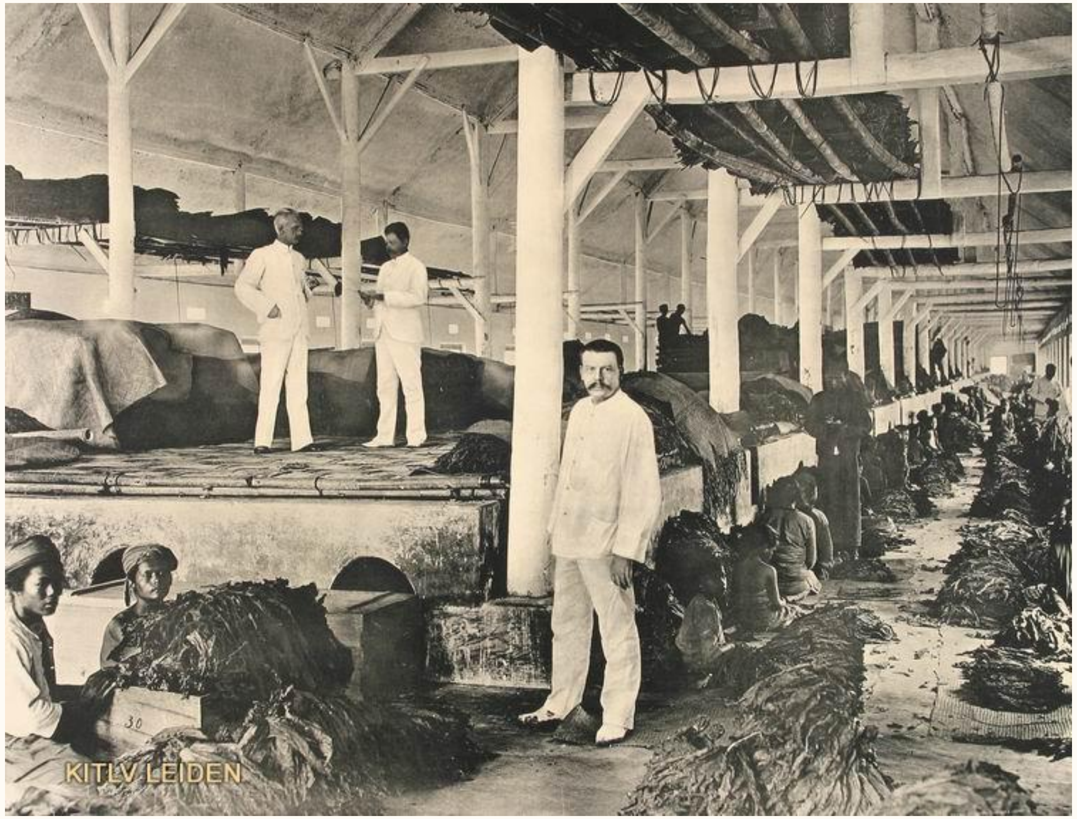
Exclusion & Segregation

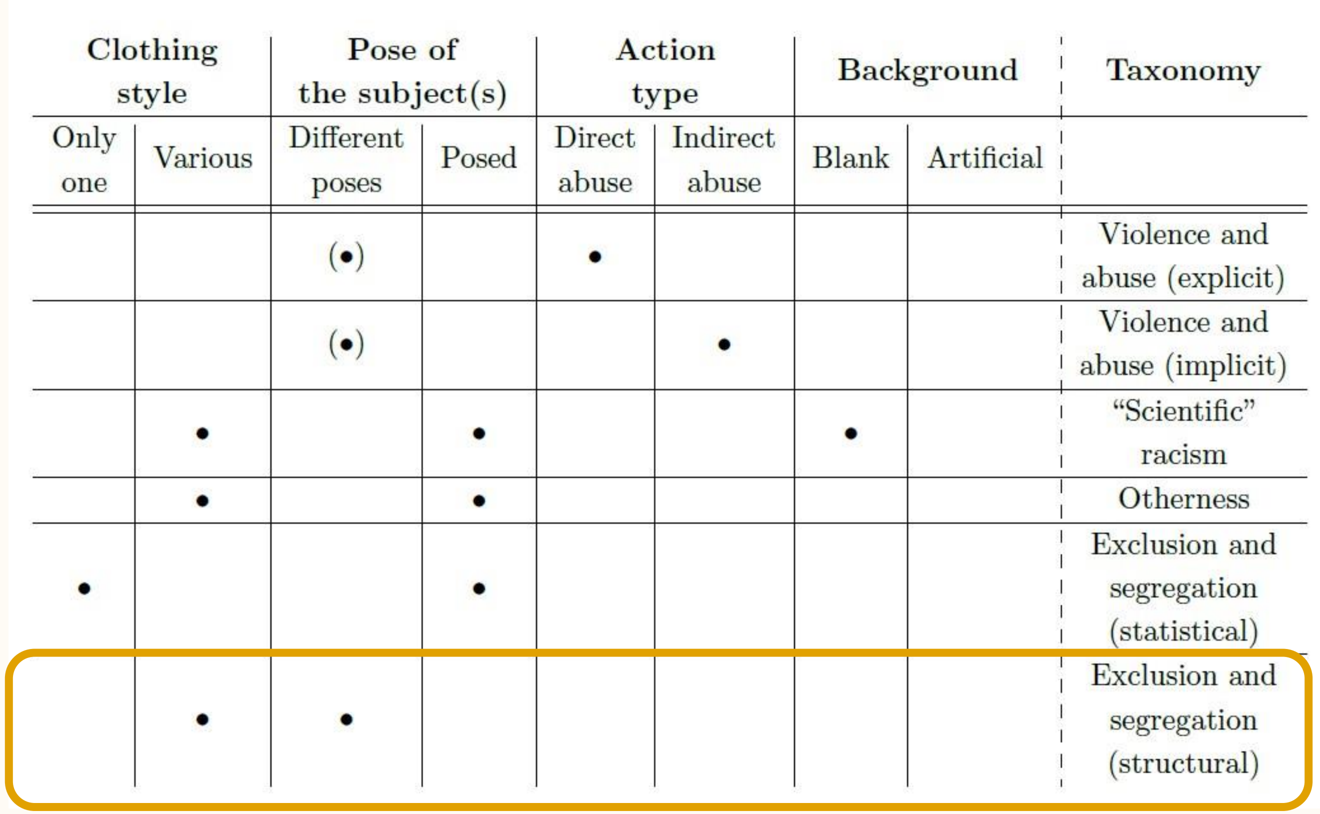
01 Taxonomy development



Clothing style		Pose of the subject(s)		Action type		Background		Taxonomy
Only one	Various	Different poses	Posed	Direct abuse	Indirect abuse	Blank	Artificial	
		(•)		•				Violence and abuse (explicit)
		(•)			•			Violence and abuse (implicit)
	•		•			•		“Scientific” racism
	•		•					Otherness
•			•					Exclusion and segregation (statistical)
	•	•						Exclusion and segregation (structural)

01 Taxonomy development





Clothing style		Pose of the subject(s)		Action type		Background		Taxonomy
Only one	Various	Different poses	Posed	Direct abuse	Indirect abuse	Blank	Artificial	
		(•)		•				Violence and abuse (explicit)
		(•)			•			Violence and abuse (implicit)
	•		•			•		“Scientific” racism
	•		•					Otherness
•			•					Exclusion and segregation (statistical)
	•	•						Exclusion and segregation (structural)

02 Data and methods

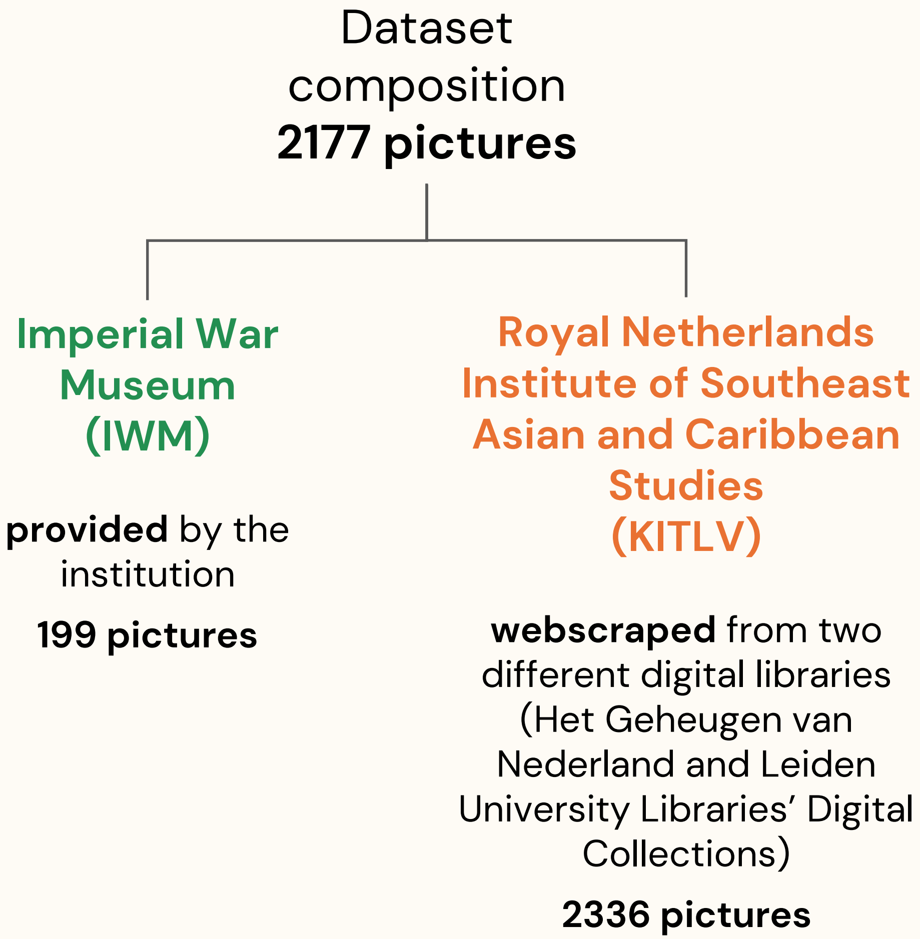
- Raw data collection from two different archival sources
- Data annotation through Label Studio (Image Classification template) using the taxonomy
- Data cleaning

- Dataset creation

Class	Samples	Percentage
Not-sensitive content	1939	76,58%
Dubious content	330	13,03%
Sensitive content	263	10,39%

Imbalance!

- **Stratified random sampling** in three sets (train, validation, test) with a 70-15-15 proportion



Dataset composition
2177 pictures

Imperial War Museum (IWM)

provided by the institution

199 pictures

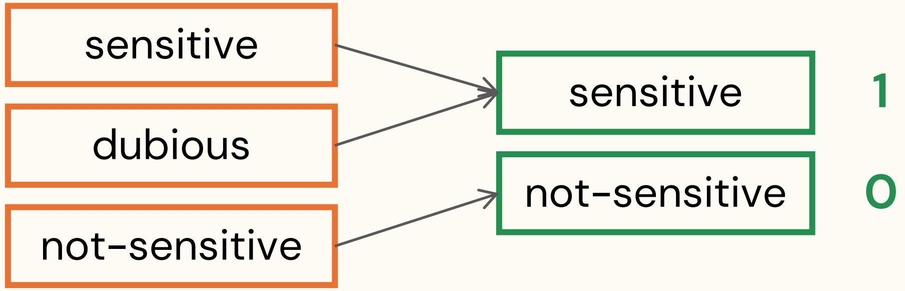
Royal Netherlands Institute of Southeast Asian and Caribbean Studies (KITLV)

webscraped from two different digital libraries (Het Geheugen van Nederland and Leiden University Libraries' Digital Collections)

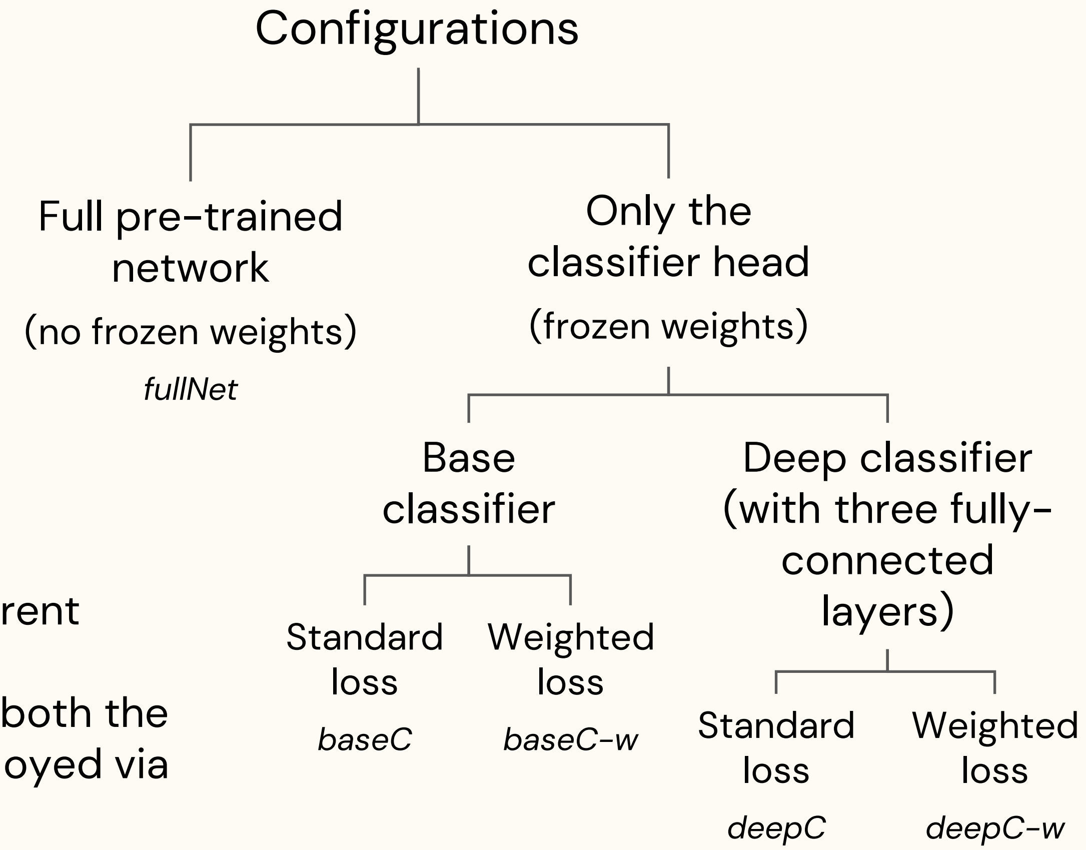
2336 pictures

02 Data and methods

- Image Classification task
- Simplification of the sensitive content definition: **binary definition and binary classification**



- Transfer learning (**ResNet50**) experimenting with different configurations
- The best performing configuration is fine-tuned using both the train and validation set and the resulting model is deployed via the test set



02 Training setup

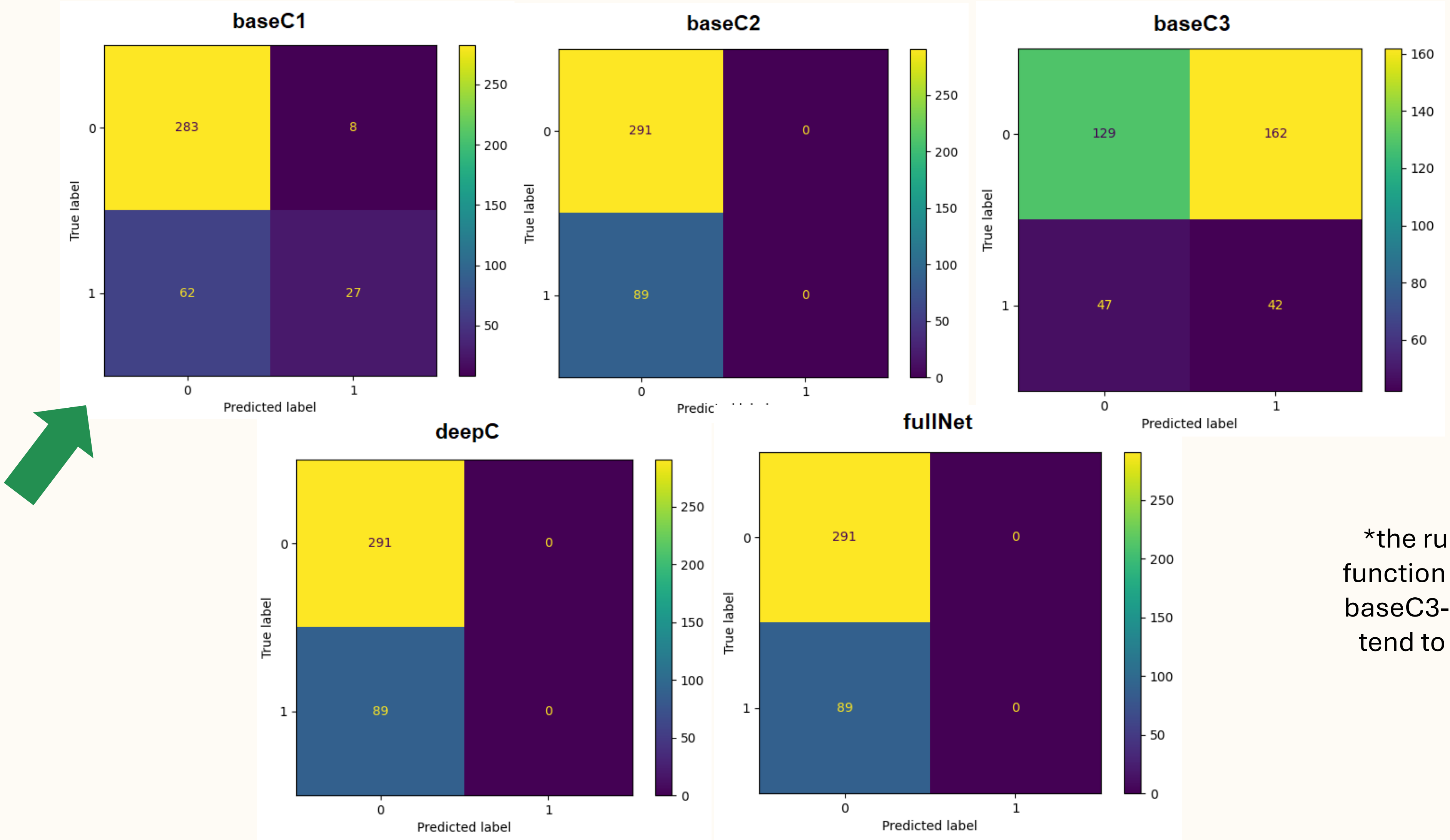
- Experimental training done on train+validation sets to improve the hyperparameters' configuration
- Metrics: accuracy, precision (m/M), recall (m/M), **f1-score** (m/**M**)
- Confusion matrix

Hyperparameter	Value
Batch size	32
Number of training epochs	15
Early stopping	Loss score (patience: 5)

	baseC1	baseC1-w	baseC2	baseC2-w	baseC3	baseC3-w
Learning rate	1e-2	1e-2	1e-4	1e-4	1e-6	1e-6
Weight decay	1e-3	1e-3	1e-5	1e-5	1e-7	1e-7
Weighted classes	No	Yes	No	Yes	No	Yes
	deepC1	**deepC1-w**				
	fullNet	**fullNet-w**				

02 Training

	baseC1	baseC2	baseC3	deepC	fullNet
F1-score (Macro)	0,662	0,433	0,419	0,433	0,433



baseC1

baseC2

baseC3

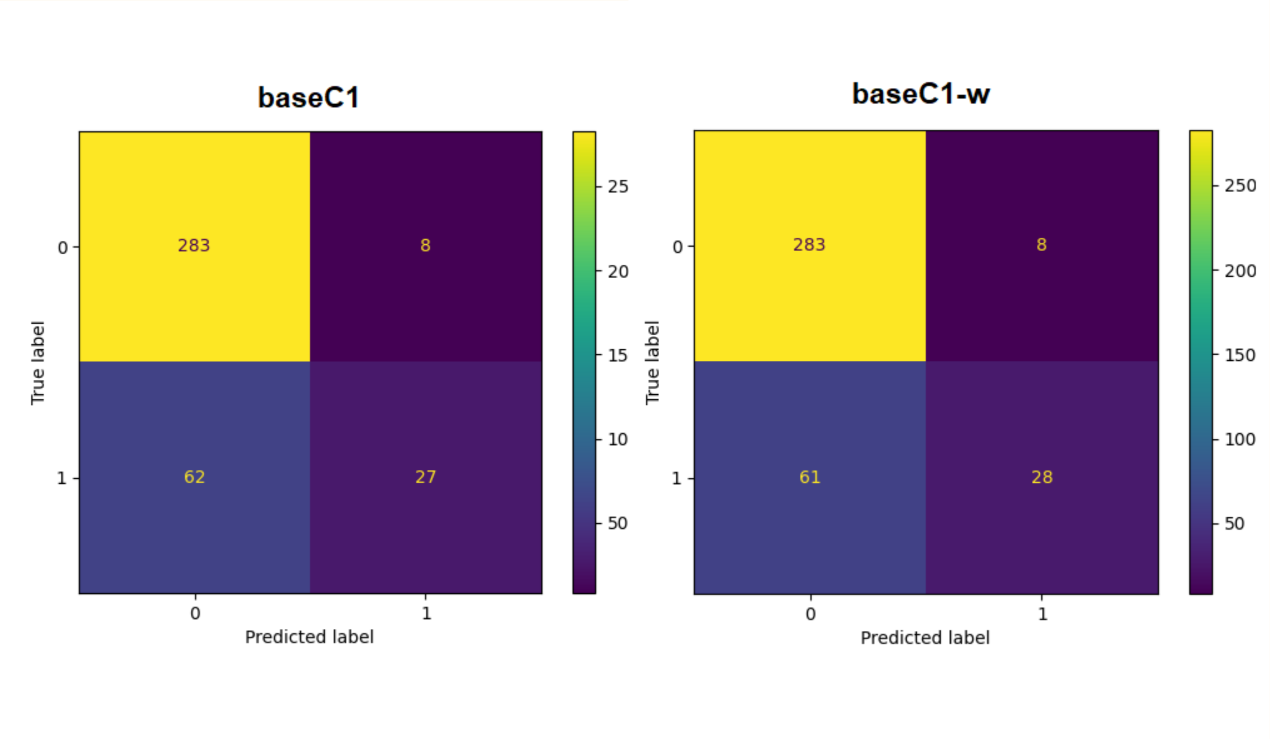
deepC

fullNet

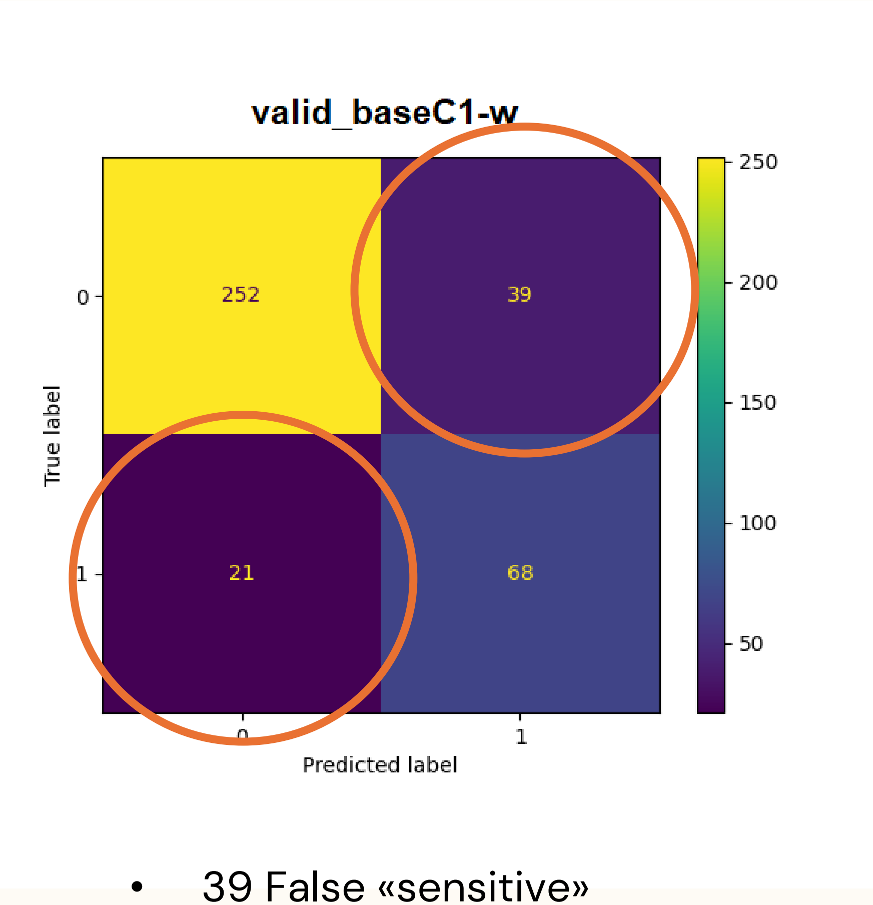
*the runs with weighted loss function (baseC1-w, baseC2-w, baseC3-w, deepC-w, fullNet-w) tend to have similar matrices

03 Results analysis

	baseC1	baseC1-w	valid_baseC1-w
F1-score (Macro)	0,662	0,669	0,793



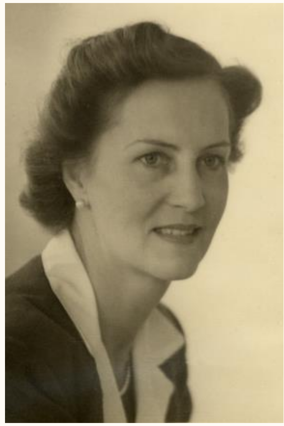
larger train dataset
larger training epochs



- 39 False «sensitive»
- 21 False «not-sensitive»

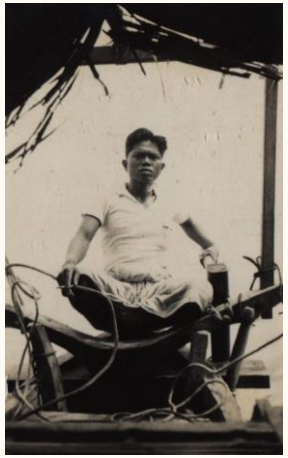
03 Results analysis

- **5 different misclassified photo styles** (studio portraiture, populated landscapes, landscapes, CH, lifestyle documentary) + «other»
- **4 possible motivations**: errors in the annotation phase, few training epochs, feature similarity, emphasis on larger features



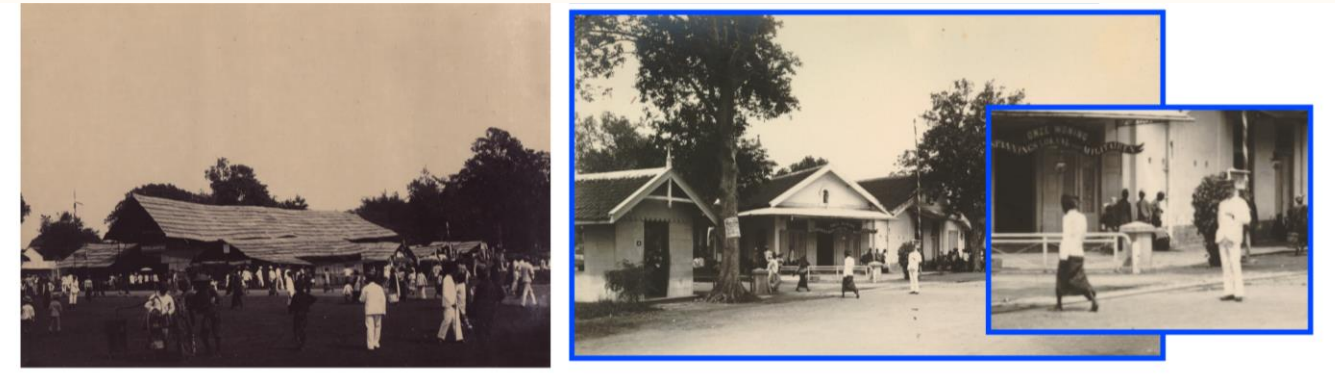
false «sensitive»

error during the annotation

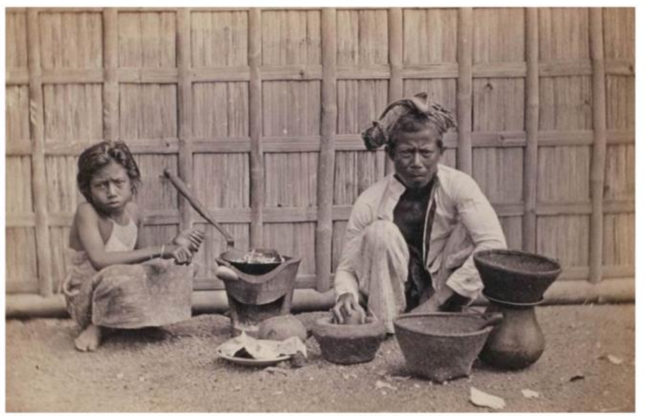


false «sensitive»

feature similarity



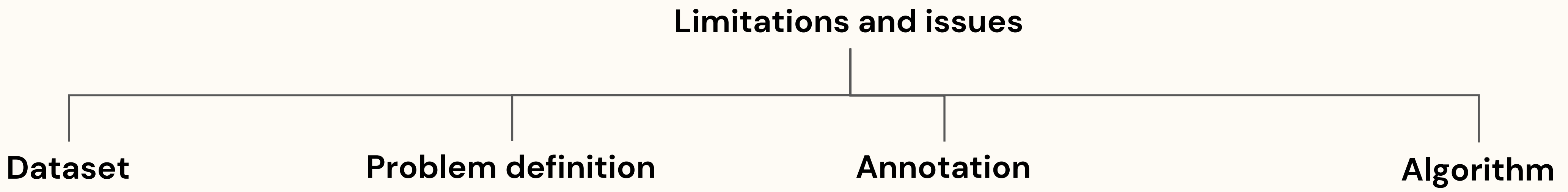
false «sensitive» – emphasis on larger features



false «not-sensitive» – few training epochs

04 Discussion

Feasibility of the application of binary Image Classification algorithms for automatic sensitive content detection **at the cost of simplifying the issue**



Limitations and issues

Dataset

Dimensions
Quality
Balance

Cluttered images

Opening access to colonial archives

Problem definition

Complexity and subjectivity
Opacity of classes' boundaries

Intraclass variation
Similarity across classes

Gather further insights from diverse stakeholders

Annotation

Small number of annotators
No specific preparation on the topic

Annotation pilot

Algorithm

Inadequacy of the Image Classification algorithm (*feature scale variation*)
Smallscale hyperparameter tuning

Further hyperparameter tuning
Exploration of different approaches: Object Detection; multimodal ML algorithms (CLIP, GLIP)

Alma Mater Studiorum – Università di Bologna

Revealing Contested Memory: Automatic sensitive content detection in colonial photographic archives

Thank you!

Defended by Orsola Maria Borrini
Supervisor Prof. Giovanni Colavizza
Co-supervisor Prof. Charles Jeurgens (University of Amsterdam)

Session III
Academic Year 2022/2023