Towards Explainable Multimodal Detection of Misogynistic Memes

Pinel Johanne **Sassi Ambre** **Thevenoz Julien**
Group 24

Abstract

Misogynistic memes contribute to a toxic digital environment and real-world harm. Recent advances in explainable AI have introduced techniques that improve the interpretability of multimodal models by making their decision processes more transparent and human-understandable. We apply explainability technique (LIME, CLIP surgery, and saliency maps) to the text and image modalities separately of a multimodal classifier for misogynous meme detection. We evaluated our model's classification performance across the MAMI and MEME datasets and proposed explanations for its limited explaining capacity.

1. Introduction

Online platforms are increasingly plagued by misogynistic content, which contributes to a toxic digital environment and real-world harm. Misogyny, a form of hate speech directed specifically at women, undermines gender equality and individual well-being [**?**]. 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 apply explainability techniques such as LIME [1], CLIP surgery [2] and saliency maps [3] to the text and image modalities of a multimodal classifier for misogynistic meme detection [4]. Our goal is to visualize and understand model decisions in sensitive societal contexts, ensuring they are trustworthy and aligned with human reasoning.

2. Related Work

Multimodal misogyny detection gained popularity with SemEval-2022 Task 5 [5], which introduced the MAMI dataset for binary classification (Task A) and fine-grained categorization into shaming, stereotype, objectification, and violence (Task B). Among the top-performing models, SRCB [6] achieved the best overall results, while TIB-VA [4] led the fine-grained classification and is openly available. We build upon TIB-VA, as subcategory classification provides a natural foundation for explainability. Beyond the competition, Jindal et al. [7] proposed MISTRA, a hybrid architecture combining ViT, DistilBERT, CLIP, and BLIP with dimensionality reduction and feature alignment for improved fusion. Rehman et al. [8] introduced a context-aware framework leveraging attention mechanisms and graph neural networks, highlighting interpretability as a key direction for future work.

Recent advances in explainable AI have introduced techniques that improve the interpretability of multimodal models by making their decision processes more transparent and human-understandable. CLIP Surgery enhances the visual interpretability of CLIP by modifying its inference pathway to yield sharper and more localized class activation maps (CAMs), without requiring retraining [9]. Gradient-based saliency maps visualize which input pixels most influence model predictions by computing the gradient of the output with respect to the input, offering intuitive insight into visual reasoning [3]. LIME provides local post-hoc explanations by approximating the model around a given prediction with an interpretable surrogate model, applicable to both text and image modalities [1].

3. Method

3.1. Model architecture

For this project we use the TIB-VA[4] Model from the SemEval-2022 Task 5 competition, to determine if the meme were misogynistic or not. The evaluation of memes is carried out through two tasks. Task A determines whether a meme is misogynistic or not, while Task B identifies which subcategory of misogyny the meme falls into. The subcategories are: shaming, stereotype, objectification, and violence.

The model takes the image and the overlaid text of the meme as separate inputs and uses CLIP (ViT-L/14), pre-trained on 400 million image-text pairs, to extract their respective representations. The text is first processed by the CLIP Text Encoder, which outputs 768-dimensional embeddings for each token. These embeddings are then passed through a

256-unit Long Short-Term Memory (LSTM) layer to capture contextual information. In parallel, the image is processed by the CLIP Image Encoder and projected into a 256-dimensional space via a fully connected layer. Both textual and visual features undergo dropout (rate 0.2), are concatenated, and then passed through another fully connected layer. Finally, sigmoid layers are used to predict whether the meme is misogynistic (Task A) and which subcategory or subcategories it belongs to (Task B).

3.2. Training procedure

The training procedure of this model is based on the Adam optimizer with an initial learning rate of c. The batch size is set to 64 and the model is trained with 20 epochs. To have a better convergence, the learning rate is cut in half every 5 epochs. Furthermore, 10% of the training data is used as a validation set for tuning the hyperparameters. This model is implemented using the Pytorch library in Python.

3.3. Data set

For this project, we used two different datasets containing misogynistic memes.

The first one is the **MAMI dataset** [5], which was the focus of the SemEval-2022 competition where the code was originally developed. This dataset is designed for the automatic identification of misogynous memes using both textual and visual information. It includes two main subtasks: Task A, a binary classification of misogyny presence, and Task B, which requires recognizing specific types of misogyny. Since the code was created for this challenge, it is well adapted to the MAMI dataset and its associated tasks.

We also tested the model on a second dataset, the **MEME dataset** [10], which contains 800 memes along with text transcriptions and is designed for automatic detection of multimodal misogynistic content. To evaluate the model's generalization beyond the MAMI dataset, we applied it to MEME. Since the labelling scheme differed, we manually annotated 17.5% of the dataset to enable proper evaluation.

3.4. Explainability procedure

For the explainability of the memes we choose to explore 3 different methods, each of them or only for the images.

CLIP-Surgery : We selected the memes that were labeled as misogynistic by the model and passed them through CLIP Surgery to better understand what the model was focusing on. CLIP Surgery slightly alters how CLIP processes inputs during inference to generate clearer and more focused visual explanations, without requiring any fine-tuning. It helps reduce noise and makes the model's focus more aligned with the actual content.

Saliency maps : We implemented saliency maps ourselves to identify which parts of a meme most influence the model's prediction. By calculating the gradient of the output relative to the input pixels, saliency maps highlight key regions that contribute to the final decision, making it easier to interpret and validate the model's behavior in complex multimodal tasks like misogyny detection. **LIME on text :** LIME relies on perturbing interpretable input features and observing changes in the model's output.

Figure 1 presents an overview of our model architecture along with the different implementations.

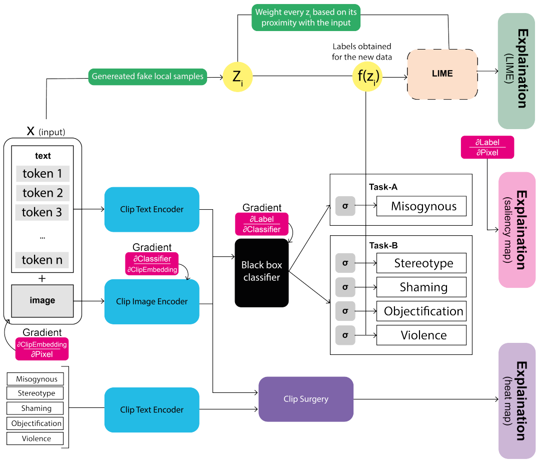

Figure 1. Overview of our model and its explainability implementations

4. Validation

4.1. Model performances

We evaluated our model's classification performance and explainability capacity across the MAMI and MEME datasets, the results are presented in Table 1. Since the classifier component of our model is inherited from the TIB-VA architecture, it retains the same level of classifying ability. Surprisingly the model performs better on the MEME dataset than on MAMI, although it wasn't trained on this dataset. On further inspection of the MEME dataset, we realized that its content are on average significantly more hateful and less nuanced than the MAMI dataset, which could explain why they are classified more precisely. Overall, these results validate the robustness and generalizability of the base classifier of the model.

However, the post-hoc explainability layers we implemented, specifically CLIP-Surgery, saliency maps and LIME, failed to provide meaningful insights. Outputs were often noisy and lacked interpretive clarity as illustrated on Figure 2.

This may be due to architectural incompatibilities and modality-specific challenges. CLIP-surgery assumes a ViT-B/16 backbone and is tuned to the spatial attention structure and patch granularity of that architecture. The use of ViT-L/14 in our classifier introduces deeper layers and different positional embeddings, which can distort the token-level spatial alignment required for meaningful visual explanations. Moreover, CLIP was pre-trained on broad internet image–text pairs and not specifically on hate-speech like content, meaning abstract concepts like "shaming" or "misogyny" may not be reliably represented in the shared embedding space, reducing the validity of the semantic alignment for our use case.

Similarly LIME is not well-suited for deep multimodal embeddings such as those used by CLIP. It relies on perturbing interpretable input features and observing corresponding output changes. However CLIP encodes text into dense, abstract representations via a Transformer making it difficult to isolate specific input components. Because individual token are not independently interpretable in CLIP's joint embedding space, perturbations often yield nonsensical embeddings and unpredictable model behavior. As for saliency map, according to [2, 9], gradient-based methods fail with CLIP-based models because CLIP's image-level training lacks region-text alignment, so gradients highlight irrelevant areas. In addition, inconsistent self-attention and redundant features further reduce interpretability. Another explanation might be that late concatenation of image and text embeddings weakens direct pixel-to-output attribution, causing saliency maps to focus on non-meaningful regions.

Model	Data set	Task A F1 Score [%]	Task B F1 Score[%]
TIB-VA	MAMI (Test set)	73.40	73.10
SRCB	MAMI (Test set)	83.4	73.1
Ours	MAMI (Validation)	89.40	73.12
Ours	MAMI (Test set)	72.87	71.10
Ours	MEME (Test set)	96.60	76.47

Table 1. Our experimental results for the different datasets

4.2. Limitations

The most important limitation is that all explored explainability solutions focus on either text or image and do not consider the interplay between the modalities. This strongly limits their explaining power since many memes present benign text and image which only become offensive when combined. Furthermore since the model is capable of integrating both modalities and recognizing such offensives memes, it means that the current solutions can never pretend to explain the complete behavior of the model.

The second one is that no dedicated metrics were implemented to evaluate the quality of the explanations, as preliminary inspection already suggested that they were not informative. This anecdotal validation is admittedly weak, and a more rigorous evaluation framework is required in future work to systematically assess explanation quality.

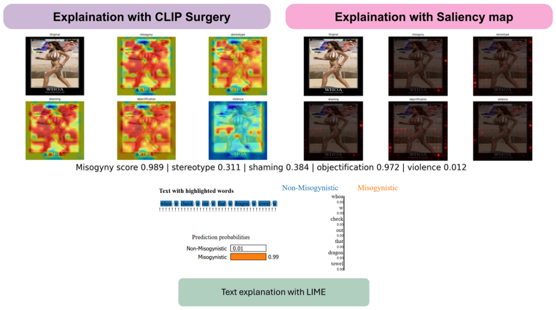

Figure 2. Overview of our explanability results with a misogynistic meme

4.3. Future work

Modifying the current architecture to support BERT-style embeddings could integrate better with LIME, thus offering more interpretable token-level explanations. Also, incorporating CLIP Surgery into the model pipeline (not just as an ad-hoc explainability tool) could improve the coherence of visual embeddings and facilitate the application of gradient-based methods such as saliency maps. Finally leveraging the self-attention weights in CLIP to generate patch similarity heatmaps may provide more fine-grained and spatially aligned visual explanations.
Evaluation metrics following the Co-12 principles proposed in [11] should be devised to quantify the explanaibility methods' performances. As an example, Contrastivity tests such as the Target Sensitivity Check could help to verify if the generated explanations meaningfully differ between each misogyny sub-type label (i.e. task B) and Completeness tests such as the Fidelity and the Deletion Check would quantify how much of the model's reasoning is captured by the explanation methods.

5. Conclusion

In this project, we explored the integration of post-hoc explainability methods into a state-of-the-art multimodal classifier for misogynistic meme detection. While the base classifier demonstrated strong performance across both the MAMI and MEME datasets, our attempts to apply explainability techniques (CLIP-Surgery, saliency maps, and LIME) yielded lacking results. The lack of interpretability can be attributed to architectural mismatches, the abstract nature of CLIP embeddings, and the separation of modalities in both modeling and explanation. Our findings highlight the challenges of adapting existing interpretability tools to complex multimodal settings and underscore the need for future work on modality-aware explanation frameworks.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?" Explaining the Predictions of Any Classifier," *arXiv preprint arXiv:1602.04938*, 2016.

[2] Y. Zhao *et al.*, "Gradient-based visual explanation for transformer-based clip," in *Proceedings of Machine Learning Research*, vol. 235, 2024.

[3] Y. Wang, T. Zhang, X. Guo, and Z. Shen, "Gradient based feature attribution in explainable ai: A technical review," *arXiv preprint arXiv:2403.10415*, 2024. License: CC BY 4.0.

[4] S. Hakimov, G. S. Cheema, and R. Ewerth, "Tib-va at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes," *arXiv preprint arXiv:2204.06299*, Apr. 2022.

[5] E. Fersini, D. Nozza, P. Rosso, *et al.*, "Semeval-2022 task 5: Multimedia automatic misogyny identification," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, M. Palmer, N. Schneider, S. Singh, and S. Ratan, eds.), (Seattle, United States), pp. 533–549, Association for Computational Linguistics, July 2022.

[6] J. Zhang and Y. Wang, "Srcb at semeval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, M. Palmer, N. Schneider, S. Singh, and S. Ratan, eds.), (Seattle, United States), pp. 585–596, Association for Computational Linguistics, July 2022.

[7] N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, and B. R. Chakravarthi, "MISTRA: Misogyny Detection through Text–Image Fusion and Representation Analysis," *Natural Language Processing Journal*, vol. 7, p. 100073, June 2024.

[8] M. Z. U. Rehman, S. Zahoor, A. Manzoor, M. Maqbool, and N. Kumar, "A context-aware attention and graph neural network-based multimodal framework for misogyny detection," *Information Processing & Management*, vol. 62, p. 103895, Jan. 2025.

[9] Y. Li, H. Wang, Y. Duan, J. Zhang, and X. Li, "A closer look at the explainability of contrastive language-image pre-training," *arXiv preprint arXiv:2304.05653*, 2024.

[10] F. Gasparini, G. Rizzi, A. Saibene, and E. Fersini, "Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content," *Data in Brief*, vol. 39, p. 107720, 2021.

[11] M. Nauta *et al.*, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–42, 2023.