

Motivation

- The observed **tendency of disinformation (disinfo) to adopt a sensationalist story format** indicates that detecting sensational content is essential for **spotting news items requiring immediate fact-checking**
- Text-based approaches can **identify provocative, exaggerated, or emotionally charged language** to grab attention
- But, **sensational images** designed to grab immediate attention & provoke intense emotional reactions also act as a **potent catalyst to spread disinfo**
- Tools for **sensational image detection** could act as a **filtering mechanism** to assist fact-checkers in **identifying check-worthy content & flagging disinfo**

Current landscape on visual content analysis

- Two classes of methods: one dealing with **NSFW content detection**, and one targeting **visual sentiment analysis**
- First class aims to spot **visual material inappropriate for public, professional or general viewing or violating safety guidelines**
- Second class aims to **identify the emotional tone/attitude expressed in the visual content**
- Existing methods can support the **detection of sensational images for flagging disinformation, only to a limited extent**

Problem formulation & data selection

Problem statement

- Sensational image detection aims to spot **shocking, provocative or emotionally charged visual content, intended to grab viewers' attention & trigger strong emotional responses**

Source dataset: VisualNews dataset

- Contains > 1M news images, along with articles, captions & other metadata from 4 news agencies
- Used for news image captioning; **extended to support other tasks** (e.g., OOC multimodal data detection)
- Is **highly-aligned with our needs**, since we are dealing with the detection of sensational images in news items

Data selection

- Based on a **list of sensational visual concepts & events**, formed by 5 journalists & 7 fact-checkers from 6 institutions located in 4 different EU countries
- Covers **194 visual concepts & events from 5 topics**: i) war & conflict, ii) refugees & migrants, iii) climate change & environmental disasters, iv) racism, v) religion
- Automated pre-selection of **candidate sensational images** for annotation
 - 5K images with the highest relevance to the concepts & events of the formed list, according to a method for ad-hoc text-driven video search¹
 - 3K additional images from under-represented topics
- Automated pre-selection of **candidate non-sensational images** for annotation
 - 5K images with the lowest relevance to the concepts & events of the list
 - 3K additional images based on a curated list of visually-relevant non-sensational concepts & events

Sens-VisualNews dataset

Data annotation

- Performed by 3 annotators who worked independently & had to specify whether visual content was sensational, non-sensational or ambiguous, using a custom-made annotation tool (Fig. 1)
- Collected labels per image were aggregated based on majority vote; ties or “ambiguous” decisions were handled by image removal
- 24% of candidate “sensational” images were flipped to the “non-sensational” class

Sens-VisualNews dataset: 9,576 images from various news topics (Fig. 2) equally divided in two classes (Fig. 3); split in dev. (~10% of samples) & test set (~90% of samples)

Strict subset: 4,542 images with unanimous annotator decisions, again equally divided in two classes and split in dev. & test set

Dataset for download: <https://github.com/IDT-ITI/Sens-VisualNews>

Sensational Image Annotation Tool

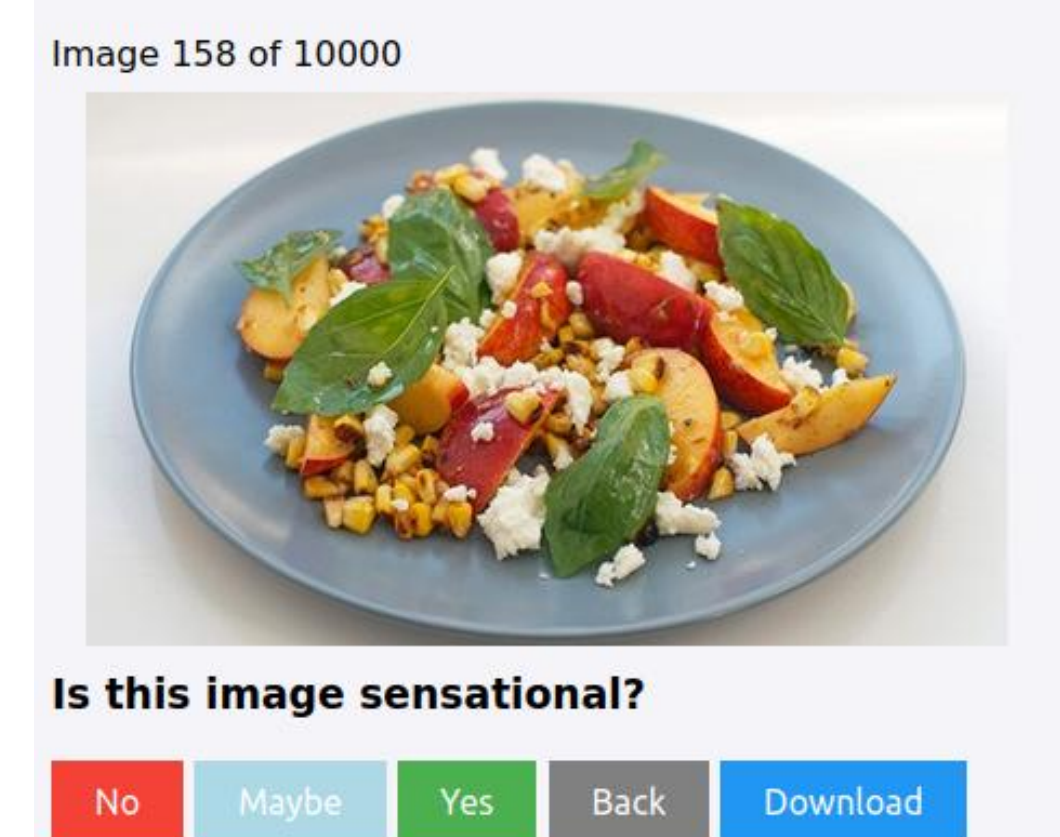


Fig. 1 A screenshot of the data annotation tool.

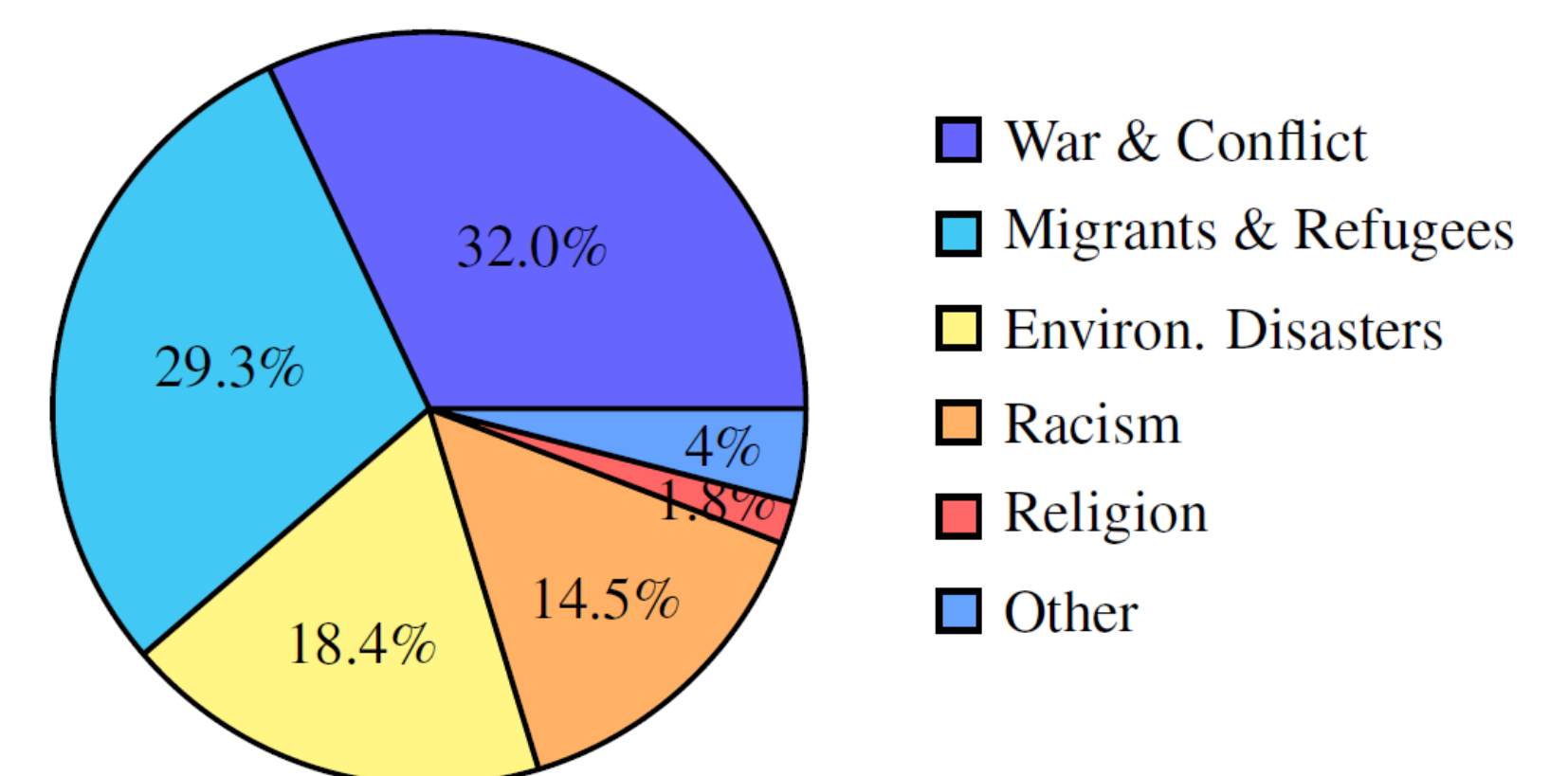


Fig. 2 News topics distribution in the “sensational” class of the Sens-VisualNews dataset.



Fig. 3 Examples of sensational (top row) & non-sensational (bottom row) images from the Sens-VisualNews dataset.

Experimental results

Evaluation protocol

- Evaluated various MLLMs (see Table 1) based on top-1 accuracy on the test set, under zero-shot & fine-tuned evaluation settings
- Fine-tuned the models with the highest zero-shot performance, using a subset of images (dev. set) & Low Rank Adaptation

Prompting strategies & prompt sensitivity on the dev. sets

- Examined **3 different prompts & 2 prompt formats** for each prompt (prepending/appending the visual tokens in the prompt)
 - “Is this image sensational? A sensational image evokes strong emotions (e.g. fear, anger, anxiety, disgust, shock). Answer with a single yes or no.”
 - “Does this image contain shocking, provocative or emotionally charged features to grab the viewer’s attention? Answer with a single yes or no.”
 - “Does this image trigger strong emotional responses (e.g. fear, anger, anxiety, disgust, shock)? Answer with a single yes or no.”
- All models show noticeable sensitivity to the used prompt; no strict connection between model size & sensitivity

Performance using the best prompting approach per model

Zero-shot setting

- Closely tied to the num. of params
- Higher on strict set due to decreased label ambiguity/noise
- Models from **Qwen3 VL & LLaVa OV 1.5** families are the most competitive

Fine-tuned setting

- Significantly improved in all cases
- Qwen3 VL 8B** gains +3.0% top-1 accuracy & shows the highest performance on the created Sens-VisualNews dataset

Model	Params	Test Top-1 (%)	
		Strict	Full
Baseline (VLM-based)	11B	84.2	81.2
SmolVLM2	2.2B	68.1	62.2
InternVL 3.5	1B	76.4	76.1
	2B	86.2	77.9
	4B	88.8	83.4
	8B	92.1	83.8
Qwen3 VL	2B	92.7	86.3
	4B	90.8	85.0
	8B	<u>92.8</u>	<u>87.0</u>
	2B (ft)	95.1	89.4
	4B (ft)	<u>95.4</u>	89.4
LLaVA OV	0.5B	90.8	81.3
	7B	91.8	81.5
LLaVA OV 1.5	4B	89.7	86.7
	8B	93.9	87.6
	4B (ft)	<u>95.4</u>	89.7
	8B (ft)	94.6	<u>89.8</u>

Table 1. Performance comparison on the test sets of the full dataset and its strict subset, across zero-shot & fine-tuned settings. Best scores in bold, second-best underline. Gray highlight indicates fine-tuning