Leveraging Large-scale Multimedia Datasets to Refine Content Moderation Models

Ioannis Sarridis, Christos Koutlis, Olga Papadopoulou, and Symeon Papadopoulos

Media Analysis, Verification and Retrieval Group (MeVer) https://mever.gr
CERTH-ITI

Content Moderation

CM importance:

- Protect platforms’ audience from harmful content

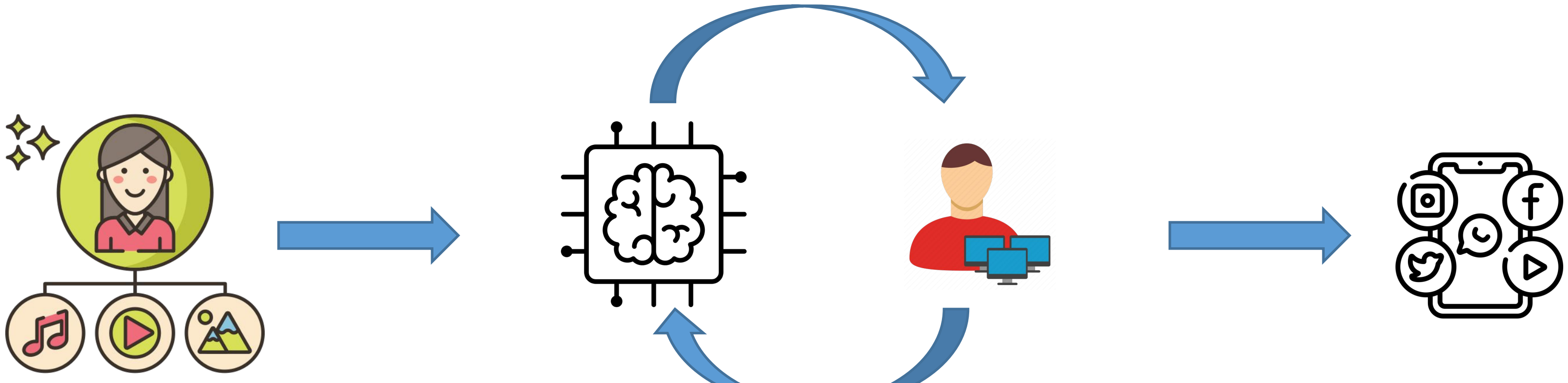
CM in popular platforms:

- Non-transparent systems comprising AI models and human moderators

CM in decentralized platforms (e.g., MediaVerse):

- Transparency in terms of policies and AI systems

MediaVerse



User generated Content

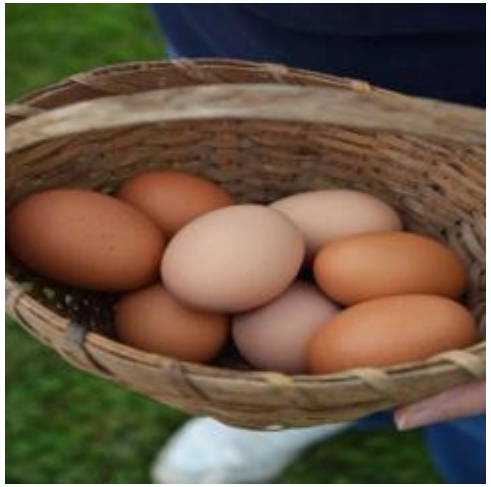
AI models

Human

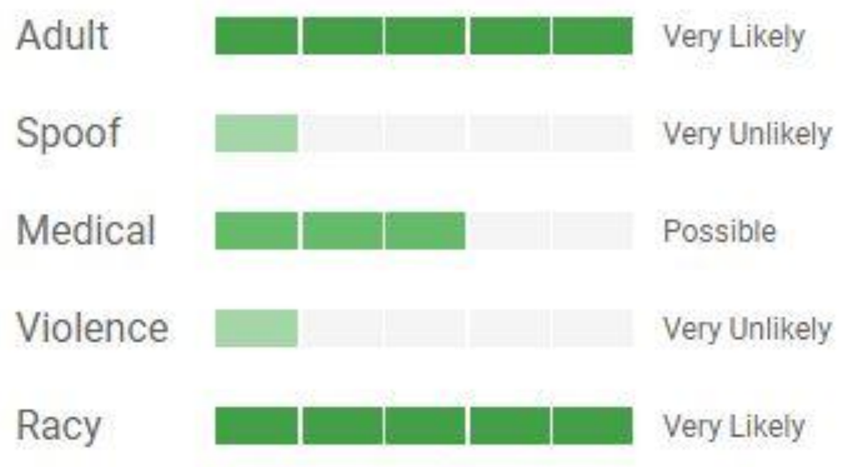
Web

AI models reliability

Why human annotators are still necessary?



Google Vision AI
https://cloud.google.com/vision



Adult	Very Likely
Spoof	Very Unlikely
Medical	Possible
Violence	Very Unlikely
Racy	Very Likely

Motivations and Contributions

Motivations:

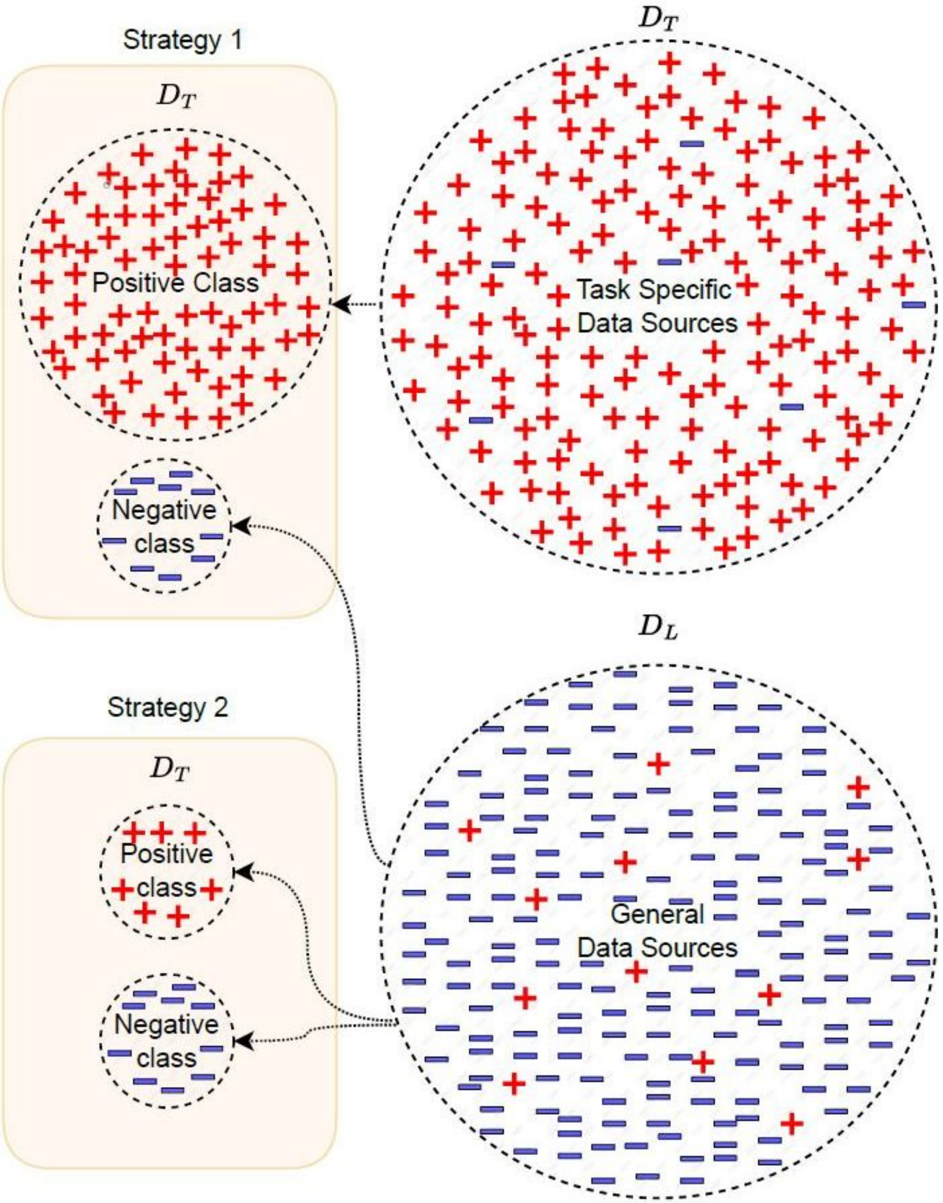
- Lack of adequate task-specific training data
- Manually annotation impact on the annotators’ emotional well-being

Contributions:

- A framework for
  - **collecting** and **annotating** task-specific content moderation data
  - **minimizing** the human annotators’ involvement
- Consideration of **two model adaptation strategies**

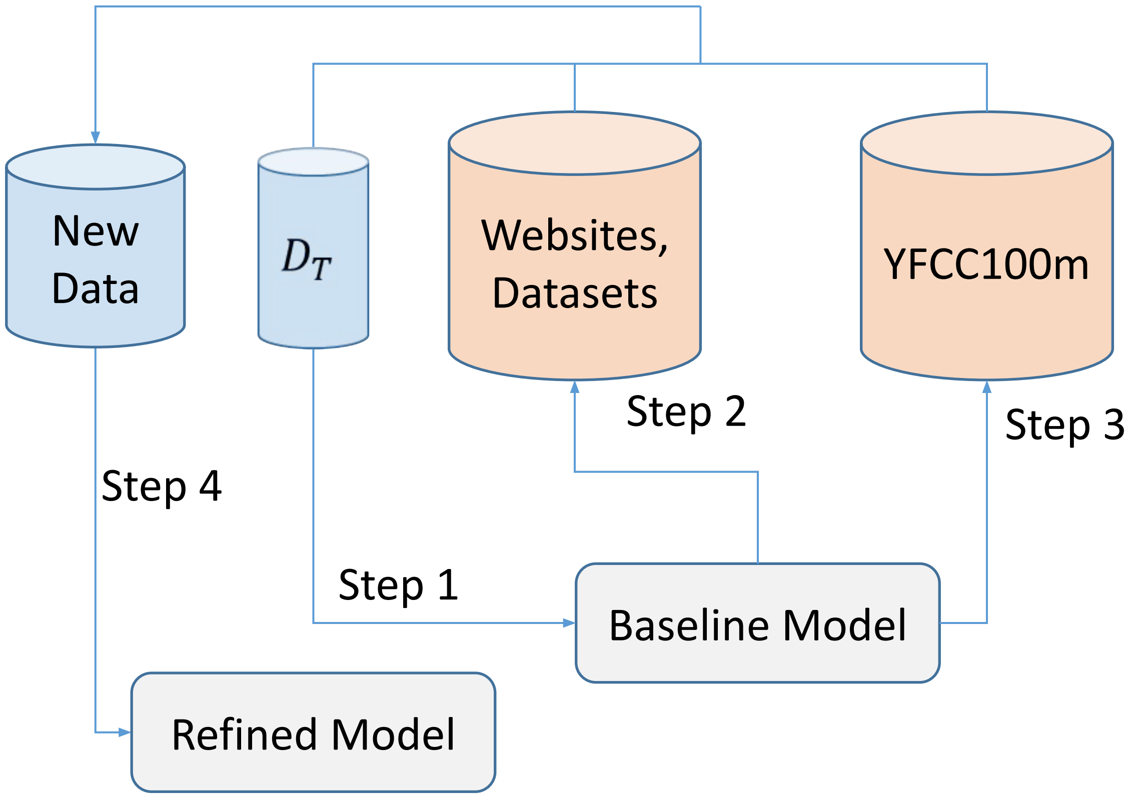
Model adaptation strategies

- Task-specific data sources D_T: websites, datasets, etc.
- General data sources D_L: Large public multimedia datasets
- Strategy 1:
  - Positive Data: Task-specific data sources
  - Negative Data: General data sources
  - Task: NSFW detection
- Strategy 2:
  - Positive Data: General data sources
  - Negative Data: General data sources
  - Task: Disturbing content detection



Framework – Strategy 1 – NSFW

- **Step 1**. Train a baseline model utilizing existing task-specific datasets
- **Step 2**. Expand positive data: web (e.g., pornography websites) and datasets (e.g., NudeNetData)
- **Step 3**. Expand negative data: YFCC100m samples classified as positive by the baseline model (i.e., hard-negatives)
- **Step 4**. Retrain the model utilizing the new training data

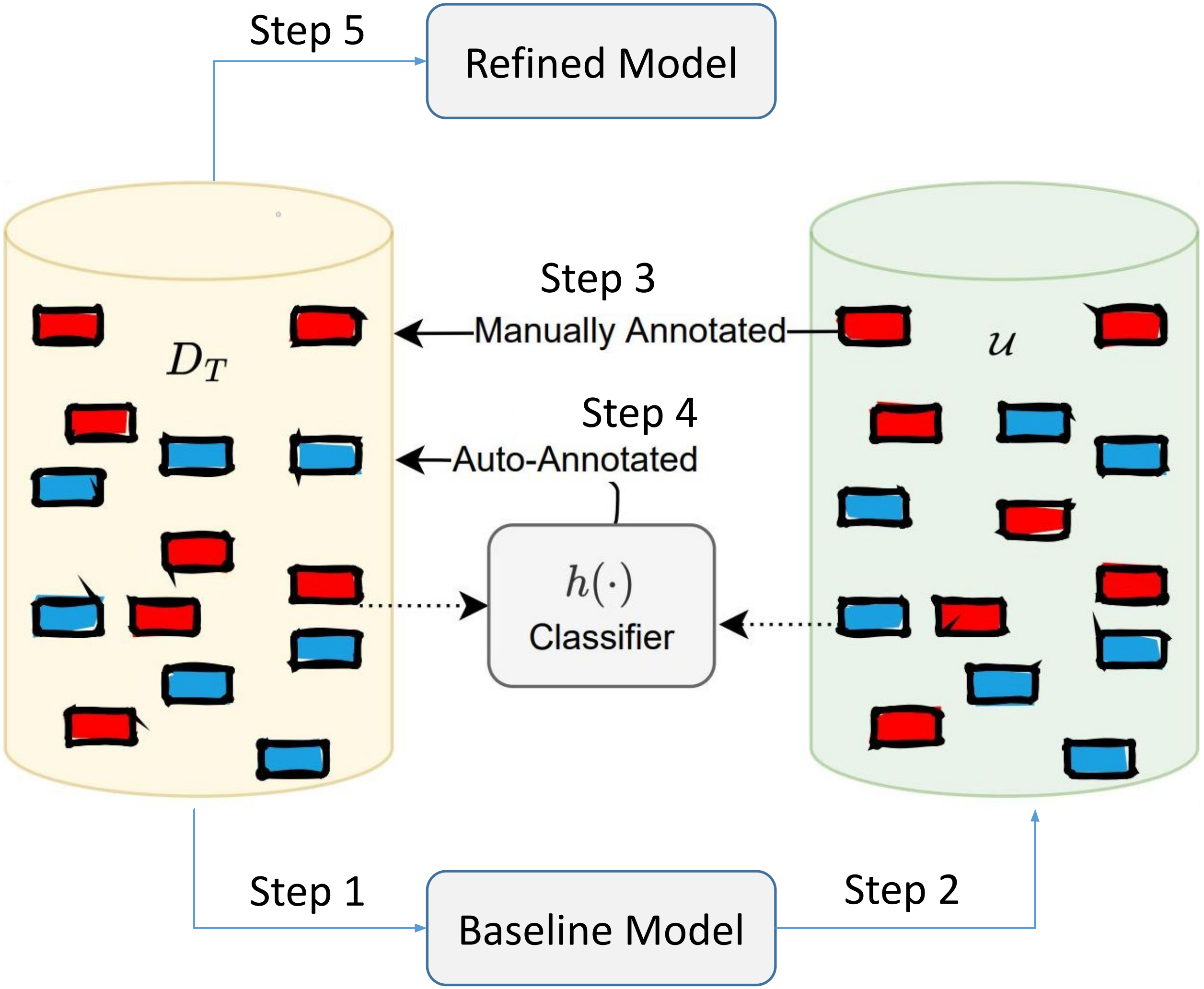


Framework – Strategy 2 – Disturbing Content

- **Step 1**. Train a baseline model utilizing existing task-specific datasets (i.e., D_T)
- **Step 2**. Keep the YFCC100m samples classified as positive by the baseline model (i.e., $\mathcal{U}$)
- **Step 3**. Manually annotate a few samples per class (i.e., $\mathcal{M}$)
- **Step 4**. Auto-annotate $\mathcal{U}$ samples

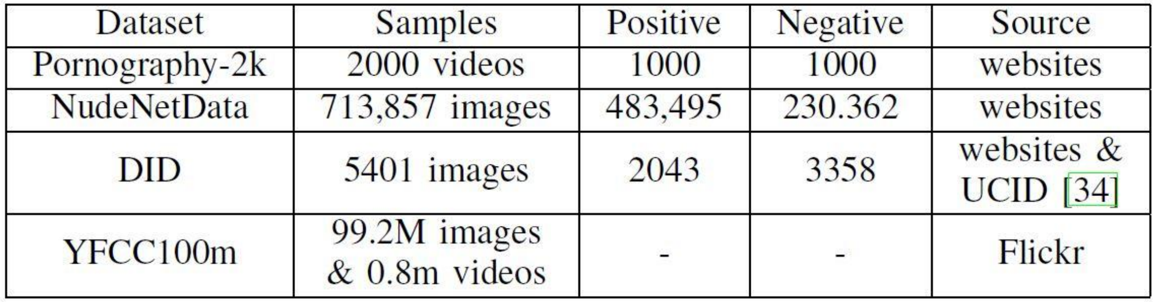
$$s_{i,j} = \frac{f^l(\mathbf{M}_i)f^l(\mathbf{U}_j)}{||f^l(\mathbf{M}_i)||\ ||f^l(\mathbf{U}_j)||}$$ & radius-NN (i.e., $h(\cdot)$)

- **Step 5**. Retrain the model utilizing the new training data



Experimental Setup

- Datasets:
  - NSFW: Pornography-2k
  - Disturbing content: DID
  - NSFW task-specific data source: NudeNetData
  - General data source: YFCC100m



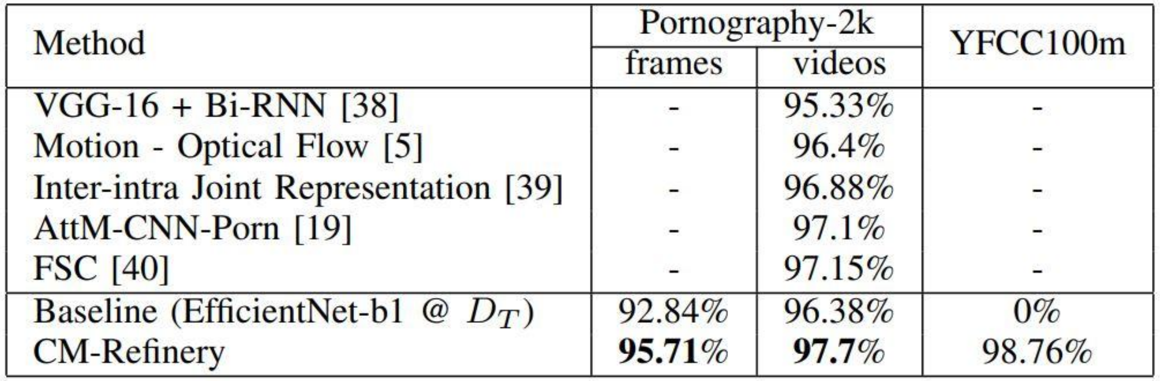
Dataset	Samples	Positive	Negative	Source
Pornography-2k	2000 videos	1000	1000	websites
NudeNetData	713,857 images	483,495	230.362	websites
DID	5401 images	2043	3358	websites & UCID [34]
YFCC100m	99.2M images & 0.8m videos	-	-	Flickr

- Model Architecture: EfficientNet-b1

- Performance evaluation:
  - Accuracy on Pornography-2k frames and videos
  - Accuracy on DID images

Results

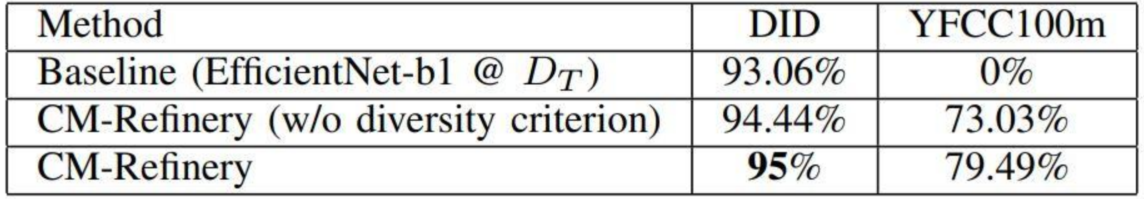
Fully automated annotation



Method	Pornography-2k		YFCC100m
	frames	videos	
VGG-16 + Bi-RNN [38]	-	95.33%	-
Motion - Optical Flow [5]	-	96.4%	-
Inter-intra Joint Representation [39]	-	96.88%	-
AttM-CNN-Porn [19]	-	97.1%	-
FSC [40]	-	97.15%	-
Baseline (EfficientNet-b1 @ D_T)	92.84%	96.38%	0%
CM-Refinery	**95.71%**	**97.7%**	98.76%

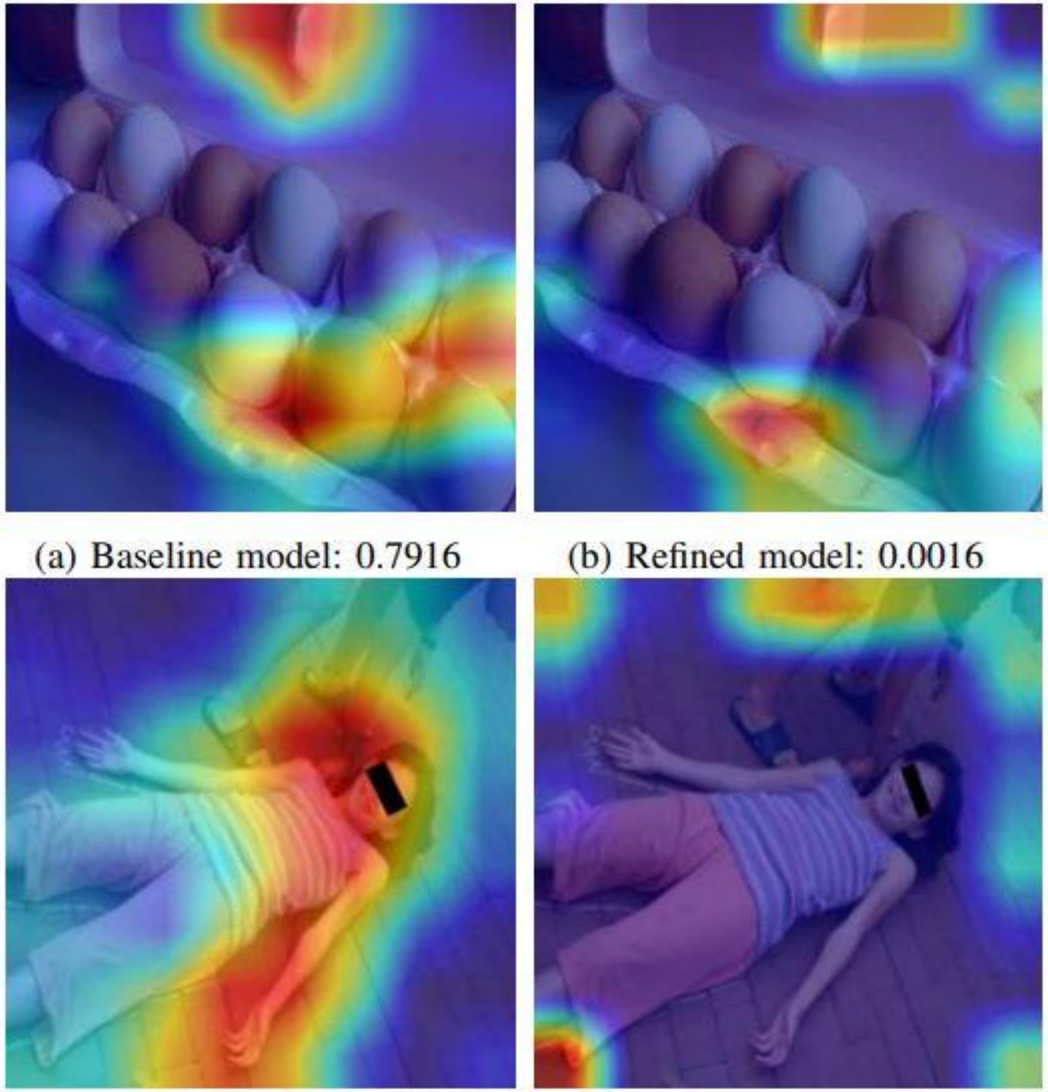
TABLE II: Performance comparison on D_T: Pornography-2k.

Human exposure to harmful data reduced by x13.4



Method	DID	YFCC100m
Baseline (EfficientNet-b1 @ D_T)	93.06%	0%
CM-Refinery (w/o diversity criterion)	94.44%	73.03%
CM-Refinery	**95%**	79.49%

TABLE III: Results of conducted experiments on D_T: DID.



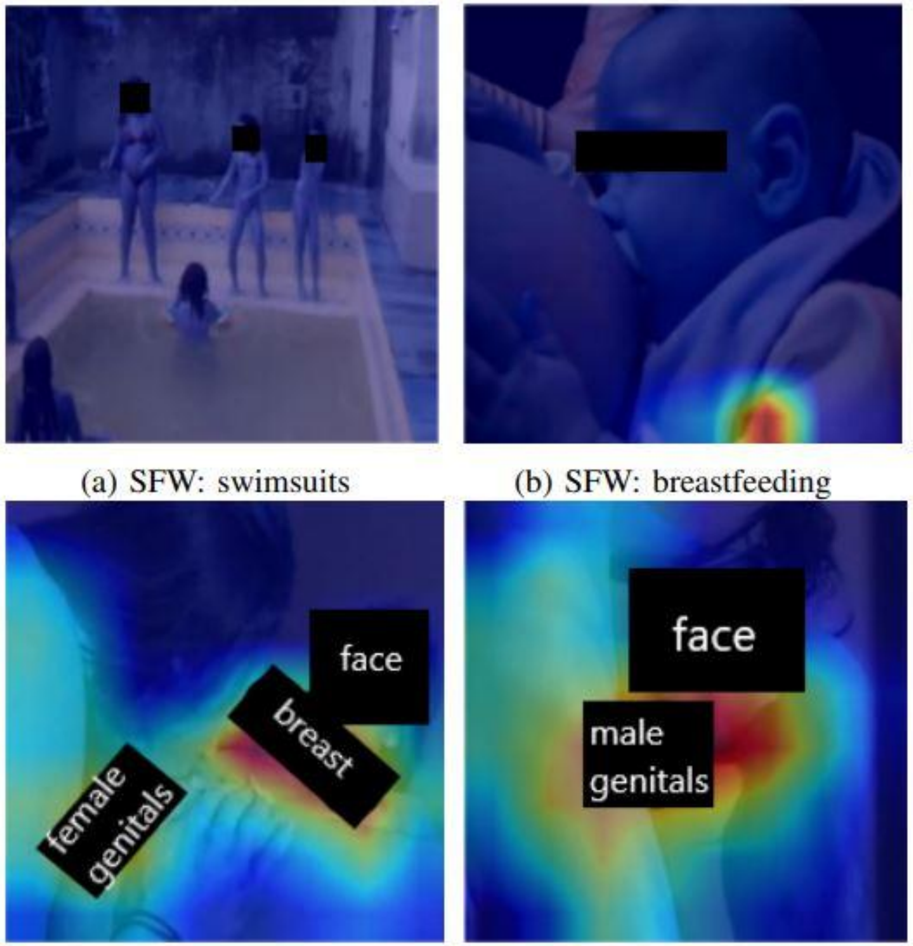
(a) Baseline model: 0.7916

(b) Refined model: 0.0016

(c) Baseline model: 0.9503

(d) Refined model: 0.005

Qualitative analysis

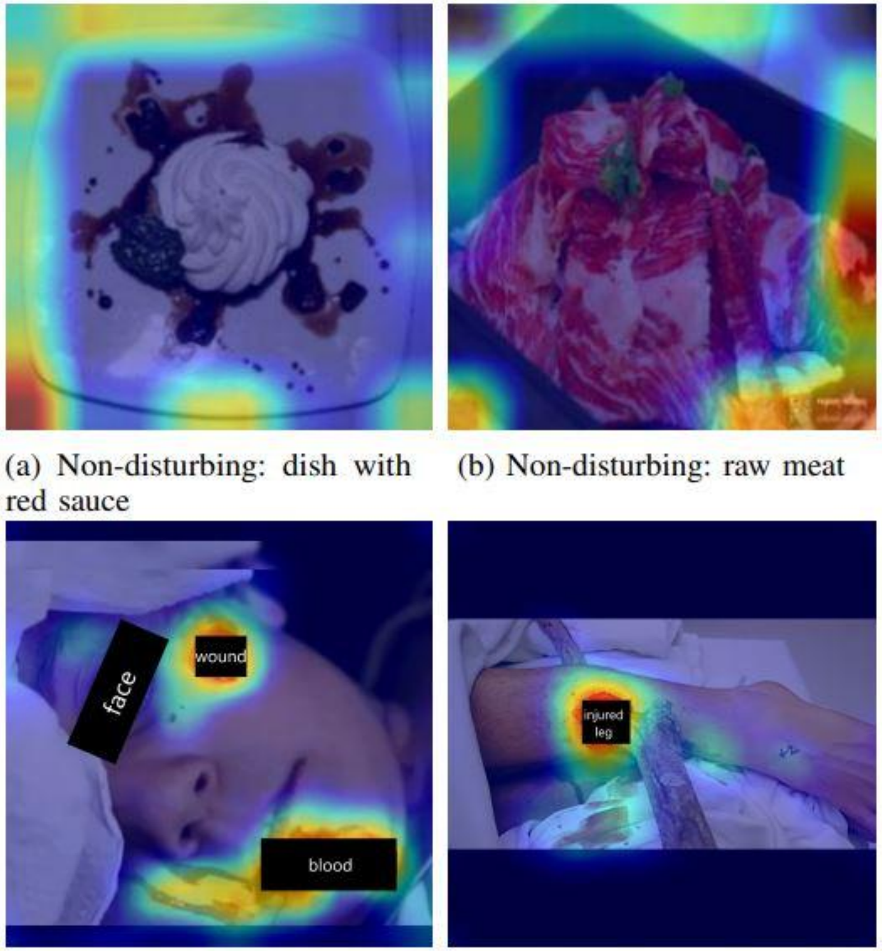


(a) SFW: swimsuits

(b) SFW: breastfeeding

(c) NSFW: female breast and genitals

(d) NSFW: male genitals



(a) Non-disturbing: dish with red sauce

(b) Non-disturbing: raw meat

(c) Disturbing: blood

(d) Disturbing: severe wound

The manually assigned labels describe what the images depict.

Refined Model vs Commercial Services

(a) DeepAI: 0.7466 (NSFW), Google: 5/5 (NSFW), Ours: 0.1206 (SFW)

(b) DeepAI: 0.9417 (NSFW), Google: 3/5 (NSFW), Ours: 1e-5 (SFW)

(c) DeepAI: 0.9532 (NSFW), Google: 5/5 (NSFW), Ours: 0.1248 (SFW)

(d) DeepAI: 0.0431 (SFW), Google: 4/5 (NSFW), Ours: 0.6845 (NSFW)

(e) DeepAI: 0.0138 (SFW), Google: 1/5 (SFW), Ours: 0.3453 (SFW)

(f) DeepAI: 0.3484 (SFW), Google: 3/5 (NSFW), Ours: 0.8513 (NSFW)

Future Work

- How the subjective nature of content moderation affects the AI models?
    - Few-shot approaches

- How the AI models can deal with the policy changes?
    - Frameworks that consider the policy changes

- How the bias in AI models affects their decisions?
    - Assess and mitigate bias in content moderation AI models

Thank you!

Media Analysis, Verification and Retrieval Group (MeVer)

🌐 https://mever.gr

https://twitter.com/meverteam

✉ gsarridis@iti.gr

CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

iti Information Technologies Institute

MediaVerse
https://mediaverse-project.eu