MilaNLP at SemEval-2022 Task 5: Using Perceiver IO for Detecting Misogynous Memes with Text and Image Modalities

Giuseppe Attanasio and **Debora Nozza** and **Federico Bianchi**
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
{giuseppe.attanasio3,debora.nozza,f.bianchi}@unibocconi.it

Abstract

Warning: This paper contains examples of language that some people may find offensive or upsetting.

In this paper, we describe the system proposed by the *MilaNLP* team for the Multimedia Automatic Misogyny Identification (MAMI) challenge. We use Perceiver IO as a multimodal late fusion over unimodal streams to address both sub-tasks A and B. We build unimodal embeddings using Vision Transformer (image) and RoBERTa (text transcript). We enrich the input representation using face and demographic recognition, image captioning, and detection of adult content and web entities. To the best of our knowledge, this work is the first to use Perceiver IO combining text and image modalities. The proposed approach outperforms unimodal and multimodal baselines.

1 Introduction

Monitoring and detecting hateful content online is of paramount importance to limit the spread of hate, misconception, and prejudice. Content on social media platforms poses several challenges, from fast-paced, large-scale generation requiring automatic solutions, to ever-changing information mediums, like internet memes.

Misogynous memes are an unfortunately common phenomenon. Based on sexist preconceptions, these memes target and degrade women for humor. This intricate nature makes them hard to detect with classical computational models as hate is conveyed by associating known visual concepts with specific textual wording. In the Multimedia Automatic Misogyny Identification task (Fersini et al., 2022) novel systems are required to detect misogynous memes in English. The task is divided into two sub-tasks. Sub-task A requires solving the sole misogynous meme identification (i.e., a binary task). Sub-task B requires recognizing more specific categories, namely *stereotype*, *shaming*, *objectification*, and *violence*. Figure 1 (top) shows an example from the dataset.

F: White, 20-29, Female - Count: 1
W: "No men meme"
C: "A poster of a woman in a kitchen"
A: False

Figure 1: Sample from the training set (top) annotated with *Misogynous* and *Stereotype* labels. We enriched the meme with additional information (bottom), namely detected faces (F), web entities (W), caption (C), and adult content (A).

We propose a novel architecture where unimodal[1] components extract salient information from the meme. We present all information to a late fusion layer that distills it into a latent representation. We use renowned unimodal encoders networks and Perceiver IO (Jaegle et al., 2022) as the late fusion layer. Notably, while jointly learning from all modalities, Perceiver IO easily extends to multi-task learning. To the best of our knowledge, this work is the first to use Perceiver IO combining text and image modalities. We hence effectively address, with the same architecture, both sub-tasks A and B.

[1]We use *unimodal* whenever a single modality is involved, e.g., when dealing with Image content only. By extension, *multimodal* refers to a mixture of unimodals, e.g., Image and Text.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 654 - 662

The proposed system outperforms both unimodal and multimodal baselines. The results show that Perceiver IO is an effective and efficient method to jointly fuse input representations from different modalities in multi-task setups. However, we achieved sub-par performance against other competing solutions. We ranked 25th (13 F1 points worse than the best system) out of 69 competing teams on sub-task A and 15th (4 F1 points worse than the best system) out of 42 competing teams on sub-task B. Our system is not specialized in either of the sub-tasks and hence it under-performs against task-engineered solutions. We report a brief error analysis in Section 5.

We release the code to replicate our experiments at https://github.com/MilaNLProc/milanlp-at-mami.

2 Background

In the last years, the task of hate speech detection has attracted considerable attention from the Natural Language Processing and Computer Vision communities. Among the research work in this area, only a limited number of approaches have focused on the problem of misogyny detection, which is a concrete problem in social media platforms. Nozza (2021) shows that hate speech detection models do not transfer across different hate speech targets, further demonstrating the need for ad-hoc misogyny detection approaches and datasets. Indeed, the corpora made available as part of shared tasks (Fersini et al., 2018, 2020b; Basile et al., 2019; Mulki and Ghanem, 2021) enabled a variety of NLP approaches to the problem of automatic misogyny detection on Twitter posts (Indurthi et al., 2019; Fersini et al., 2020a; Attanasio and Pastor, 2020; Lees et al., 2020, inter alia).

The Multimedia Automatic Misogyny Identification task focuses on the problem of misogyny detection with the new perspective of multimodality. The most similar research effort in the direction of hateful memes detection is the Hateful Memes Challenge (Kiela et al., 2020) and the MMHS150K corpus (Gomez et al., 2020).

On the other hand, multi-modal models that combine image and text are now becoming incredibly popular in Natural Language Processing (Cao et al., 2020; Lu et al., 2019; Tan and Bansal, 2019; Radford et al., 2021; Bianchi et al., 2021; Su et al., 2020, inter alia) due to their capabilities to solve zero-shot tasks.[2]

In this paper, we focus on the use of Perceiver IO (Jaegle et al., 2022) to combine the information coming from different sources such as the meme image, the text, and other features.

3 System overview

Following recent work addressing similar tasks (Pramanick et al., 2021; Zhu, 2020; Lee et al., 2021, inter-alia), we decompose the multimodal learning of hateful memes into two stages. First, we embed with unimodal encoders different input sources. Next, we adopt a multimodal late fusion to jointly learn from different modalities. With this setup, we tackle both sub-tasks A and B with no additional data other than the one provided for the task.

We build unimodal streams using pretrained, modality-specific encoder networks. Each encoder contributes an input representation to the subsequent fusion layer. We then use Perceiver IO (Jaegle et al., 2022) as a late fusion approach over the concatenation of modality-specific representations. Perceiver IO produces a structured, multi-dimensional output. We leverage this ability to jointly learn misogyny and other relevant aspects (e.g., aggressiveness, objectification, etc.) from the input. With that, we effectively solve both sub-tasks A and B of the challenge using a single model in a multi-task learning setting.

The system architecture is shown in Figure 2. To the best of our knowledge, this is the first attempt to use Perceiver IO as a multimodal late fusion layer for multi-task learning.

In the following sections, we further describe what type of unimodal source we considered (Section 3.1) and how we adapted Perceiver IO to multimodal, multi-task learning (Section 3.2).

3.1 Unimodal streams

The provided memes are characterized by two features: the meme image and the transcription of the over-imposed text. Building on ideas from recent work (Blaier et al., 2021; Pramanick et al., 2021), we enrich the meme with semantic information including image caption, face and demographics, detection of adult content, and web entities. We report a sample in Figure 1.

[2] See also Frank et al. (2021) for a detailed study and ablation analysis on the capabilities of these models.

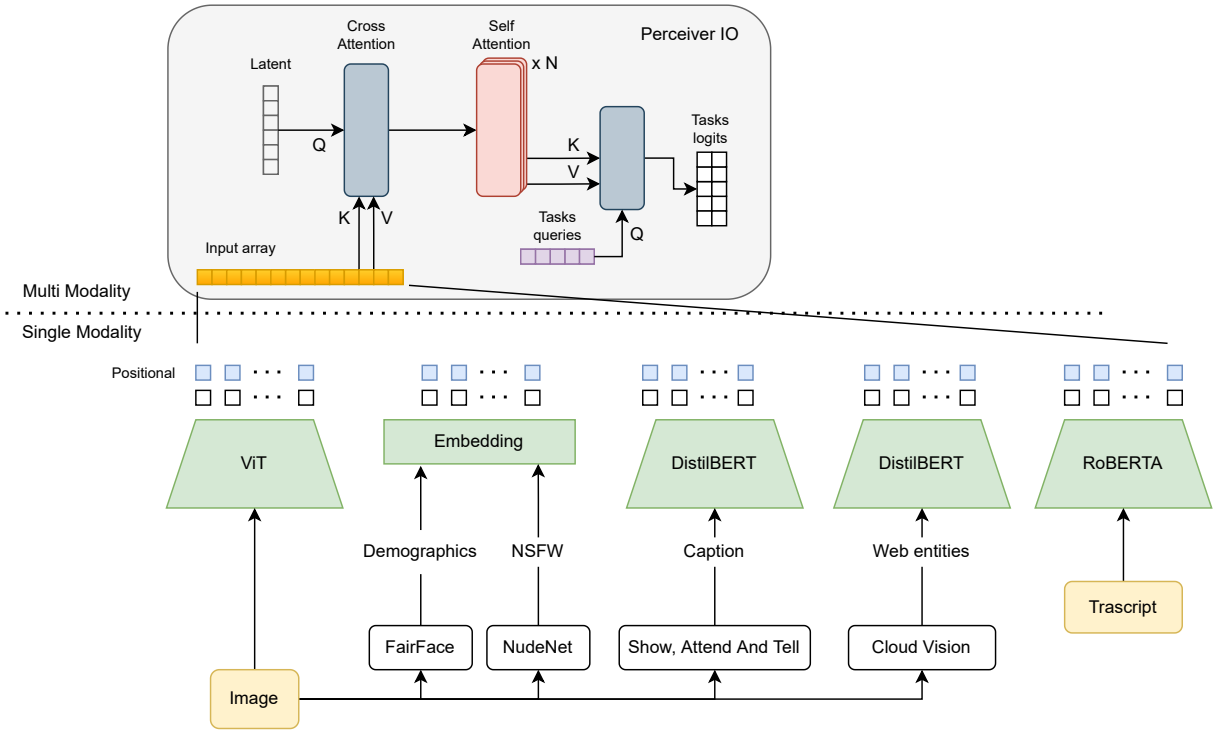


Figure 2: System overview. Light yellow boxes are the raw meme components. Green shapes are unimodal encoders. White boxes are external models used to enrich the meme. The darker yellow array represents the input array to the Perceiver IO fusion layer.

We derive a different input stream from each of the features of the semantically-augmented meme (Figure 2, bottom).

Image We encode raw images using a pretrained Vision Transformer (Dosovitskiy et al., 2021) (ViT). In this stream, the input image is divided into patches of 16x16 pixels and fed through a stacked transformer architecture. We use the last hidden token embeddings as the output of the stream.

Text Transcript We encode the text transcript using a pretrained RoBERTa (Liu et al., 2019) and use the last hidden token embeddings as the output of the stream. We choose RoBERTa based on its performance and its hurtful sentence completion (HONEST) score (Nozza et al., 2021, 2022b).

Face and Demographics Some images contain one or more faces. We use a pretrained FairFace (Karkkainen and Joo, 2021) model to detect faces and demographics. For each face found, we extract three categorical attributes, namely *Age*[3], *Gender*, and *Ethnicity*. Figure 1 (bottom, F) reports an example of face and demographics extracted with FairFace from a training sample. We encode face demographics with a simple embedding layer.

[3]FairFace produces age ranges, e.g., *60-75*, *75+*.

Adult Content We use NudeNet[4], a pretrained classification and detection model for nudity detection and censoring, to detect if the image contains adult content. We encode this information with additional embeddings.

Image caption We include an automatically generated image caption to have a textual description of the meme content. We use a pretrained Show, Attend and Tell model (Xu et al., 2015) to generate the caption for both training and test memes. Figure 1 (bottom, C) reports an example of an automatically generated caption. We encode the caption using a pretrained DistilBERT (Sanh et al., 2019).

Web Entities We use Google Cloud Vision API to detect web entities starting from the meme image.[5] As a summary of the extracted entities, we use the textual field `best guess labels` provided by the API. We encode this text using a pretrained DistilBERT. Note that this is a different model than the one used for captions.

[4]https://github.com/notAI-tech/NudeNet

[5]https://cloud.google.com/vision/docs/detecting-web

3.2 Late fusion with Perceiver IO

We concatenate all the information extracted by unimodal streams into a single input array (Figure 2, top). Ideally, this array contains all the raw unrouted information necessary for the task. We use a Perceiver IO-based late fusion layer for information distillation and multi-task learning.

Perceiver IO builds on Perceiver (Jaegle et al., 2021) a modality-independent neural architecture with two crucial advantages over unimodal models. First, it is designed not to leverage inductive biases as in modality-specific architectures. Because of that, it effectively learns from the input of any shape and nature. Second, Perceiver distills inputs in a much smaller latent representation for memory efficiency. On top of that, Perceiver IO (Jaegle et al., 2022) adds a decoding step to produce a structured output of arbitrary size and semantics. The output is generated using decoder queries cross-attending the latent representation. The number of queries and their dimension defines the output shape.

We use this feature of Perceiver IO to address both sub-tasks A and B at once. Specifically, we define five different *task queries*, one per characteristic of the misogynous meme identification problem, i.e., *misogyny*, *shaming*, *aggressiveness*, *objectification*, and *violence*. In this multi-task setup, we generate logits and extract probability distributions for each of the five aspects.

4 Experimental setup

The provided training set counts 10,000 samples. Class labels are balanced on misogyny ($p_1 = 0.5$) but unbalanced on *shaming* ($p_1 = 0.18$), *stereotype* ($p_1 = 0.28$), *objectification* ($p_1 = 0.22$), and *violence* ($p_1 = 0.095$). We validated our models and baselines using three-fold cross-validation over the training set. We measure performance with F1 Macro on the binary misogyny detection task.

For ViT, RoBERTa, and DistilBERT, we use implementations and checkpoints from the transformers library (Wolf et al., 2020) and HuggingFace Hub. We use monolingual English checkpoints for text models. For Perceiver IO, we use the lucidrains's PyTorch implementation.[6]

Unimodal encoders For ViT, we used the `google/vit-base-patch16-224-in21k` checkpoint. We used the standard feature processor which entails 1) resizing to a maximum shape of 224x224 and channel normalization. We also augmented the image using color jitter ($hue = 0.1$), random horizontal flip ($p = 0.5$), affine transformations,[7] contrast ($p = 0.3$) and equalization ($p = 0.3$) variations. We discard the CLS last token embedding and use the sequence of the remaining 196 tokens as the image representation.

For RoBERTa, we used the `roberta-base` checkpoint and its standard tokenizer padding and truncating up to a maximum of 32 tokens. We performed text augmentation via token removal ($p = 0.15$).

For DistilBERT, we used the `distilbert-base-uncased` checkpoint and its standard tokenizer padding and truncating up to a maximum of 32 tokens. Note that this configuration is duplicated for both the text transcript and web entity streams.

Perceiver IO We did not use a Perceiver IO pretrained checkpoint. We manually fine-tuned the number of latent variables in $\{64, 128, 256, 512, 1024\}$, latent dimension in $\{128, 256, 512, 1024\}$, and number of self-attention layers in $\{6, 12\}$, and settled on the configuration `[265, 512, 6]`. Self-attention layers are applied sequentially and do not share weights. We use 5 decoder queries to produce a structured output of shape $(5, 2)$. We project the output using a final linear layer and consider the result as the tasks logits. We also tried to use a different linear projection layer per task but with no measurable improvement in performance.

Training setup We trained the entire system end-to-end, i.e., we jointly optimized all unimodal encoders *and* Perceiver IO. Following Bianchi et al. (2021), we tried to freeze the encoders for an arbitrary number of steps and then unfreeze them, with no significant improvement.

We manually tuned the learning rate in the range $[10^{-5}, 10^{-6}]$. We then used 10^{-5} with linear decay and 10% of total steps as warmup steps. We set weight decay to 10^{-2}. We trained the system for four epochs using Focal Loss (Lin et al., 2020) to account for class imbalance. We set alpha to 0.25 and gamma to 2.

[6] https://github.com/lucidrains/perceiver-pytorch

[7] For the full set of affine transformations please refer to our repository.

Model	F1
Lexicon BoW + LR	66.4
RoBERTa	77.4
Vision Transformer	73.1
Perceiver IO	82.9
Perceiver IO (FWCA)	**83.3**

Table 1: F1 performance (%) on our three-fold cross-validation setup of unimodal (top) and multimodal (bottom) models in the misogyny identification task (sub-task A).

Model	F1
Best system	**83.4**
Perceiver IO (FWCA)	69.9
Provided baseline	64.0
Best system	**73.1**
Perceiver IO (FWCA)	69.3
Provided baseline	62.1

Table 2: Performance on sub-task A (top) and sub-task B (bottom) compared to best systems and *baselines* provided by task organizers.

5 Results

The provided test set counts 1,000 samples. Target labels are balanced in terms of misogyny but imbalanced for the rest of the categories. Unbalancing is slightly more marked than the training set. We report the prior frequency of the positive class as p_1 in Table 3.

In the following, we compare the performance of the proposed system which encodes detected faces (F), web entities (W), the image caption (C), and adult content detection (A). We refer to the system as Perceiver IO (FCWA). It achieves an overall F1 (macro) score of 69.9% in sub-task A and an F1 (weighted) score of 69.3% in sub-task B. We rank 31th (13 F1 points worse than the best system) on sub-task A and 18th (4 F1 points worse than the best system) on sub-task B. The system outperforms several baselines provided by the task organizers. On sub-task A, we outperform 1) sentence embeddings from a pretrained Universal Sentence Embedding model, 2) an image classifier fine-tuned from a VGG model, and 3) a classifier based on the concatenation of the first two representations plus a single layer neural network. On sub-task B, we outperform 1) a multi-label model based on the concatenation of deep image and text embeddings and 2) a hierarchical multi-label model based on text representations. Results are reported in Table 2.

Category	F1	P	R	p_1
misogynous	69.91	75.07	71.10	0.50
shaming	65.98	65.69	66.31	0.15
stereotype	67.79	68.58	67.34	0.35
objectification	70.06	71.72	69.30	0.35
violence	74.29	82.73	70.27	0.15

Table 3: Performance on the test set in terms of F1 (%), Precision (P), and Recall (R) with macro averaging. Prior frequency (p_1) of the positive class in the training set.

5.1 Quantitative analysis

In the proposed dataset, misogyny characteristics (e.g., *shaming* or *objectification*) apply only to misogynous content. Hence, we consider performance on sub-task A (binary misogyny detection) as a proxy for the overall quality of the model configuration. We report performance on our cross-validation over the training set in Table 1.

We compared our system with internal unimodal baselines. For the textual modality, we tested a Bag-of-Word (BoW) representation[8] extracted from the HurtLex lexicon (Bassignana et al., 2018) fed to a Logistic classifier and RoBERTa. For the visual modality, we tested Vision Transformer. Our system outperforms by a large margin these unimodal baselines.

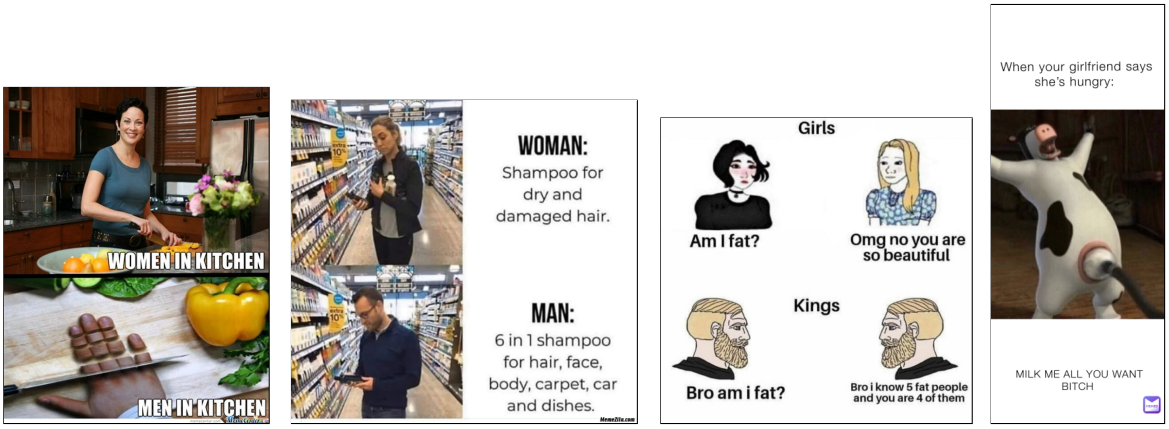
Further, we validated the intuition of semantically enriched memes with input ablation. Results are reported in Table 1. Specifically, we removed all streams but the encoders of the raw image and the transcript (Perceiver IO). Cross-validation results show that our enriched input (Perceiver IO (FWCA)) improves classification performance.

5.2 Error analysis

In the post-evaluation phase of the task, we studied the labels predicted by our system on the 1,000 test memes. Results are reported in Table 3 separately by category.

The system effectively learned all categories (F1 is always greater than 65%) but with differences. The misogynous identification task has a 70% F1

[8]We use a binary presence matrix, i.e., the rows are the transcript of the meme, the columns each of the terms of the lexicon, and the cell is 1 if the transcript contains the term at least once or 0 otherwise.



(a) (b) (c) (d)

Figure 3: A sample of false positives from the test set on the misogyny identification task.

(a) (b) (c) (d)

Figure 4: A sample of false negatives from the test set on the misogyny identification task.

score composed of a promising 75% in precision and 71% recall. While performance between *stereotype* and *objectification* are close, the system underperformed in the *shaming* category. We think this behavior is due to the broad kind of both visual and textual content that can convey *shameful* messages and is hence more difficult to learn. Finally, the model performed best in *violence*, where it is probably simpler to identify visual clues that let intend a violent message.

We also manually inspected the errors of our model. We conducted the analysis separately by false positives (Figure 3) and false negatives (Figure 4). In the following, we speculate on the possible causes of these errors.

Weak Adult Content detection We noticed several wrong annotations extracted by NudeNet in both false positives and false negatives (e.g., all memes in Figure 3 are labeled as NSFW). We believe this might have introduced noise in our training that further propagated in classifying test memes.

Bias towards composite memes Several memes have a composite nature. They contrast some behavior or reaction of two groups (e.g., boys and girls) using a predefined structure. A typical example is achieved by organizing one group on top of another (Figure 3b and 3c).

We noticed several composite memes among false positives. We argue this kind of meme has a strong association bias with the positive class (*misogynous = 1*). Indeed, we believe that, in absence of relevant information, the system leverages the structure of the memes and wrongly produces a positive prediction.

Hard stereotypes Several memes contain non-hateful wording and image content. However, they convey misogynous messages because the combination of image and text leverages a well-known stereotype about women. We argue that the correct classification of these memes must involve either a solution explicitly modeling the stereotype or sufficient training data to become *aware* of it. We call "hard stereotypes" those memes whose stereotypical message is not backed by sufficient training data. We noticed several hard stereotypes in our misclassified memes. Figure 4 shows four examples of those.

Gender Bias We noticed several memes misclassified as false negatives depicting only a man in the

scene. We believe the system might have learned a spurious bias associating the presence of men in the image with the negative class (*misogynous = 0*).

5.3 Further considerations

Quality of generated information In our experiments, automatically generated image captions often describe coherently the content of the image, although we do not expect them to generalize well to complex and diverse concepts.

Instead, FairFace extracts almost always precise face counts and age, ethnicity, and gender for each face even in crowded images (there are cases of 14 people in a single image). Web entities are often significant but provide coarse-grained information (e.g., "internet cat meme").

Unconditioned prediction During the post-evaluation phase, we identified a potential weakness in the proposed architecture.

Decoding queries in Perceiver IO are independent by design. This behavior enhances the model's generalization capabilities as it needs to learn internal representations useful for all outputs. However, it also prevents any conditioning between them. In our task, the *misogynous* aspect influences the remaining ones in the sense that only misogynous memes are characterized by one or more categories for sub-task B. We are not explicitly modeling this condition and hence probably hindering the performance.

6 Conclusion

We addressed both sub-tasks A and B of the Multimedia Automatic Misogyny Identification shared task with a novel Perceiver IO-based system. We take advantage of pretrained encoders and external services to extract and enrich with salient information the input meme. Then, we use Perceiver IO as a multimodal, multi-task late fusion layer of several unimodal streams. To our knowledge, this is the first time Perceiver IO has been used to combine text and image modalities.

The proposed system outperforms unimodal and multimodal baselines but underperforms against more specialized, task-specific competitor systems. We ranked 25th out of 69 competing teams on sub-task A and 15th out of 42 competing teams on sub-task B. In future work, we will explore improved input preprocessing (e.g., for the OCR-based text provided), and model ensembling. Additional effort should be put into identifying and mitigating unintended bias that may be present in our multimodal misogyny detection model following approaches proposed for text modality (Nozza et al., 2019; Attanasio et al., 2022a,b). The development of these multi-modal hate speech classifiers can be useful for the automatic evaluation of large pre-trained models (Nozza et al., 2022a).

Acknowledgements

This project has partially received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA). The authors are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. Giuseppe Attanasio did part of the work at the DataBase And Data Mining Group at the Polytechnic University of Turin. Computing resources were provided by the SmartData@PoliTO center on Big Data and Data Science.

References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022a. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL2022 (Forthcoming)*. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022b. Benchmarking Post-Hoc Interpretability Approaches for Transformer-based Misogyny Detection. In *Proceedings of the First Workshop on Efficient Benchmarking in NLP*. Association for Computational Linguistics.

Giuseppe Attanasio and Eliana Pastor. 2020. PoliTeam @ AMI: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identificationin italian tweets. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 48–54. Accademia University Press.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*.

Efrat Blaier, Itzik Malkiel, and Lior Wolf. 2021. Caption enriched samples for improving hateful memes detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9350–9358, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision – ECCV 2020*, pages 565–580, Cham. Springer International Publishing.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Giulia Boifava. 2020a. Profiling Italian misogynist: An empirical study. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France. European Language Resources Association (ELRA).

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2018)*, 12:59.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020b. AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2022. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. *Disentangling Hate in Online Memes*, page 5138–5147. Association for Computing Machinery, New York, NY, USA.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw @ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of Seventh Evaluation Campaign of Natu-*

ral Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Bologna, Italy. CEUR.org.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. 2020. Focal loss for dense object detection. volume 42, pages 318–327, Los Alamitos, CA, USA. IEEE Computer Society.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Hala Mulki and Bilal Ghanem. 2021. Working notes of the workshop arabic misogyny identification (armi-2021). In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 7–8, New York, NY, USA. Association for Computing Machinery.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, , and Dirk Hovy. 2022a. Pipelines for Social Bias Testing of Large Language Models. In *Proceedings of the First Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022b. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2048–2057. JMLR.org.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.