# Towards Explainable Multimodal Detection of Misogynistic Memes

Johanne Pinel, Ambre Sassi, Julien Thévenoz
Group 24

## Problem definition

Online platforms are increasingly plagued by misogynistic content, which contributes to a toxic digital environment and real-world harm. While deep learning models have shown promising performance in detecting such harmful content, they often function as black boxes, providing little insight into their decision-making processes. We explore the use of post-hoc explainability techniques on a multimodal classifier for misogynistic meme detection, with the goal of understanding and visualizing how deep learning models make decisions in sensitive societal contexts to ensure that they are trustworthy and aligned with human reasoning.

## Key Related Works

- TIB-VA at SemEval-2022 Task 5: A Multimodal Architecture for the Detection and Classification of Misogynous Memes
- Explainable Classification of Internet Memes
- A Closer Look at the Explainability of Contrastive Language-Image Pre-training
- From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI

## Method

### TIB-VA classifier :

- **CLIP**: Pre-trained model encoding aligning images and text with semantical alignment. Image embedding : ViT-L/14 and Text embedding : Transformer
- **Multimodal Neural Network**: Custom model for binary classification from CLIP embeddings.
- **Transformer-based Tokenization:** Text preprocessing for CLIP using a Transformer-compatible tokenizer.

### Evaluated 2 explainability methods :

- **CLIP-Surgery** : highlight images patches which are semantically aligned with concept of misogyny, shaming, etc.
- **Saliency maps** : highlight patches which most contributed to the output score

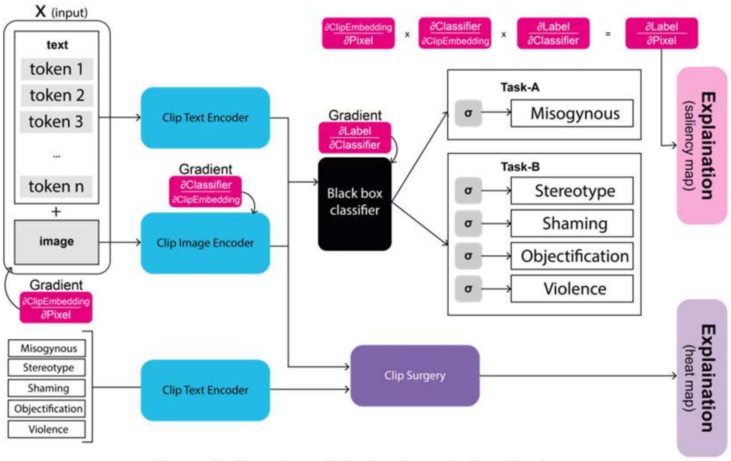


Figure 1 : Overview of the implemented methods

## Dataset(s)

- MAMI dataset : 8,000+ memes labeled for misogyny and subcategories (violence, shaming, stereotyping, objectification).
- MEME dataset: Benchmark dataset of 800 memes with text transcriptions for automatic detection of multi-modal misogynistic content

EE-559: Deep Learning, 2025

## Validation

### Method

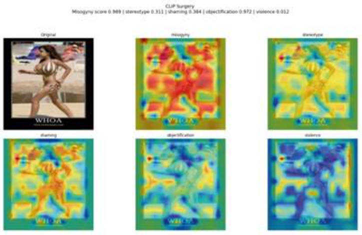
- Performance testing on MAMI and MEME datasets
- **Metrics** to assess performance: Accuracy and F1-score

| Category | Overall Misogyny | | Subcategory (F1 Scores) | | | |
|---|---|---|---|---|---|---|
| Dataset | Accuracy [%] | F1 Score [%] | Shaming [%] | Stereotype [%] | Objectification [%] | Violence [%] |
| MAMI Validation | 89,40 | 89,40 | 71,19 | 74,83 | 79,65 | 66,79 |
| MAMI Testing | 74,10 | 72,87 | 66,98 | 67,46 | 77,89 | 72,06 |
| MEME Testing | 96.62 | 96.60 | 68.30* | 83.43* | 72.60* | 81.55* |

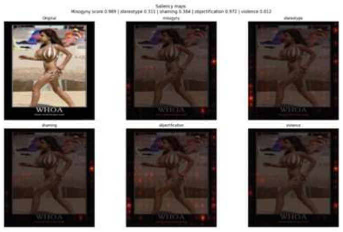
* Trained on a self-annotated subset (17.5%)

- Results of the different explainability methods:

Explaination with CLIP Surgery



Explaination with Saliency map



### Discussion

Classifier performs well but post-hoc explainability methods (saliency maps, CLIP-Surgery) produce noisy and uninformative outputs. This aligns with known limitations of CLIP-based models : poor attention localization and semantic misalignment hinder interpretability. Architectural differences in the image encoders and missing tuning steps likely contributed to the failure.

- **Possible Improvements**
  - Try different explainability methods (SHAP, self attention)
  - Quantifiable metrics for explainability following Co-12 principles [3]

## Limitations

- **Weak current results:** So far, the explanations produced are not sufficiently informative or reliable.
- **Limited explanatory power due to modality separation:** Our methods analyses text and image separately, not considering their interaction. Even if separate modality-explanation worked correctly, it couldn't explain memes where hate speech arises from combination of benign text & images.

## Conclusion

While the misogynistic meme classifier works well on different datasets, its explainability capabilities were unconclusive due to lack of consideration of multi-modality and architecture difficulties. The project also laid the groundwork for integrating explainability techniques such as saliency maps and vision-text alignment via CLIP-Surgery.
Future work should explore cross-modal explanation methods and develop human-grounded evaluation metrics to further improve both performance and interpretability.

### References

[1] S. Hakimov, G. Cheema, and R. Ewerth, "TIB-VA at SemEval-2022 Task 5: A Multimodal Architecture for the Detection and Classification of Misogynous Memes," 2022. Accessed: May 25, 2025. [Online]. Available: https://aclanthology.org/2022.semeval-1.105.pdf

[2] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, Xiaomeng Li, "A closer look at the explainability of Contrastive language-image pre-training," Pattern Recognition,Volume 162, 2025, 111409, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2025.111409.

[3] Nauta, Meike & Trienes, Jan & Pathak, Shreyasi & Nguyen, Elisa & Peters, Michelle & Schmitt, Yasmin & Schlötterer, Jörg & Van Keulen, Maurice & Seifert, Christin. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys. 55. 10.1145/3583558.