Information Processing and Management 60 (2023) 103474

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Recognizing misogynous memes: Biased models and tricky archetypes

Giulia Rizzi [a,b], Francesca Gasparini [a], Aurora Saibene [a], Paolo Rosso [b], Elisabetta Fersini [a,*]

[a] DISCo, University of Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy
[b] Universitat Politècnica de València, Camino de Vera, Valencia, Spain

ARTICLE INFO

Keywords:
Misogyny identification
Meme
Bias estimation
Bias mitigation

ABSTRACT

Warning: This paper contains examples of language and images which may be offensive.

Misogyny is a form of hate against women and has been spreading exponentially through the Web, especially on social media platforms. Hateful content towards women can be conveyed not only by text but also using visual and/or audio sources or their combination, highlighting the necessity to address it from a multimodal perspective. One of the predominant forms of multimodal content against women is represented by memes, which are images characterized by pictorial content with an overlaying text introduced a posteriori. Its main aim is originally to be funny and/or ironic, making misogyny recognition in memes even more challenging. In this paper, we investigated 4 unimodal and 3 multimodal approaches to determine which source of information contributes more to the detection of misogynous memes. Moreover, a bias estimation technique is proposed to identify specific elements that compose a meme that could lead to unfair models, together with a bias mitigation strategy based on Bayesian Optimization. The proposed method is able to push the prediction probabilities towards the correct class for up to 61.43% of the cases. Finally, we identified the most challenging archetypes of memes that are still far to be properly recognized, highlighting the most relevant open research directions.

1. Introduction

In the last few years, with the spread of social media, the opportunities to share people's experiences and opinions have increased exponentially. At the same time, the phenomenon of hatred is also growing accordingly: social media users share hateful content, often involuntarily or unawarely, towards different targets and minorities (Kocoń et al., 2021). One of the more targeted groups is represented by women. This phenomenon has grown in recent years generating significant research activity. In 2016, Poland published a book analyzing online harassment, abuse and violence (Poland, 2016), highlighting how Internet, thanks to its anonymity, is a fertile ground for hate speech and violence especially against women. In 2018 the European Parliament Department for Citizens' Rights and Constitutional Affairs reported a study on cyber violence and hate speech online against women that analyzes the societal context and root causes (Van Der Wilk et al., 2018). More recently, Musso, Proietti, and Reynolds (2020) demonstrated how violence against women is cross-national and transdisciplinary, and does not change significantly in different geo-cultural realities. In particular, online Violence Against Women with respect to classical forms of violence allows an increase in reproducibility, ubiquity and uncontrollability of its dissemination. In a survey related to violence against women, the European

[*] Corresponding author.
E-mail addresses: g.rizzi10@unimib.it (G. Rizzi), francesca.gasparini@unimib.it (F. Gasparini), aurora.saibene@unimib.it (A. Saibene), prosso@dsic.upv.es (P. Rosso), elisabetta.fersini@unimib.it (E. Fersini).

https://doi.org/10.1016/j.ipm.2023.103474
Received 21 March 2023; Received in revised form 17 July 2023; Accepted 31 July 2023
Available online 9 August 2023

Union Agency for Fundamental Rights reported that 20% of young women (18–29) have experienced cyber sexual harassment in Europe (Nevala, 2014; Rights, 2014). This phenomenon increased during the pandemic of COVID-19, because the quarantine shifted the orientation of community sexual violence to technology-facilitated sexual one as reported in Almenar (2021), and Jatmiko, Syukron, and Mekarsari (2020). As further evidence of the spread of this problem, the report published by UNESCO in 2022 (Collett et al., 2022) revealed that, according to a survey based on more than 900 journalists and media workers in 125 countries, nearly 73% of the interviewed women have experienced online violence.

Hateful content towards women can be conveyed not only through text, but also using visual and/or audio sources or their combination, highlighting the necessity to deal with a multimodal problem. One of the predominant forms of multimodal content against women is represented by memes, which are images characterized by pictorial content with an overlaying text introduced a posteriori (according to YPulse social media behavior survey (YPulse, 2019) performed in 2019, 75% of 13–36-year-old and 79% of 13–17-year-old share memes). While several approaches have been presented in the state of the art for tackling the problem of automatic misogyny identification applying text-based models (Badjatiya, Gupta, Gupta, & Varma, 2017; Fersini et al., 2020; Pamungkas, Basile, & Patti, 2020), up to our knowledge methods that rely only on image-based analysis have not been yet developed to detect misogyny on multimodal contents. Furthermore, few research investigations have been devoted to the multimodal perspective (Fersini et al., 2022). Detecting misogynous memes is therefore in its infancy (Fersini, Gasparini, & Corchs, 2019), where simple unimodal and multimodal approaches have been investigated to understand the contribution of textual and visual cues. Inspired by the previous findings highlighted in Fersini et al. (2019), where textual information has been shown to be a good marker that significantly represents and predicts misogynous content of memes, and considering the increasing attention in bias analysis related to machine learning models, the proposed paper contributes as follow:

- **RQ1 — Which modality contributes more to recognize misogynous memes?** Unimodal and multimodal models have been presented to automatically detect misogynous content, demonstrating that textual information representing the pictorial content of a meme can provide additional insights when combined with the text of the meme itself. In particular, the proposed unimodal and multimodal approaches have been designed to exploit on one hand the existing pre-trained (visual and/or textual-based) models and on the other hand to guarantee a reduced number of parameters to induce the classification models.
- **RQ2 — What are the elements that compose a meme that could lead a model to be biased?** A bias estimation strategy is proposed to identify specific elements that compose a meme that could lead to unfair models, revealing some known *stereotypes* and new patterns. In particular, some terms and image tags, such as *dishwasher* for misogynous memes and *memeshappen* for not misogynous ones, have been identified as elements that could have a strong impact on the model bias.
- **RQ3 — Is Bayesian Optimization an effective technique to mitigate the bias of a model affected by the elements previously identified?** A Bayesian Optimization (BO) strategy has been investigated for tuning the model hyperparameters to mitigate the model bias, showing remarkable capabilities in specific settings. In particular, the proposed multimodal approaches, when BO is adopted, are able to reduce the bias on controversial memes while still maintaining good recognition performance on the rest.
- **RQ4 — What are the archetypes of memes that represent an open challenge for misogynous detection systems?** A detailed error analysis has been performed to identify archetypes of memes that represent an open challenge for misogynous detection systems that still needs to be addressed for future research. As main findings, several challenging types of memes have been identified, highlighting that the most relevant open issue refers to memes that include implicit sexual references and therefore the necessity to approach the unexplored problem of multimodal compositional reasoning.

The paper is organized as follows. In Section 2 an overview of the state of the art is presented, outlining both unimodal and multimodal approaches adopted to address the misogyny identification task. In Section 3 the adopted benchmark dataset is presented. In Section 4 both unimodal and multimodal approaches are introduced to determine which source contributes more on the detection of misogynous memes. In Section 5, a bias estimation technique is proposed to identify specific elements that compose a meme that could lead to unfair models, together with a bias mitigation strategy based on Bayesian Optimization. The most frequent archetypes underlying misclassified misogynous memes have been identified in Section 6, which describes several main open research directions. Finally, a discussion about impacts on practical and theoretical aspects, as well as on the limitations of the proposed approach, is reported in Section 7, while in Section 8 conclusions are drawn.

2. Related work

The state of the art of automatic misogyny identification in online environments is quite prolific, but still in its infancy. As summarized in Rodríguez, Díaz-Ramírez, Miranda-Vega, Trujillo, and Mejía-Alvarez (2021), contributions addressing hate content, in particular misogynous content, focus on different domains: online detection, offline detection, safety and education. Especially, the majority of contributions in online detection use Artificial Intelligence, mainly Machine Learning algorithms, to detect hateful content. Different approaches have been developed, based also on the different modalities where misogyny can be manifested: text, images, etc.

The majority of works address misogyny detection from a **linguistic perspective**. The first contribution in this area is represented by Anzovino, Fersini, and Rosso (2018); the authors propose a corpus of misogynous tweets and an exploratory investigation on language features and machine learning models for detecting and classifying misogynous language. Moreover, considerable contributions in the field are represented by AMI (Automatic Misogyny Identification) IberEval 2018, AMI EVALITA 2018, and AMI at Evalita2020 (Fersini et al., 2020). In fact, the first two shared tasks mainly focused on tackling the problem of misogyny in Twitter, in three different languages, namely English, Italian, and Spanish. Instead, AMI at Evalita2020 challenge was based on Italian tweets only and organized in two subtasks: (1) identifying misogyny and aggressiveness and (2) evaluating the fairness of the model.

In this context, several approaches have been proposed to address the problem of automatic misogyny identification. The majority of the approaches are based on pre-trained models (Butt, Ashraf, Sidorov, & Gelbukh, 2021; Dutta, Majumder, & Naskar, 2021; Ta, Rahman, Najjar, & Gelbukh, 2022), such as BERT (Devlin, Chang, Lee, & Toutanova, 2019) and USE (Cer et al., 2018), while a few others are based on lexical-resources (Frenda et al., 2018; García-Díaz, Cánovas-García, Colomo-Palacios, & Valencia-García, 2021; Jiang, Yang, Liu, & Zubiaga, 2022; Pamungkas, Basile, & Patti, 2022) or on a combination of the above (Calderón-Suarez, Ortega-Mendoza, Montes-Y-Gómez, Toxqui-Quitl, & Márquez-Vera, 2023). Transfer learning mechanisms have been adopted, to face the lack of labeled data. In particular, Samghabadi et al. (2020) developed an end-to-end neural model using attention on top of BERT incorporating a multi-task learning paradigm to simultaneously face aggression and misogyny detection tasks. Bashar, Nayak, and Suzor (2020) demonstrated that a pre-trained LSTM model could be properly adapted using transfer learning to detect misogynistic tweets using a small training dataset. More recently, Calderón-Suarez et al. (2023), adapted cross-domain and embedding-based methods to transfer knowledge from song phrases to social media text that could convey misogynistic messages. In a recent work (Pamungkas et al., 2020), a joint-learning model based on LSTM and BERT has been proposed to work with low-resource languages. Their main contribution relates to the adoption of different datasets on sexism, hate speech, and offensive language to detect misogynous content through cross-domain classification methods. The experiments have shown that lexical features such as sexist slurs and woman-related words are among the most predictive features to detect misogyny.

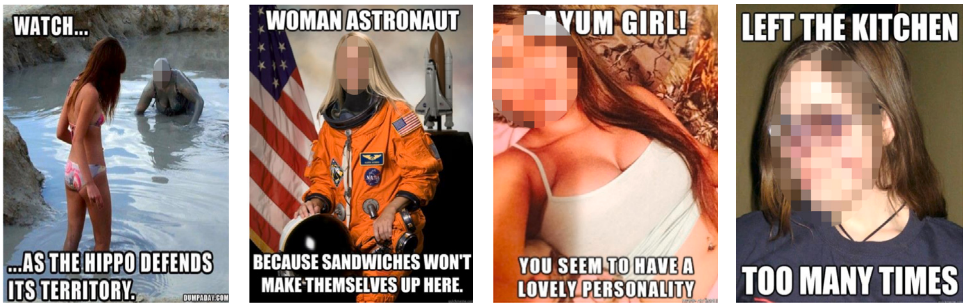
Despite several multimodal content being spread on the web with hateful and harmful intentions, most of the developed unimodal approaches are based on text analysis, while a few works faced the problem considering only a **visual perspective**. Ling et al. (2021), for instance, studied which visual elements make the memes viral on social media, but without considering their potential offensive content. Gandhi et al. (2020) investigated from a visual perspective, how to detect offensive and non-compliant content/logos in product images. Focusing on misogyny, and in particular on sexism recognition, a preliminary work was done in Gasparini, Erba, Fersini, and Corchs (2018), where a visual binary classifier trained has been adopted in conjunction with a textual one to detect sexist content in multiple languages. It may be worth noting that in pornography detection, where women could also be the object of offensive attacks, visual classifiers have been widely adopted (Gangwar, González-Castro, Alegre, & Fidalgo, 2021; Hor et al., 2021; Lin, Qin, Peng, & Shao, 2021; Tabone et al., 2021). Investigations on pornographic detection typically use deep learning techniques such as convolutional neural networks and object detection models to identify male and female genital organs and sexual activity (AlDahoul et al., 2021).

Recently, researchers have begun to address the problem related to automatic misogyny identification from a **multimodal** point of view, where misogyny can be manifested by a combination of textual and visual content. Today, one of the most widespread forms of multimodal communication, especially on social media, is given by memes which, despite their ironic intent, can also be used to communicate hateful content towards different targets. An important contribution in this scenario has been given by the recent *Hateful Memes (HM) Challenge: Detecting Hate Speech in multimodal Memes* (Kiela et al., 2020) promoted by Facebook, where one of the targets of hateful memes was a woman. The competition aimed to automatically identify hateful content in memes from a multimodal point of view. Recent investigations (Hee, Lee, & Chong, 2022; Kiela et al., 2020) indicate that the best multimodal approach is based on the combination of textual and visual features given by transformer-based architectures (i.e. ViLBERT Lu, Batra, Parikh, & Lee, 2019 and Visual-BERT Li, Yatskar, Yin, Hsieh, & Chang, 2020). Those approaches are compared to other unimodal and multimodal strategies and proposed as a baseline for the challenge itself, confirming the need for a multimodal analysis to detect hate content. Despite the good performance reached by those approaches (∼75% of accuracy), those models still perform poorly with respect to human ability, underlying the necessity of new machine learning models.

Less attention has been paid to the specific form of hatred that targets women, where only a few papers address the problem of misogyny detection in memes either from a unimodal or multimodal point of view. A first approach to identifying hate content towards women is represented by Fersini et al. (2019), which evaluates both a unimodal and multimodal approach to understand the impact of the two components (image and text). Their research demonstrates that a late-fusion multimodal strategy is the best approach to face the complexity of the represented message. Further investigations from the same authors (Fersini, Rizzi, Saibene, & Gasparini, 2021) have led to introduce a multimodal approach that considers both visual (in the form of captioning) and textual information. The suggested approach obtains better classification results compared both to unimodal classifiers and to the multimodal state-of-the-art benchmark: Visual-BERT.

The main contribution of addressing the problem of automatic detection of misogyny in memes is represented by the *SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification* (MAMI) (Fersini et al., 2022), which explores the detection of misogynous memes on the web by proposing two sub-tasks: one focused on misogynous content detection in memes (Sub-task A) and the other devoted to identify types of misogyny (Sub-task B). Regarding the models adopted by the participants, the majority of them exploited pre-trained models, distinguished in text-based, mostly grounding on BERT-like architectures and image-based models, where the most adopted ones are based on Visual-BERT (Li et al., 2020).

In addition to the shortcomings of the models proposed in the state of the art, several authors highlight how the proposed models may be subject to **bias** that could affect the real performance of the models (Angwin, Larson, Mattu, & Kirchner, 2016; Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Dixon, Li, Sorensen, Thain, & Vasserman, 2018; Park, Shin, & Fung, 2018; Shen, Li, Bouadjenek, Mai, & Sanner, 2023; Song, Giunchiglia, Li, Shi, & Xu, 2023; Spinde et al., 2021; Wiegand, Ruppenhofer, & Kleinbauer, 2019; Yalcin & Bilge, 2022). We can distinguish bias estimation and mitigation strategies, according to the considered source of information. For what concerns the bias that can be introduced from the textual sources, the majority of the investigations related


WATCH...
...AS THE HIPPO DEFENDS ITS TERRITORY.
DUMPADAY.COM
WOMAN ASTRONAUT
BECAUSE SANDWICHES WON'T MAKE THEMSELVES UP HERE.
AYUM GIRL!
YOU SEEM TO HAVE A LOVELY PERSONALITY
LEFT THE KITCHEN
TOO MANY TIMES

(a) Shaming (b) Stereotype (c) Objectification (d) Violence

Fig. 1. Examples of misogynous memes.

to hateful content are focused on two different types of bias, i.e. racial bias (Elsafoury, Wilson, Katsigiannis, & Ramzan, 2022; Garg, Schiebinger, Jurafsky, & Zou, 2018; Manzini, Chong, Black, & Tsvetkov, 2019) and gender bias (Caliskan, Ajay, Charlesworth, Wolfe, & Banaji, 2022; Chaloner & Maldonado, 2019; Field & Tsvetkov, 2020; Kaneko, Imankulova, Bollegala, & Okazaki, 2022; Razo & Kübler, 2020). In both cases, the bias estimation metrics and the related mitigation policies are based on a predefined set of target (racial or gender) seed words that are used to quantify and minimize the bias at the dataset or model level. Regarding the bias that can be introduced from the visual point of view, analogous approaches have been proposed. In fact, the majority of the investigations related to the image processing research areas are focused on racial bias (Hirota, Nakashima, & Garcia, 2022; Zhao, Wang, & Russakovsky, 2021) and gender bias (Kyriakou, Barlas, Kleanthous, & Otterbacher, 2019; Schwemmer et al., 2020) that can be generated by state-of-the-art image tagging or captioning models. Analogously to what is performed for the text, also in image tagging and image captioning, the bias estimation and mitigation techniques are based on associations between a given target group, such as asian, black people, or women, and a specific target semantic concept, such as occupation, physical attribute or activity. A more general approach, related to the representation bias in image recognition models, is presented in Li and Vasconcelos (2019).

In a multimodal setting, metrics to estimate the bias of a misogyny identification model, and techniques to mitigate it are still missing.

3. MAMI benchmark dataset

In order to address the problem of Misogyny Identification in Memes, we adopted the benchmark recently introduced for the Multimedia Automatic Misogyny Identification (MAMI) challenge hosted at SemEval 2022 (Fersini et al., 2022). The dataset is composed of 10,000 memes for training and 1000 memes for testing, collected by focusing on the following main types of misogyny:

- *Shaming*: The practice of criticizing women who violate expectations of behavior and appearance regarding issues related to gender typology (such as "slut shaming") or related to physical appearance (such as "body shaming") (Van Royen, Poels, Vandebosch, & Walrave, 2018). This category focuses on content that seeks to insult and offend women because of some characteristics of the body or personality.
- *Stereotype*: a stereotype is a fixed, conventional idea or set of characteristics assigned to a woman (Eagly & Mladinic, 1989). A meme can use an image of a woman according to her role in society (role stereotyping), or according to her personality traits and domestic behaviors (gender stereotyping).
- *Objectification*: A practice of seeing and/or treating a woman like an object (Szymanski, Moffitt, & Carr, 2011).
- *Violence*: A meme that indicates physical and/or a call to violence against women (Andreasen, 2021).

Examples of the above-mentioned types of misogynous memes are presented in Fig. 1. More details about the meme selection strategy and annotation are reported in Fersini et al. (2022).

4. Unimodal vs multimodal approaches

In order to analyze different modalities and their contribution to identify misogynous memes, both unimodal and multimodal approaches are investigated. The proposed approaches have been designed to take advantage of pre-trained (visual and/or textual-based) models induced on large datasets, thus a small number of parameters needs to be updated in the classification layer to recognize misogynous and not misogynous memes. We report in the following subsections the investigated models and their experimental comparison.

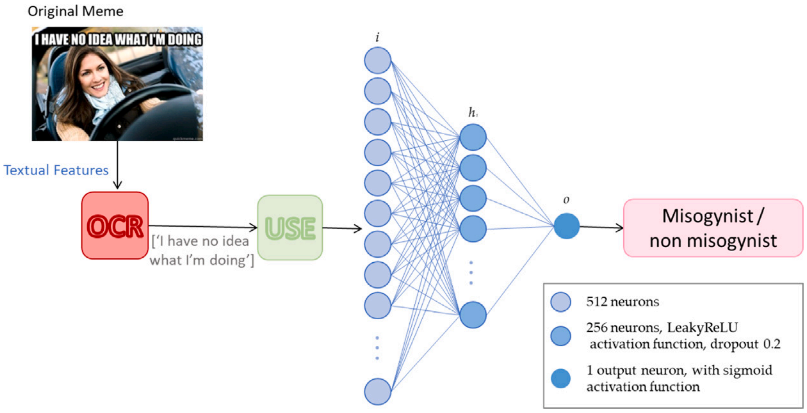

Original Meme
I HAVE NO IDEA WHAT I'M DOING
Textual Features
OCR
['I have no idea what I'm doing']
USE
i
h
o
Misogynist / non misogynist
512 neurons
256 neurons, LeakyReLU activation function, dropout 0.2
1 output neuron, with sigmoid activation function

Fig. 2. Text-based model schema.

4.1. Unimodal models

Text-based. The textual transcription of the text contained in the memes is used to train the first baseline unimodal model. In particular, the Text-based model takes as input the text embedding obtained via Universal Sentence Encoder (USE) (Cer et al., 2018), and it learns a very simple neural network characterized by a decreasing size of neurons within a single hidden layer. In particular, the proposed model is based on the following architecture:

- Input layer determined by the sentence encoder, with 512 input neurons;
- Hidden layer (dense) with 256 internal neurons, with LeakyReLu activation function, dropout 0.2;
- Output layer (dense) with 1 output neuron, with sigmoid activation function.

The objective function is based on the *binary cross-entropy loss* and it is minimized by means of the *Adam* (Kingma & Ba, 2015) optimization algorithm. The training phase is of 100 epochs and a batch size of 64 instances. The architecture of the proposed text-based unimodal classifier is reported in Fig. 2.

Image tag-based. A further unimodal classifier that can be simply designed is based on the identification of objects contained in the memes, i.e. one or multiple human-readable concepts. In order to extract such image tags, the Clarifai API (Clarifai, 2023) has been used. Starting from these tags, a subset of 14 human-readable concepts (tags) has been identified to capture specific characteristics typically related to misogynous content. The subset of tags has been selected considering experts' observations on a benchmark dataset (Fersini et al., 2022) of memes and the automatically assigned tags. We report in Table 1 the association between the selected tags and the type of misogyny.

Table 1
Selected tags to represent the different misogyny expressions.

Type of misogyny	Tags
Shaming	woman, animal, cartoon
Stereotype	car, kitchen, kitchen utensil, crockery, broom, dishwasher, woman, child, nudity, dog
Objectification	woman, man, nudity, cartoon, cat
Violence	woman, man

In particular, for capturing *stereotypes*, few tags have been proposed to capture those visual elements that could be considered as a proxy of women's conventional ideas: the fact that women are typically perceived as housekeepers (e.g. in the kitchen, doing laundry, cleaning or cooking), as mothers (with children) led us to consider those tags typically associated to these roles such as broom, dishwasher and child. In order to capture memes related to *shaming* behaviors, the proposed tags have been identified to capture those visual elements that could be considered as a proxy of insulting women by comparing them with others (e.g. overweight women compared with big animals, 'ugly' woman compared with horses, etc.). For what concerns memes denoting women *objectification*, we propose a few tags that symbolize a woman with her physical attributes or depict her as a thing external to the thinking mind. To this purpose, tags capturing nudity or depicting women as objects by means of cartoons or comparison with cats are proposed. Regarding memes related to *violence* towards women, they typically contain men and women as subjects, suggesting the corresponding visual tags. The underlying model is based on a convolutional neural network with a customized inceptionV2 architecture (Ioffe & Szegedy, 2015).

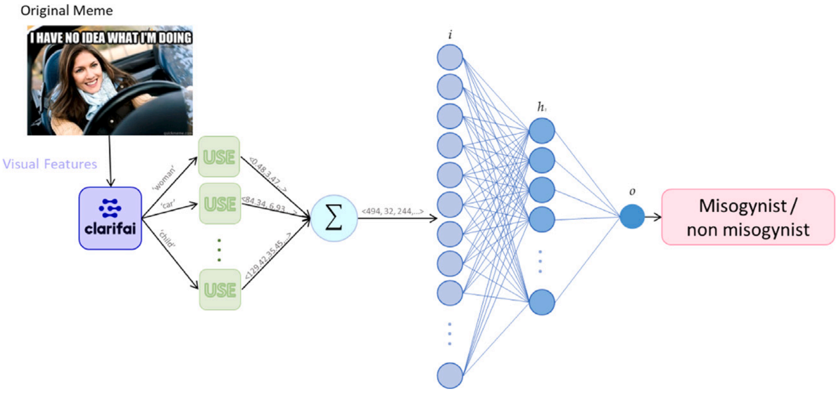


Fig. 3. Image Tag-based model schema.

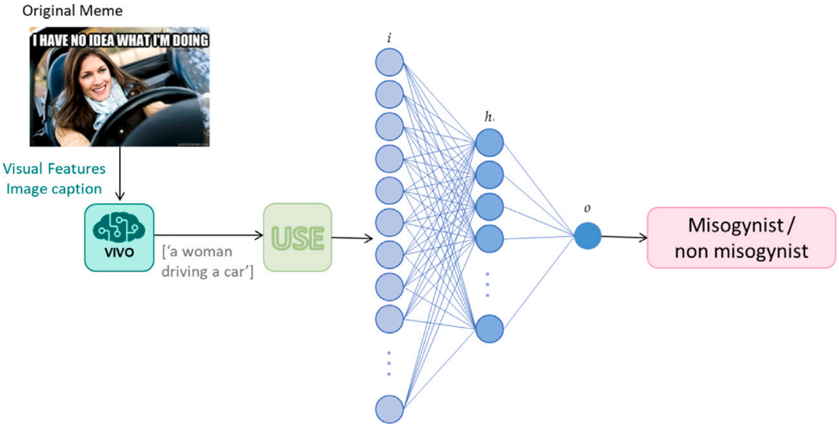

Original Meme
I HAVE NO IDEA WHAT I'M DOING
Visual Features
Image caption
VIVO
['a woman driving a car']
USE
i
h
o
Misogynist / non misogynist

Fig. 4. Image Caption-based model schema.

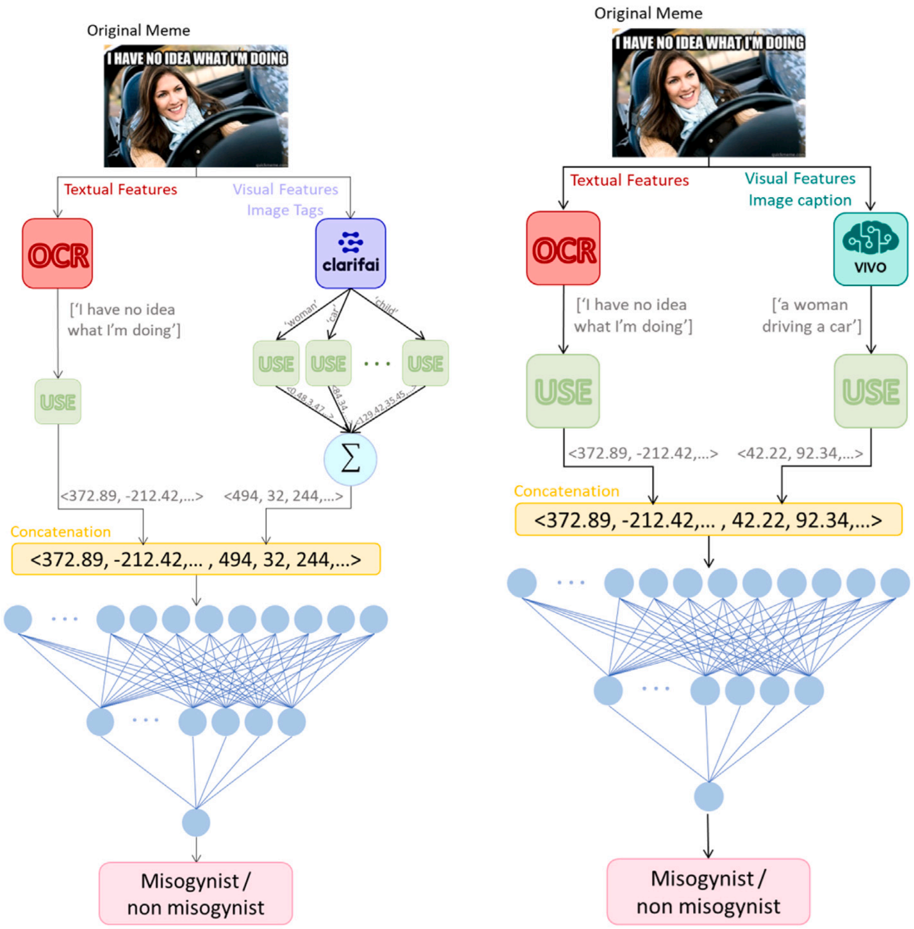
In order to identify the selected tags, we exploited the model included in the Clarifai API, which provides as output 9098 concepts that have been mapped as shown in Appendix B. Once the proposed tags have been extracted, they are represented as dense feature vectors by exploiting the pre-trained Universal Sentence Encoder (Cer et al., 2018). In particular, since memes are typically composed of multiple tags, we aggregate the tag embeddings in a vector representation of the meme by a simple average as follows:

$$\vec{\mathbf{x}}_m = \frac{\sum_{i=1}^{|T_m|} \vec{\mathbf{v}}_i}{|T_m|} \tag{1}$$

where $\vec{\mathbf{x}}_m$ is the final embeddings of the meme, $\vec{\mathbf{v}}_i$ is the tag embedding vector associated to each tag i identified in a meme and T_m is the entire set of tags in the meme. The obtained latent representation of a meme ($\vec{\mathbf{x}}_m$) is provided as input to the following tag-based model reported in Fig. 3, which has the same architecture as the text-based one.

Image caption-based. The last unimodal classifier is based on image captioning, which allows the generation of a textual description for each meme. In order to generate such captions, we adopted the *VIsual VOcabulary (VIVO)* (Hu et al., 2021) state-of-the-art model. VIVO performs a pre-training step in the absence of caption annotations by breaking the dependency of paired image-caption training data and using a large amount of paired image-tag data. VIVO exploits a visual vocabulary to train a Transformer model (initialized using BERT-base (Devlin et al., 2019), with an additional linear layer) to align image-level tags with their corresponding image region features. Once a caption is generated by the VIVO model, its latent representation is computed through USE and is provided as input to a subsequent neural architecture that has the same characteristics as the previous models. The Caption-based model is reported in Fig. 4.

Text-BERT. BERT (Devlin et al., 2019) is currently one of the most extensively used approaches for text analysis, and it has also been used for hateful content detection, getting the best results in several tasks. BERT is intended to jointly condition on both left and right context in all the layers in order to pre-train deep bidirectional representations from the unlabeled text. The pre-trained BERT model can be extended with just one extra output layer for a variety of tasks, including question-answering and language inference. Even though several BERT-based models have emerged in recent years, BERT continues to be one of the most widely used baseline



(a) Multimodal Text-Tags Model (MTT)

(b) Multimodal Text-Caption Model (MTC)

Fig. 5. Structure of the proposed models.

models. In this paper, we adopted Text-BERT, the unimodal model adopted in Singh et al. (2020). In our investigation, it takes as input the transcription of the text within each meme to create its embedding representation. Text-Bert has been fine-tuned on the MAMI benchmark dataset to address the identification of misogynous and not misogynous memes.

4.2. Multimodal approaches

Misogyny in a meme can be represented by a single modality (text or image), but in most cases, the misogynous message is conveyed by their combination. In fact, in many memes, the misogynous message can be understood by joining the meaning of what is represented by text and images. In those cases, considering only a single modality is not enough, because both modalities are necessary to correctly comprehend the whole misogynous message. To this purpose, a few multimodal approaches are investigated in the following paragraphs.

Early fusion approach. A first simple multimodal approach combines the information given by the considered sources of information by a concatenation of the corresponding latent representations. In particular, a first model called MTT (Multimodal Text and Tags) concatenates the embedding of the text within the meme, with the embedding of the tags, obtained by means of Eq. (1). A second model called MTC (Multimodal Text and Caption), analogously, concatenates the embedding of the text within the meme, with the embedding of the meme caption. In both cases, the final vectors originated by the concatenations are provided as input to the neural model reported in Figs. 5(a) and 5(b). Similarly to the unimodal approaches, the model architecture is composed of an input layer of 1024 neurons, a hidden layer of 512 neurons with LeakyReLU activation function and dropout of 0.2, and finally an output layer with Sigmoid activation function.

Table 2
Model performance using a 10-fold cross-validation on the 10000 memes of the MAMI training data. (+) refers to the misogynous class, while (–) to the not misogynous one. **Bold** denotes the best performance for each performance measure, while underlined text related to the AUC measure denotes the best corresponding unimodal and multimodal approaches. (*) denotes that the corresponding model obtains results that are statistically different with respect to all the other models.

Model		Precision		Recall		F1 measure		AUC
		P–	P+	R–	R+	F–	F+	
Unimodal	Text-based	0.78	0.77	0.77	0.78	0.77	0.78	0.853*
	Image Tag-based	0.65	0.68	0.72	0.60	0.68	0.64	0.709*
	Image Caption-based	0.59	0.69	0.78	0.46	0.67	0.55	0.675*
	Text-BERT	0.81	0.80	0.78	0.81	0.79	0.80	0.884
Multimodal	MTT	0.82	0.81	0.80	0.82	0.81	0.81	0.889
	MTC	0.80	0.79	0.79	0.80	0.80	0.80	0.876*
	Visual-BERT	**0.85**	**0.83**	**0.83**	**0.85**	**0.84**	**0.84**	**0.919***

Table 3
Model performance on the 1000 memes of the MAMI test data. (+) refers to the misogynous class, while (–) to the not misogynous one. **Bold** denotes the best performance for each performance measure, while underlined text related to the AUC measure denotes the best corresponding unimodal and multimodal approaches. (*) denotes that the corresponding model obtains results that are statistically different with respect to all the other models.

Model		Precision		Recall		F1 measure		AUC
		P–	P+	R–	R+	F–	F+	
Unimodal	Text-based	0.70	0.62	0.54	0.77	0.61	0.69	0.726
	Image Tag-based	0.63	0.60	0.53	0.69	0.58	0.64	0.651*
	Image Caption-based	0.55	0.58	**0.65**	0.48	0.60	0.52	0.609*
	Text-BERT	0.72	0.62	0.50	0.80	0.59	0.70	0.718
Multimodal	MTT	**0.80**	**0.67**	0.57	0.86	**0.67**	**0.75**	**0.786***
	MTC	0.75	0.65	0.55	0.81	0.63	0.72	0.757*
	Visual-BERT	**0.80**	0.61	0.44	**0.89**	0.57	0.73	0.773*

We experimented USE and BERT for text embedding. According to the results reported in Appendix A, USE has promising AUC values with a small variability in terms of Interquartile Range. Therefore USE has been selected as the embedding strategy enclosed in the multimodal models.

Visual-BERT. One of the most widely used multimodal approaches available in the state of the art, which has been used for hateful content detection, is represented by Visual-BERT (Kiela et al., 2020; Li et al., 2020). Visual-BERT is a model trained on the COCO dataset[1] to jointly learn a representation that captures both visual and language characteristics. To accomplish this task, it uses a BERT-like transformer (Devlin et al., 2019) to create an embedding for each image–text pair. In order to predict whether a meme is misogynous or not, the pre-trained Visual-BERT model has been fine-tuned on the MAMI challenge dataset. Notice that the default configuration for Visual-BERT COCO has been used in all the experiments. Image features have been extracted using MMF (Singh et al., 2020) based on ResNet-152.

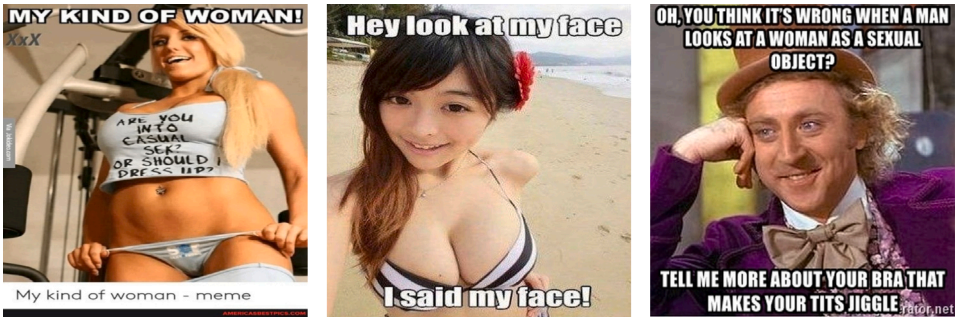
4.3. Model comparison

In order to understand which modality contributes more to recognize misogynous memes (**RQ1**), we compared the proposed unimodal and multimodal approaches. We report in Tables 2 and 3 the model performance on a 10-fold cross-validation on the training data (10,000 memes) and on the test set (1000 memes) respectively of the MAMI benchmark dataset. The performance on the test set has been measured by splitting the training data in 10 folds, using 9 folds for training, 1 for validation and the test set for the performance estimation.

The performance is reported in terms of Precision (P), Recall (R) and F1-Measure (F) for both labels, i.e. misogynous (+) and not misogynous (-). Additionally, we also report the AUC value for each compared model.

Considering Table 2, a T-test has been performed to compute the statistical equality with a pairwise analysis. The paired Student's t-test is based on the null hypothesis that the two analyzed models are identical in performance. The obtained results show that, according to the AUC measure, all the proposed models are statistically different considering a value of alpha equal to 0.05, with the exception of Text-BERT and MTT for which we cannot reject the null hypothesis. By analyzing the AUC measure reported in Table 2, we can easily highlight that among the unimodal models, the ones based on text (Text-based and Text-BERT) strongly

[1] https://cocodataset.org/#home


MY KIND OF WOMAN!
XxX
ARE YOU INTO CASUAL SEX? OR SHOULD I DRESS UP?
My kind of woman - meme
AMERICASBESTPICS.COM
Hey look at my face
I said my face!
OH, YOU THINK IT'S WRONG WHEN A MAN LOOKS AT A WOMAN AS A SEXUAL OBJECT?
TELL ME MORE ABOUT YOUR BRA THAT MAKES YOUR TITS JIGGLE

(a) Misogynous meme correctly classified by the Image Caption-based model.

(b) Misogynous meme correctly classified by the Image Tags-based model.

(c) Misogynous meme correctly classified by the Text-BERT model.

Fig. 6. Examples of some test set misogynous memes correctly classified by the unimodal models.

outperform both the Image Tags- and Image Caption-based models, showing that the textual component is more informative than the other sources of information. However, focusing on the precision and recall measures related to both misogynous (+) and not misogynous (–) labels, we can point out that the image-based unimodal models have unbalanced precision and recall for the two considered classes. In particular, those models have high precision but low recall on one class and low precision but high recall on the other. Instead, the text-based unimodal models (i.e., Image Text-based and Text-BERT) have a good proportion of precision and recall on both classes.

Fig. 6 depicts some examples related to the typical misogynous memes that are correctly classified by the unimodal models.

In particular, Fig. 6(a) reports a representative example for the Image Caption-based model. The extracted caption is *"a woman in a tank top"*, which is able to capture the pictorial content of the meme. In fact, the presence of women with revealing outfits seems to be easily coupled with misogynous content and well-modeled by this unimodal approach. A similar meme can be also observed in Fig. 6(b), which is associated with the caption *"a person smiling at the camera"*, and the tags *"woman"* and *"nudity"*. In this case, the produced caption is not able to capture the keywords usually associated with misogyny, while the Tags-based model is able to detect specific information from the pictorial content related to typical misogynous memes (e.g., nudity and cleavage). Finally, Fig. 6(c) is a clear example of a successful Text-BERT classification, where the misogynous information is mainly expressed by the superimposed text. The lexical components of the text are in fact related to misogyny and report an interaction between a man and a woman while introducing the concept of sexual objectification. In this case, the pictorial information, decoupled from the text, is unable to provide information on misogyny, having that the extracted caption refers to the depicted actor *"Gene Wilder in a suit"* and that the tag is only *"man"*.

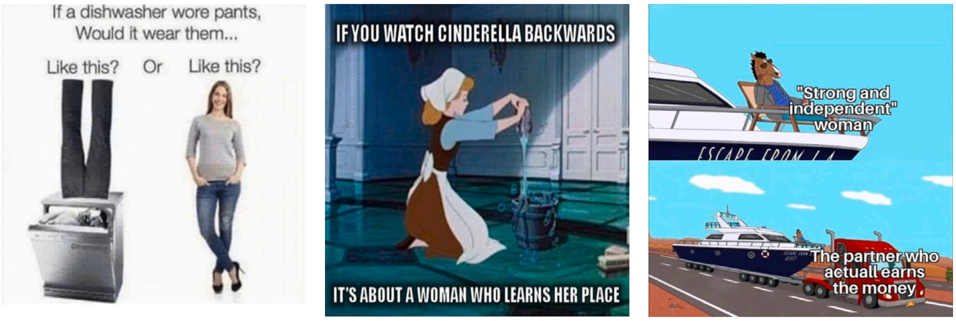
Focusing more on the comparison of all the considered models, we can also determine that the multimodal approaches outperform the unimodal ones. Although in some cases, unimodal approaches could be sufficient because a single modality could be enough to correctly predict the target label, a multimodal approach could push the performance even further, especially when the misogynous message of the memes needs to be grasped by taking into account both sources. We can also point out that Visual-BERT is not only the model with the highest performance on (cross-validated) training data but also achieves a balanced trade-off between precision and recall both on misogynous and not misogynous memes.

In order to understand why the multimodal approaches work better than the unimodal ones, we report a few misclassified examples in Fig. 7. For instance, the meme shown in Fig. 7(a) has been misclassified only by the unimodal text-based approach, because the text within the meme is harmless when decontextualized from the whole scene. Instead, the pictorial content seems to be able to vehicle misogyny due to the presence of a woman and a dishwasher, which are commonly present in misogynous memes.

Similarly, the meme shown in Fig. 7(b) has been misclassified by the visual-based unimodal models only. In this case, while the text within the meme is sufficient to detect the misogynous content, the proposed image is a harmless cartoon scene that cannot lead the visual-based models to recognize the misogynous message. Finally, the meme shown in Fig. 7(c) has been misclassified by all the unimodal models, because the misogynous message can be grasped by a deep understanding of what is conveyed by image and text together, which are harmless if considered alone.

Table 3 reports the results on the test dataset. Notice that also in this case a T-test has been performed to compute the statistical equality with a pairwise analysis. The obtained results show that, according to the AUC measure, all the proposed models are statistically different considering a value of alpha equal to 0.05, with the exception of Text-based and Text-BERT for which we cannot reject the null hypothesis.

Moreover, we can observe two main aspects: (1) the Text-based approaches are still the best performing models among the unimodal ones, while for the multimodal settings, the Multimodal Text and Tags (MTT) approach performs better than the others (including Visual-BERT); (2) all the models making inference on the test set, analogously from the experiments on the cross-validate training data (reported in Table 2), have high precision but low recall on one class and low precision but high recall on the other, finally leading to recognition capabilities in terms of F1-Measure that are definitely higher on the misogynous class than the not

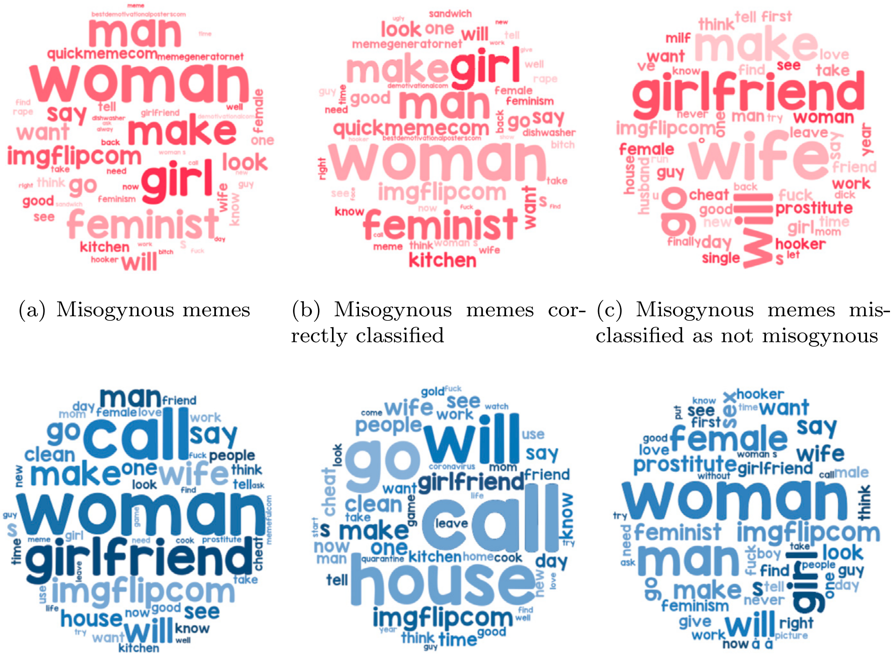

If a dishwasher wore pants,
Would it wear them...
Like this? Or Like this?
IF YOU WATCH CINDERELLA BACKWARDS
IT'S ABOUT A WOMAN WHO LEARNS HER PLACE
"Strong and independent" woman
ESCAPE FROM L.A.
The partner who actually earns the money

(a) Misogynous meme misclassified only by the unimodal text-based model.

(b) Misogynous meme misclassified only by the visual-based unimodal models.

(c) Misogynous meme misclassified by all the unimodal models.

Fig. 7. Examples of misclassified misogynous memes available in the (cross-validated) training dataset.



(a) Misogynous memes

(b) Misogynous memes correctly classified

(c) Misogynous memes misclassified as not misogynous

(d) Not misogynous memes

(e) Not misogynous memes correctly classified

(f) Not misogynous memes misclassified as misogynous

Fig. 8. Word clouds of lemmas on the training dataset. Word clouds reflect the relative frequency of the words in each considered set or subset of memes. Therefore, frequencies are not directly comparable.

misogynous one. This suggests a potential distortion in the model mainly due to a strong association of words within the memes to the misogynous class in the training data. In order to verify this hypothesis, we performed a frequency analysis of the lemmas (i.e., the canonical form of each word also known as uninflected form) related to the text within the memes, the text of the caption and text of tags in the MAMI training dataset.

Therefore, the most frequent terms, distinguishing between misogynous and not misogynous memes, have been computed. The achieved results are represented in the word clouds in Fig. 8. Figs. 8(a) and (d) represent the word clouds of the misogynous and not misogynous memes on the entire training dataset. Figs. 8(b) and (e) represent the set of memes correctly classified by all the multimodal models according to the two class labels. Figs. 8(c) and (f) summarize the set of memes misclassified by all the multimodal models, again distinguishing between misogynous and not misogynous instances. Notice that a similar analysis has been performed both for image tags and for image captions. The obtained word clouds can be found in Appendix C.

Looking at the word cloud in Fig. 8 (f), which refers to not misogynous memes misclassified as misogynous, we can identify some terms, such as *girl*, that are frequent in the misogynous data (and therefore reported also in Fig. 8(a)). Additionally, the presence of the same terms in 8 (b), denotes that misogynous memes containing such terms have been correctly classified. Therefore, the considered multimodal models seem to be prone to predict the misogynous label by the presence of those terms that are commonly shared in Figures (a), (b) and (f). Moreover, an analogous behavior can be identified considering the not misogynous memes, whose word clouds are reported in Fig. 8(d). For instance, the term *girlfriend* which is frequent in the set of misogynous memes misclassified as not misogynous (Fig. 8(c)), is equally represented in the not misogynous meme word clouds (Figs. 8(d) and (e)).

One of the possible reasons behind the previously described false positives and false negatives could be the ***selection bias*** (Ousidhoum, Song, & Yeung, 2020) in the keyword-based collection of the dataset. Selection bias takes place when the chosen data are not reflective of real-world data distribution. There are several types of selection bias (Hibberts, Burke Johnson, & Hudson, 2012), among which: (i) Coverage bias, (ii) Non-response bias, and (iii) Sample bias. Coverage bias manifests when data are not selected in a representative fashion (e.g. a population has zero chance of being included in the sample). Non-response bias refers to an unrepresentative dataset due to participation gaps in the data-collection process. Sample bias happens because proper randomization is not achieved when collecting data. In fact, sample bias (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021) arises due to the nonrandom sampling of subgroups. As a consequence of sample bias, the trends estimated for one population may not generalize to data collected from a new population. Sample bias refers to a correlation between a (subset of) feature(s) and the label. If this correlation only occurs in a set of examples but not in the normal population, that set is biased.

Although as main findings related to **RQ1** we can argue that – despite the textual component brings a significant contribution to identify misogynous and not misogynous memes – a multimodal approach is necessary to address the task of misogyny identification, we also need to measure and mitigate a potential selection bias inherited by the model due to some specific textual and visual cues that can impact on the model capacity in a real environment.

5. Measuring and mitigating bias

Sample bias can be intuitively seen as an unintended behavior of the model that happens due to a strong misleading association of some features to the target classes. In order to validate the hypothesis that the considered models could suffer from such distortion, we propose: (i) a strategy to discover specific cues called *candidate biased elements* that could lead the model to perform an unintended behavior, (ii) a multimodal bias metric to quantify the distortion of the models, and finally, (iii) a related mitigation strategy based on Bayesian Optimization.

5.1. Identifying candidate biased elements

Misogynous and not misogynous memes, as highlighted by the word cloud analysis, are likely characterized by different elements strongly associated with the corresponding label. This suggests a deeper analysis of the selection bias that may affect the models due to the presence of different elements in misogynous and not misogynous memes respectively. In our case, the presence of specific elements can lead the model to an erroneous behavior by predicting a specific label due to the presence of those elements. This distortion in the investigated data-derived models can be in fact caused by an imbalance in the distribution, in relation to the prediction label, of specific terms or visual elements, which will be defined in general as *candidate biased elements* (and in particular as *candidate biased terms* and *candidate biased tags*).

Candidate biased elements. One approach for identifying candidate biased elements within a dataset is to compute the Polarized Weirdness Index (PWI) (Poletto, Basile, Sanguinetti, Bosco, & Patti, 2021), which is the ratio of the relative frequency of a given element (e.g. term) with respect to a class label over the relative frequency of that element with respect to the rest. Although PWI could be easily extended to address more than one modality, it has three main significant limitations: (1) it evaluates the distribution of a term across the entire corpus disregarding the presence of the element within a given sentence, (2) it ignores the context in which the element is used and therefore it does not consider the occurrence of other terms within the same sentence and (3) the metric is unbounded (ranging from zero to infinity), making its understanding a bit complex. For the sake of completeness, we have analyzed the terms with the highest PWI score computed on the considered MAMI dataset. In particular, we identified among the most biased terms (i.e. with a coefficient equal to infinite) a few erroneous cases. For instance, in the top list with the highest PWI measure, we found the word *saber* with a coefficient equal to +infinite, denoting a word that has a strong polarization with respect to the misogynous class. However, this word is referenced in a single misogynous meme, with a relatively low frequency. When computing the corresponding PWI metric, its coefficient goes to infinite. Analogous examples are related to the terms *WWIII* and *batter* which have a very low frequency in the corpus. However, since they are always associated with the misogynous label they are qualified as biased terms even if they have a very low impact on any predictive model (due to their low frequency). An even more extreme case is when a given term occurs one time and it is associated with a specific class. Also in this case, the PWI coefficient goes to infinite. To overcome these limitations, we propose a novel estimation for identifying candidate biased elements.

Given a multimodal dataset D (in our case the MAMI dataset of memes), e is a visual or textual element belonging to the set $\mathcal{T}$ that comprises all the terms and tags of D. A bias score $S(e)$ can be estimated for each element e according to the following formula:

$$S(e) = \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|} P(c^{+} \mid T_m) - P(c^{+} \mid \{T_m - e\}) \tag{2}$$

where $\mathcal{M}$ is the set of memes containing e, c^{+} represents the misogynous label and T_m denotes the set of terms and tags in a given meme m.

Table 4
Terms in the MAMI dataset with the highest positive and the lowest negative bias scores.

Biased terms			
Misogynous	Score	Not Misogynous	Score
Dishwasher	0.342	Memecrunch	−0.186
Chick	0.293	Weak	−0.187
Whore	0.289	Communism	−0.199
Demotivational	0.287	Valentine	−0.207
Diy	0.282	Anti	−0.212
Promotion	0.276	Template	−0.215
Bestdemotivationalposters	0.273	Developer	−0.219
Motivateusnot	0.269	Ambulance	−0.239
Imgur	0.268	Mcdonald	−0.289
Motifake	0.259	Memeshappen	−0.303

Table 5
Tags in the MAMI dataset with the highest positive and the lowest negative bias scores.

Biased tags			
Misogynous	Score	Not Misogynous	Score
Dishwasher	0.361	Crockery	−0.145
Broom	0.345	Kitchen-utensil	−0.155
Nudity	0.333	Animal	−0.190
Woman	0.165	Dog	−0.205
Car	0.022	Cat	−0.283

$P(c^+ \mid T_m)$ represents the probability of an instance (meme) m of being associated with the positive class label, given the terms and tags T_m within the instance (meme) itself, and can be computed as follows:

$$P(c^+ \mid T_m) = \frac{P(c^+) \prod_{i=1}^{|T_m|} P(t_i|c^+)}{\sum_{l \in \{c^+, c^-\}} P(l) \prod_{i=1}^{|T_m|} P(t_i|l)} \tag{3}$$

Analogously, $P(c^+ \mid \{T_m - e\})$ denotes the probability of an instance (meme) m of being associated with the positive label c^+, given the text (tags) present in the instance (meme), excluding the evaluated element e except for the term (tag) in analysis. This probability can be computed as follows:

$$P(c^+ \mid \{T_m - e\}) = \frac{P(c^+) \prod_{\substack{i=1 \\ t_i \neq e}}^{|T_m|} P(t_i|c^+)}{\sum_{l \in \{c^+, c^-\}} P(l) \prod_{\substack{i=1 \\ t_i \neq e}}^{|T_m|} P(t_i|l)} \tag{4}$$

The normalization at the denominator in Eqs. (3) and (4) allows considering every meme as equally important despite the number of contained elements. The proposed bias score ranges into the interval $[-1; +1]$. The higher positive the score, the more likely the element would induce bias towards the positive class (misogynous). On the other hand, the lower negative the score, the more likely the element would be associated with the negative class (not misogynous). Terms or tags with a score close to zero, are considered neutral regarding their impact on the choice of the label.

Tables 4 and 5 list the most important elements obtained by computing the bias score reported in Eq. (2), distinguishing in biased terms and biased tags. In particular, the terms that achieved the highest positive and the lowest negative score have been reported.[2] in Table 4. As we can see, the list of terms with the highest score for the misogynous class (Table 4) is composed of tokens that are typically associated with some specific misogyny categories like *dishwasher* and *chick* for stereotype and *whore* for objectification. The remaining tokens are websites that have been used to collect only misogynous memes, but not the not misogynous ones. Regarding the list of terms with the highest negative bias score for the not misogynous class, it is composed of tokens that are very general and commonly used in a variety of popular memes.

Concerning the visual component, a list of the most biased tags is reported in Table 5. We can observe that tags referring to common meme templates (like *cat* and *dog*) achieved a low negative score. On the other hand, tags associated to strong stereotypical visions of women achieved high scores (e.g. *dishwasher* and *broom*).

To summarize the main findings related to **RQ2**, we identified a set of relevant elements that are part of the memes both from a textual and a visual point of view, and that can lead models to produce biased predictions. In particular, the identified elements mostly referred to stereotypical ideas of women and to the sources from which the memes have been collected.

[2] The entire set of terms and tags with the corresponding bias scores have been reported at https://github.com/MIND-Lab/Debiasing-Misogynous-Meme-Recognition-Systems/Results The PWI index has been also made available for a comparative evaluation.

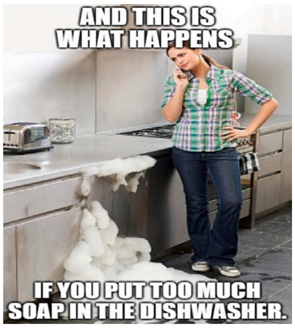


Fig. 9. An example of a meme in the synthetic dataset.

The five elements that achieved the highest positive and the lowest negative score have been selected as *candidate biased terms* and *candidate biased tags*. These elements have been used to create a synthetic dataset to measure the level of bias introduced in the models, to finally mitigate it. Biased elements related to the captions have not been considered for creating the synthetic dataset, having that such elements are not necessarily included in the caption due to the probabilistic generative process underlying the image-caption model.

5.2. Measuring the bias

In order to measure the bias of the models when making predictions, a *synthetic dataset* has been created. The identified biased elements have been used both to select memes from the web and to create new memes with specific characteristics that can effectively help to demonstrate the bias of the models given the presence of such elements.

In particular, let E^+ be the set of all the biased candidate elements with a positive score, which qualifies elements that are expected to introduce the bias towards the misogynous class. Also, let E^- be the set of all the biased candidate elements with a negative score, which qualifies elements that are expected to introduce the bias towards the not misogynous class.

Given a specific $e^+ \in E^+$, we collected misogynous and not misogynous memes according to the following criteria:

- a not misogynous meme is part of the synthetic dataset if it contains e^+ and it does not contain any element in E^-. This is to evaluate the impact of e^+ in introducing a bias towards the misogynous class in not misogynous memes;
- a misogynous meme is part of the synthetic dataset if it contains e^+ and it does not contain any other element in E^+. This is to verify if the model, given the presence of e^+, is able to perform well on misogynous memes.

Similarly, given a specific $e^- \in E^-$ we gathered misogynous and not misogynous memes according to the following strategy:

- a misogynous meme is part of the synthetic dataset if it contains e^- and it does not contain any element in E^+. This is to evaluate the impact of e^- to introduce a bias towards the not misogynous class in misogynous memes;
- a not misogynous meme is part of the synthetic dataset if it contains e^- and it does not contain any other element in E^-. This is to verify if the model, given the occurrence of e^-, is able to perform well on the not misogynous memes.

The proposed synthetic dataset is composed of 140 memes equally distributed between the class labels. An example of a selected meme is shown in Fig. 9. The meme depicts a funny situation related to the wrong usage of the dishwasher: it is not misogynous, but it contains many biased elements that will likely lead the model to perform a biased prediction. In fact, it contains the biased term *dishwasher* and the biased tags *dishwasher* and *woman* that achieved a positive score.

In order to measure if a given model is affected by sample bias we propose a **Multimodal Bias Estimation** (MBE) metric, which combines the area under the curve (AUC_{raw}) estimated on a test set belonging to the original raw dataset, which in this case is the MAMI dataset, and the area under curve estimated on the test set belonging to the synthetic dataset (AUC_{synt}):

$$MBE = \frac{1}{2}AUC_{raw} + \frac{1}{2}AUC_{synt} \tag{5}$$

where AUC_{synt} is computed as follows:

$$\begin{aligned} AUC_{synt} = {} & \frac{1}{2}\frac{\sum_{t\in T} AUC_{\text{Subgroup}}(\mathcal{M}_t) + \sum_{t\in T} AUC_{BPSN}(\mathcal{M}_t) + \sum_{t\in T} AUC_{BNSP}(\mathcal{M}_t)}{|T|} \\ & + \frac{1}{2}\frac{\sum_{i\in I} AUC_{\text{Subgroup}}(\mathcal{M}_i) + \sum_{i\in I} AUC_{BPSN}(\mathcal{M}_i) + \sum_{i\in I} AUC_{BNSP}(\mathcal{M}_i)}{|I|} \end{aligned} \tag{6}$$

$\mathcal{M}_t$ represents the subgroup of memes identified by the presence of a biased term t, T is the subset of selected biased terms. $\mathcal{M}_i$ denotes the subgroup of memes identified by the presence of a biased tag i and I denotes the subset of selected tags.

AUC_{synt} is a three per-element AUC-based measure, which considers both the biased terms and the biased tags, composed of the following estimations:

Table 6
Bias-related performance of multimodal models. **Bold** denotes the best performance for each performance measure, while (*) denotes that the corresponding model obtains results that are statistically different with respect to all the other models.

Model	AUC_{raw}	AUC_{synt}	*MBE*
MTT	**0.786**	0.657	0.721
MTC	0.757	**0.762**	**0.760***
Visual-BERT	0.773	0.617	0.695

- $AUC_{Subgroup}(\cdot)$, estimated on the subset of the synthetic dataset identified by the presence of a biased element;
- $AUC_{BPSN}(\cdot)$, computed on the background-positive subgroup-negative subset that corresponds to the subset of misogynous memes identified by the absence of the biased element and the not misogynous memes containing the biased element;
- $AUC_{BNSP}(\cdot)$, computed on the background-negative subgroup-positive subset that corresponds to the subset of not misogynous memes identified by the absence of the biased element and the misogynous memes containing the biased element.

The proposed *MBE* metric, which ranges into the interval $[0, 1]$, estimates the ability of the models on performing a good prediction on the raw test data and simultaneously achieving a significant performance on memes that by construction can lead to a biased prediction. According to Eq. (5), the higher is the MBE value, the better is the model. Analogously, according to Eq. (6), the lower is the AUC_{synt}, the higher is the bias of the model. If a model is biased, we expect to have a higher value related to the AUC_{raw} and a lower value on the AUC_{synt}.

We report in Table 6, the estimation of the *MBE* metric on the considered multimodal models, distinguishing between AUC_{raw} and AUC_{synt}. A T-test has been performed to compute the statistical equality with a pairwise analysis. The obtained results show that, according to the MBE measure, only MTC is statistically different from the other proposed models considering a value of alpha equal to 0.05. We can note that for both MTT and Visual-BERT, the AUC_{synt} is much lower than the AUC_{raw}, leading us to the conclusion that these models are strongly affected by the biased elements. For what concerns MTC, we can highlight that the model is able to achieve good performance on both AUC_{raw} and AUC_{synt}, suggesting that MTC is less biased than the other models.

In order to improve the generalization capabilities of the models affected by the presence of sample bias, we propose in the following section a corresponding mitigation strategy.

5.3. Bias mitigation

The proposed mitigation strategy has the main goal of reducing the distortion of the model mainly due to the presence of biased elements, by determining the optimal hyperparameter configuration such that the model could make unbiased predictions both on raw and synthetic memes. To this purpose, a Bayesian Optimization (BO) paradigm is adopted. The proposed approach is inspired by the fairness-aware hyperparameter optimization presented in F.Cruz, Saleiro, Belém, Soares, and Bizarro (2021), where the main goal is to find an adequate fairness-performance trade-off through a multiobjective optimization problem.

After a preliminary step based on an initial configuration of the hyperparameters, BO fits a (probabilistic) surrogate model. In our case, the probabilistic surrogate model is a Gaussian Process estimator based on the Matérn kernel, that aims at minimizing the negative MBE metric (i.e., maximizing the MBE metric) as objective function. Then a new promising hyperparameter configuration is identified depending on the acquisition function, which in our case is the Expected Improvement (EI). Unlike Maximum Probability of Improvement (MPI), EI considers the size of the improvement as well as the probability of improvement, resulting in a good trade-off between exploitation and exploration. Expected improvement is defined as:

$$EI(x) = \begin{cases} (f(x^+) - \mu(x) - \xi)\Phi(Z) + \sigma(x)\phi(Z) & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) = 0 \end{cases} \tag{7}$$

with

$$Z = \begin{cases} \frac{(f(x^+)-\mu(x)-\xi)}{\sigma(x)} & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) = 0 \end{cases} \tag{8}$$

where $\mu(x)$ and $\sigma(x)$ are the mean and the standard deviation of the Gaussian Process posterior predicted at x, respectively. Φ and ϕ are the Probability Density Function and Cumulative Distribution Function of the standard normal distribution, respectively. The parameter ξ governs the amount of exploration performed during optimization, with higher ξ values resulting in more exploration. In our case, $f(x^+)$ corresponds to the evaluation of MBE and it is the best value of the objective function observed so far.

Based on the new observation and the evaluation of the objective function, the surrogate model is updated. The process returns the current best solution (a.k.a. *best seen*). Afterwards, a new promising hyperparameter configuration is selected and the entire process is iteratively repeated. Typically, the termination criteria are budget-related.

Since the main aim is to obtain the optimal hyperparameter configurations without over-fitting the models, we performed the BO process according to the 10-fold cross-validation settings depicted in Fig. 10. The entire process starts by splitting both training (raw) and synthetic datasets in 10 folds. At each iteration, the following three main steps are performed:

1. **Bayesian Optimization**. The BO procedure is exploited to identify the best subset of hyperparameters:

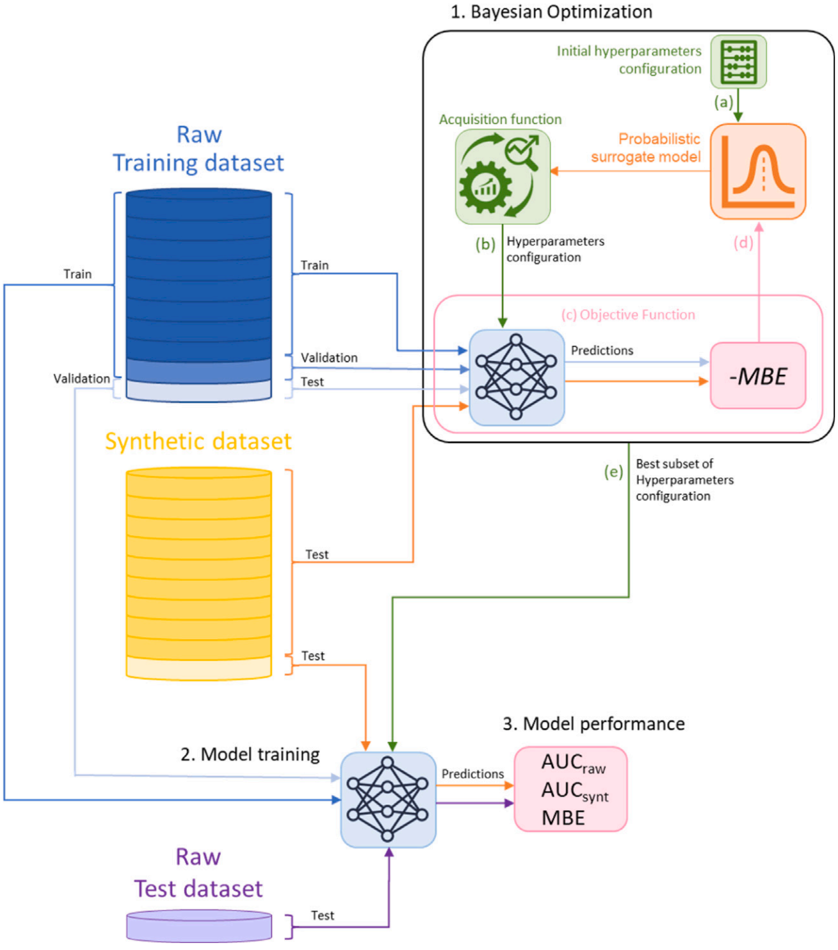


Fig. 10. Schematic representation of the Bayesian Optimization approach.

(a) An initial hyperparameter configuration (within the defined bounds) is defined and used to fit the probabilistic surrogate model.
(b) The acquisition function is called to identify the new promising hyperparameter configuration.
(c) The objective function is evaluated. The hyperparameter configuration is used to train a model on 8 folds of the training data, one of which is used as validation set and one as test set. In particular, given the 10,000 memes available as training set, we effectively used 8000 memes (8 folds) to train the model, 1000 memes (1 fold) as validation and 1000 memes (1 fold) to guide the hyperparameter optimization. More precisely, the MBE score, which is the metric to be maximized, is computed on the predictions made on the last fold from training data (used as test set) and on 9 folds of synthetic data.
The negative MBE is returned.
(d) The probabilistic surrogate model is updated.
(e) The termination criterion is checked: if it is not fulfilled, the whole procedure from step (b) is reiterated. Otherwise, the Bayesian Hyperparameter Optimizer returns the best subset of hyperparameters. As termination criterion, we considered a total number of iterations equal to 10 times the number of hyperparameters to optimize minus 1, which is the minimum number of iterations to obtain statistically relevant results.

2. **Model training**. The optimal hyperparameter configuration is used to train a brand new model that should achieve good prediction capabilities on the raw and the synthetic test sets. To this purpose, the entire training dataset (with 1 fold used as validation set) is used to perform the training phase.
3. **Model performance**. To evaluate the performance of the model based on the hyperparameter set determined by the BO strategy, the predictions are made on the raw test dataset and on the remaining fold of the synthetic dataset. Then, the model performance is computed.

Table 7
Comparison between original multimodal models and their mitigated version. **Bold** denotes the best result between the mitigated model and the corresponding non-mitigated one, while underlined text related to the MBE metric denotes the absolute best multimodal approach. (*) denotes that the corresponding model outperforms its non-mitigated version and the improvement is statistically significant.

Model	AUC_{raw}	AUC_{synt}	*MBE*
MTT	0.786	0.657	0.721
M-MTT	**0.808**	**0.719**	**0.763**
R-MTT (Thresholding)	0.250	0.453	0.352
R-MTT (Ranking)	0.607	0.401	0.504
R-MTT (Per-class Ranking)	0.609	0.465	0.537
R-MTT (Uniform)	0.785	0.629	0.707
R-MTT (Sample)	0.635	0.556	0.596
S-MTT (Narrow Topic)	0.729	**0.790**	0.759
MTC	0.757	0.762	0.760
M-MTC	**0.780**	**0.793**	<u>0.786</u>*
R-MTC (Thresholding)	0.316	0.325	0.321
R-MTC (Ranking)	0.662	0.560	0.611
R-MTC (Per-class Ranking)	0.662	0.598	0.630
R-MTC (Uniform)	0.754	0.727	0.740
R-MTC (Sample)	0.662	0.550	0.606
S-MTC (Narrow Topic)	0.739	**0.818**	0.778
Visual-BERT	0.773	0.617	0.695
M-Visual-BERT	0.773	0.617	0.695
R-Visual-BERT (Thresholding)	0.462	0.617	0.539
R-Visual-BERT (Ranking)	0.495	0.546	0.521
R-Visual-BERT (Per-class Ranking)	0.493	0.547	0.602
R-Visual-BERT (Uniform)	0.500	0.476	0.488
R-Visual-BERT (Sampling)	0.490	0.370	0.429
S-Visual-BERT (Narrow Topic)	0.467	**0.635**	0.551

For what concerns the optimal hyperparameters, the following search space has been considered:

- **Learning rate**: learning value associated with the Adam optimizer; the parameter space is defined between $1e^{-5}$ and 0.1.
- **Epsilon**: a small constant for numerical stability associated with the Adam optimizer; the parameter space is defined between $1e^{-8}$ and 1.
- **Neurons**: number of neurons in the hidden layer; the parameter space is defined between 64 and 528 (dimension of the input).
- **Activation function**: activation function for the hidden layer, defined according to the following subset: *sigmoid, relu, tanh, LeakyReLU.*
- **Dropout**: dropout value for the hidden layer: the parameter space is defined between 0.1 and 0.6.

In Appendix D, we reported the details about how the hyperparameter search spaces have been selected and a few considerations about the Bayesian Optimization process related to the best mitigated model.

Table 7 presents the results of the proposed bias mitigation strategy, reporting the performance in terms of AUC_{raw}, AUC_{synt} and MEB. We also report a comparison with two of the most widely used and recent mitigation strategies available in the literature that come from the natural language and image processing research areas, i.e. REPAIR (Li & Vasconcelos, 2019) and SAMPLING (Razo & Kübler, 2020). The considered mitigation baselines are characterized by several sampling strategies proposed in their corresponding original papers and here detailed in Appendix E. For the sake of completeness, we report the results achieved by the original multimodal models (MTT, MTC and Visual-BERT) and the proposed mitigated versions (M-MTT, M-MTC and M-Visual-BERT). We will denote each mitigated version originated by REPAIR as R-*model* and by SAMPLING as S-*model*. Since the Visual-BERT model has been previously optimized on a similar task (Kiela et al., 2020), the further optimization step on the tunable hyperparameters led to the same configuration and thus the performance was invariant.

A T-test has been performed to compute the statistical equality with a pairwise analysis between each model and its mitigated version. The obtained results show that, according to the MBE measure, only the MTC and M-MTC models are statistically different considering a value of alpha equal to 0.05. In the other cases, we cannot reject the null hypothesis. First of all, we can easily note that the proposed mitigation strategy provides a significant improvement in both proposed models (M-MTT and M-MTC), increasing not only the capability of the models in terms of AUC_{raw} on the test set but also on the AUC_{synt} on the synthetic one. This suggests that the mitigated models not only improve the recognition capabilities on memes that respect the protocol adopted for collecting and creating the dataset, but they also have a reduced bias and better generalization capabilities on controversial memes, thus reducing the impact related to the presence of biased elements on the model predictions.

A second remark relates to the general capabilities of the models: the most promising architecture available in the state of the art, i.e. Visual-BERT and its corresponding mitigated version M-Visual-BERT, achieves not only lower recognition capabilities on the raw test set than the other models but it is also characterized by the lowest generalization capabilities on the synthetic memes that have been created to expose the bias of the models. On the other hand, M-MTC shows the best generalization capabilities both on (raw) memes that are similar to the ones collected for training the model and on controversial (synthetic) memes.

Table 8
Comparison between original multimodal models (MTT and MTC) and their mitigated version (M-MTT and M-MTC) in terms of AUC_{synt} on the biased terms. **Bold** elements in the mitigated models indicate that an improvement is obtained with respect to the non-mitigated one.

Biased terms	MTT	M-MTT	MTC	M-MTC
Dishwasher	0.65	**0.70**	0.74	**0.78**
Chick	0.71	0.70	0.84	**0.85**
Whore	0.48	**0.50**	0.50	0.48
Demotivational	0.66	0.61	0.75	**0.76**
Diy	0.54	0.54	0.62	0.62
Memeshappen	0.69	**0.74**	0.70	0.67
Mcdonald	0.83	**0.84**	0.87	0.83
Ambulance	0.68	**0.78**	0.71	**0.84**
Developer	0.83	**0.84**	0.84	0.80
Template	0.66	**0.69**	0.67	**0.71**
Avg.	0.67	**0.69**	0.72	**0.73**

Table 9
Comparison between original multimodal models (MTT and MTC) and their mitigated version (M-MTT and M-MTC) in terms of AUC_{synt} on the biased tags. **Bold** elements in the mitigated models indicate that an improvement with respect to the non-mitigated one is obtained.

Biased tags	MTT	M-MTT	MTC	M-MTC
Animal	0.68	**0.69**	0.77	0.73
Car	0.85	**0.91**	0.87	**0.93**
Cat	0.57	**0.61**	0.63	0.55
Dog	0.66	**0.86**	0.80	**0.94**
Crockery	0.67	**0.79**	0.90	0.89
Dishwasher	0.51	**0.64**	0.51	**0.76**
Kitchen utensil	0.36	**0.61**	0.36	**0.61**
Woman	0.67	**0.61**	0.67	**0.70**
Avg.	0.62	**0.72**	0.69	**0.77**

A third remark comes from the comparison of the proposed mitigation strategy with REPAIR and SAMPLING. In particular, it has been shown that the proposed approach performs better, in most cases, than the other baseline methodologies and can be considered statistically significant when considering mitigation that works both with text and caption sources. The poor results of some mitigation strategies (i.e. REPAIR with Per-class Ranking and Thresholding) can be due to the use of the hyperparameter values (e.g. threshold t and percentage p) suggested by the original authors that cannot be necessarily optimal for mitigating misogynous meme detection models. In particular, selecting those instances such that the corresponding weight w_i is greater than $t = 0.5$ for the Thresholding method or that samples with the largest weight w_i such that the percentage of instances per class is equal to 50% for the Per-Class strategy could lead to underfitting the problem.

In order to quantify the impact of the mitigation strategy on the considered terms and tags, we estimated the AUC_{synt} metric for each of them comparing the original multimodal models and the corresponding mitigated version. Results are reported in Tables 8 and 9. Comparing the mitigated models (M-MTT and M-MTC) with the non-mitigated ones (MTT and MTC), we can highlight that the mitigation strategy has a positive effect of reducing the sample bias on both terms and tags (the higher the performance, the lower the bias). We can also underline that the proposed strategy has a major effect on the MTT model which was more affected by the sample bias. Regarding such improvements, which are on average between 1%–2% for the biased terms and between 8%–10% for the biased tags, the mitigation strategy has a strong impact on tags rather than on terms guaranteeing a relative improvement between 11%–15% (against a relative improvement on terms that is around 1%–3%).

By analyzing Table 8, a few terms (such as *demotivational* and *memeshappen*) that are strongly associated with the misogyny label do not result in significant improvements in most of the mitigated models. This behavior is mainly due to the location where such tokens appear in the meme text. In particular, the USE model is characterized by a training phase that is based on a self-attention mechanism. The self-attention process takes word order and surrounding context into account when generating each word representation and the entire representation of the sentence. Considering that such a few elements, as for example the domains where the memes have been created, appear at the end of the sentence and therefore do not belong to any frequent context pattern, their contribution to the embeddings of the meme text is less than other words that belong to a more frequent surrounding context. This implies that the bias mitigation strategy will have a major effect on those embedding representations where the presence of a given term is frequently contextualized with respect to the adjoining words, and a minor influence on those representations where the presence of a given term is basically independent of the surrounding context.

For what concerns the overall improvement in the models driven by the tag candidates, this is mainly due to the larger percentage of training samples containing candidate biased tags than terms. In particular, the percentage of training samples containing a given candidate biased tag is much larger than the percentage of training samples containing any candidate biased term. For instance, the training dataset used to guide the mitigation strategy contains 47.76% of memes with associated the candidate biased tag *woman*. The same dataset contains only 1.22% of memes containing the (most frequent) candidate biased term *dishwasher*. Since the training

Table 10
Percentage of misogynous ($\triangle^{+}$) and not misogynous ($\triangle^{-}$) memes in the synthetic dataset actually affected by the mitigation strategy.

Modality	Biased element	% of correct probability shift		
		$\triangle^{-}$	$\triangle^{+}$	Overall
Text	dishwasher	28.57%	100%	64.29%
	Chick	71.43%	71.43%	71.43%
	Whore	14.28%	57.14%	35.71%
	Demotivational	42.86%	100%	71.43%
	Diy	42.86%	57.14%	50%
	Memeshappen	57.14%	28.57%	42.86%
	Mcdonald	85.72%	57.14%	71.43%
	Ambulance	100%	100%	100%
	Developer	57.14%	28.57%	42.86%
	Template	85.72%	42.86%	64.29%
Tag	Dishwasher	100%	–	100%
	Woman	60.60%	–	60.60%
	Car	100%	–	100%
	Crockery	–	100%	100%
	Kitchen-utensil	–	100%	100%
	Animal	–	58.82%	58.82%
	Dog	–	100%	100%
	Cat	–	50%	50%
Overall		58.57%	64.29%	61.43%

and optimization steps reward the most frequent occurrences in the training set, it results in a model that has a major improvement on the most frequent elements provided during the training and optimization process.

We also evaluated, on the best model (M-MTC), the percentage of memes for which the mitigation strategy has produced a correct shift in terms of prediction probability with respect to the target class. In particular, given a biased element e, we estimated the percentage of probability shift towards the correct class. More formally, given a biased element e, the percentage of misogynous ($\triangle^{+}$) and not misogynous ($\triangle^{-}$) memes affected by the mitigation strategy is estimated as follows:

$$\triangle^{+}= \frac{\sum_{m_e^{+}}^{|\mathcal{M}_e^{+}|} \mathbb{1}_{+}(m_e^{+})}{|\mathcal{M}_e^{+}|} * 100 \qquad \triangle^{-}= \frac{\sum_{m_e^{-}}^{|\mathcal{M}_e^{-}|} \mathbb{1}_{-}(m_e^{-})}{|\mathcal{M}_e^{-}|} * 100 \tag{9}$$

where m_e^{+} and m_e^{-} denote respectively a misogynous (+) and not misogynous (-) synthetic meme containing the biased element e, $\mathcal{M}_e^{+}$ and $\mathcal{M}_e^{-}$ represent analogously the set of misogynous (+) and not misogynous (-) synthetic memes containing the biased element e, and finally $\mathbb{1}_{-}(m_e^{+})$ and $\mathbb{1}_{-}(m_e^{-})$ are indicator functions defined as follows:

$$\mathbb{1}_{+}(m_e^{+}) = \begin{cases} 1, & \text{if } P_{M-MTC}(m_e^{+}) > P_{MTC}(m_e^{+}) \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$\mathbb{1}_{-}(m_e^{-}) = \begin{cases} 1, & \text{if } P_{M-MTC}(m_e^{-}) > P_{MTC}(m_e^{-}) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $P_{M-MTC}(m_e^{+})$ is the probability given by the mitigated model M-MTC to predict the misogynous meme m_e^{+} as misogynous, and $P_{MTC}(m_e^{+})$ is the probability given by the non-mitigated model MTC to predict the misogynous meme m_e^{+} as misogynous.

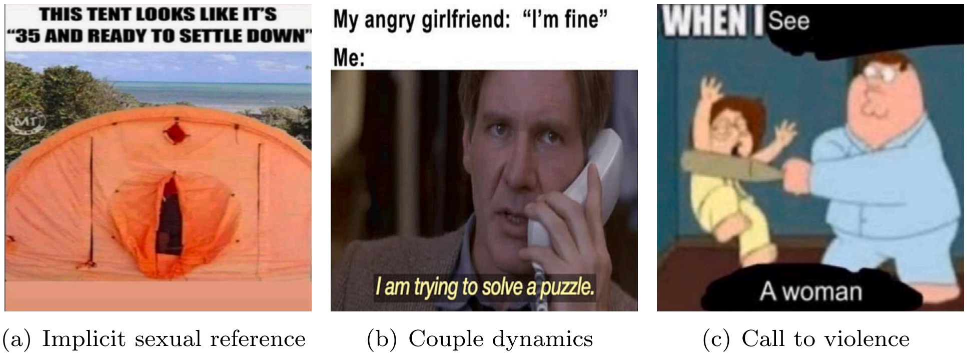
In Table 10, we report the percentage of not misogynous ($\triangle^{-}$) and misogynous ($\triangle^{+}$) memes affected by the mitigation strategy, where M-MTC and MTC are considered. The mitigated model M-MTC pushes the probabilities towards the correct class for 61.43% of the memes in the synthetic dataset.

Focusing on specific biased elements, the probability shift towards the correct class label appears to be related to the class associated with the biased elements: the prediction capabilities on misogynous memes improve more when a biased term is associated with the misogynous class (for instance *dishwasher*, *chick* and *whore*), while the prediction capabilities on not misogynous memes improve more when a biased term is associated to the not misogynous class (for instance *memeshappen*, *mcdonald* and *template*). An analogous behavior can be observed for the biased tags associated with a specific class. However, it is important to remark that although a biased element is more related to one specific class, the mitigation strategy not only introduced a correct shift towards that class but also on the other one, denoting an effective improvement on both labels.

In order to understand if the percentages representing the correctly shifted probabilities towards the correct decision are significant, we investigated different levels of significance by evaluating the relative shift in prediction introduced by the best mitigated model (M-MTC). In particular, we computed the Relative Shift (RS) as follows:

$$RS = \frac{|P_{M-MTC}(m_e) - P_{MTC}(m_e)|}{P_{MTC}(m_e)} \tag{12}$$

where $P_{M-MTC}(m_e)$ is the probability given by the mitigated model M-MTC to predict the correct class for the meme m_e and $P_{MTC}(m_e)$ is the probability given by the non-mitigated model MTC to predict the correct class for the meme m_e. We considered a probability


THIS TENT LOOKS LIKE IT'S "35 AND READY TO SETTLE DOWN"
My angry girlfriend: "I'm fine"
Me:
I am trying to solve a puzzle.
WHEN I See
A woman

(a) Implicit sexual reference (b) Couple dynamics (c) Call to violence

Fig. 11. Most frequent archetypes of misclassified misogynous memes.

shift as significant if RS is greater than a given threshold δ. Assuming δ=0.1, the percentage of significant probability shift towards the correct class is equal to 60.47%. Similarly, assuming thresholds of δ=0.2 and δ=0.3, the percentages of significant shifts are 45.35% and 41.86%, respectively.

Analogously, we estimated the Absolute Shift (AS) as follows:

$$AS = |P_{M-MTC}(m_e) - P_{MTC}(m_e)| \tag{13}$$

Also in this case, we considered a probability shift as significant if AS is greater than a given threshold ζ. Assuming ζ=0.1, the percentage of significant probability shift towards the correct class is equal to 41.86%. Similarly, assuming ζ=0.2 and ζ=0.3, the percentages of significant shifts are 19.77% and 8.14%, respectively. By considering both RS and AS, we can assert that the proposed mitigation strategy has a significant effect of the probability shift towards the correct class.

To summarize the main findings related to **RQ3**, we can affirm that the proposed mitigation strategy based on Bayesian optimization is particularly suitable to reduce the bias on controversial memes while still maintaining good identification performance on the rest. A larger synthetic dataset, built upon larger sets of biased elements, would improve even more the prediction capabilities of the models, making their mitigated version more suitable for a real-world application.

6. Meme archetypes: An open challenge

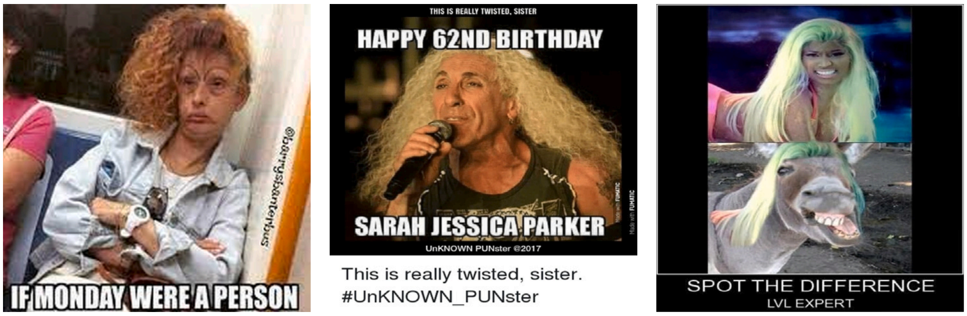
In order to provide a roadmap for future research directions, a systematic **error analysis** has been performed to understand if the memes share any common archetype that needs to be properly addressed from now onward. In particular, we analyzed the subset of misogynous memes erroneously classified as not misogynous by at least one of the multimodal mitigated approaches (M-MMT, M-MTC and M-Visual-BERT).

The most frequent archetypes in the misclassified misogynous meme set are the ones that convey a message about *implicit sexual reference* (44.35%), followed by *couple dynamics* (19.57%) and *call to violence* (17.83%). Examples of such archetypes are reported in Fig. 11.

Regarding memes denoting ***implicit sexual reference***, which in their explicit form are often censored, images of objects/things that recall feminine body parts are used. In this case, the memes are characterized by neutral text and images, which represent a misogynous message typically associated with objectification when combined. An example of memes representing an implicit sexual reference is reported in Fig. 11(a). The recognition of such archetypes is more than challenging due to the necessity of a compositional reasoning mechanism that should be able to combine two not misogynous different sources to come to a misogynous classification. *Multimodal compositional reasoning*, taking advantage of pre-trained models, is currently an open research issue (Thrush et al., 2022; Zerroug, Vaishnav, Colin, Musslick, & Serre, 2022).

Concerning memes representing ***couple dynamics***, images typically mock a stereotypical behavior of women that occurs in a relationship (e.g. husband-wife, boyfriend-girlfriend). An example of memes in this category refers to quarrels of women cheating, women being jealous or women tyranny over men's needs. An example of memes representing a stereotyped woman in a couple relationship is reported in Fig. 11(b). While measuring the presence of a *stereotypical bias* in pre-trained language, vision, and vision-language models has been addressed in the state of the art (Delobelle & Berendt, 2022; Joniak & Aizawa, 2022; Zhou, Lai, & Jiang, 2022), the identification of stereotypes into the original sources is still an open issue.

For what concerns the memes describing a ***call to violence*** against women, the most important source of information is related to the visual perspective. In fact, this archetype is represented by women with bruises and signs of physical violence and/or by men assaulting a woman, associated with a *neutral text*. An example of memes representing a call to violence against women is reported in Fig. 11(c). Explicit violence detection has been addressed for a long time, especially on video sources (Hu et al., 2022; Pang, He, Hu, & Li, 2021; Wang, Wang, & Fan, 2021). While violent scenes could be easily detected, *implicit violence* in terms of instigation to violent actions or results of abuse still needs to be properly tackled.



(a) Ugliness   (b) Popular people   (c) Comparison

Fig. 12. Frequent archetypes of misclassified misogynous memes.

Additional frequent archetypes related to misclassified misogynous memes convey a message about *ugliness* (16.52%), followed by *comparison* (6.52%) and reference to *popular people* (6.09%). Examples of such archetypes are reported in Fig. 12.

Regarding memes around the ***ugliness*** concept, women who do not adhere to the aesthetic standards promoted by society both in terms of their physical and aesthetic attractiveness are typically represented. Such types of memes contain *neutral text* coupled with an image of a woman that is far from society's standards of beauty. An example of memes representing the concept of ugliness is reported in Fig. 12(a). Although beauty is a debated concept widely investigated by philosophers, artists, and psychologists, the modeling of *aesthetic* people is an emerging visual recognition problem that still needs to be investigated and is still in its infancy (Dornaika, Moujahid, Wang, & Feng, 2020; Lebedeva, Guo, & Ying, 2022; Mao, Jin, & Du, 2009).

Regarding memes related to ***popular people*** (see Fig. 12(b)), where the hateful message is conveyed by depicting a specific person that is known for violent or abusive actions against women or by famous woman's name associated with an image of an animal or a caricature. Therefore, both image and text as well as their association may appear harmless at first glance, in some cases. The recognition of such type of archetypes could be possible only if in-depth *domain knowledge* would be available (Fortuna, Domínguez, Wanner, & Talat, 2022).

The last frequent archetype relates to what can be considered as ***comparison***, where memes typically make fun of women by taking advantage of a side-by-side comparison with objects, animals, or animated characters. An example of memes representing a stereotyped woman in a couple relationship is reported in Fig. 12(c). These archetypes are typically used to emphasize the physical or aesthetic characteristics of a woman when compared with other things. In this case, the text within the meme becomes superfluous and the comparison from a visual point of view is enough to delineate misogynous content. The recognition of such archetypes is challenging due to the necessity of addressing the detection of *metaphors* from a multimodal point of view. While metaphor detection has been widely addressed from a linguistic perspective, its understanding from a visual point of view is still in its infancy given the presence of few available datasets in the state of the art (Xu et al., 2022; Zhang, Zhang, Zhang, Yang, & Lin, 2021).

To answer **RQ4**, several archetypes that represent open challenges for misogyny identification systems have been identified. These archetypes highlight the necessity of addressing several main open research directions that can be summarized as multi-modal compositional reasoning, stereotype modeling, implicit violence detection, aesthetic people recognition, domain-knowledge integration, and metaphor identification.

7. Impact on theoretical and practical aspects and current limitations

The investigation about misogyny identification in memes has provided important results not only related to the computational perspective, but also from the understanding of the phenomenon point of view. In the following, we discuss a selection of the most relevant theoretical and practical implications, together with the most promising directions and limitations resulting from the proposed study:

- **Social effect.** Social media promotes the spreading of ideas and content, allowing users to share thoughts also in an anonymous form. As a negative consequence, haters are increasing. The impacts on women who experience being targeted by hateful content include the inability to respond to the abuse, loss of sleep, and decrease of self-confidence (Djuraskovic, 2023). Addressing this social effect is therefore mandatory. The proposed investigation has demonstrated that multimodal approaches, opportunely mitigated by any potential bias, can be developed to counteract the problem of making toxic online environments. What has been delivered with this study could have a strong impact in online social networks, where the expression of hate is a huge phenomenon constantly growing. Although on one hand, it is necessary to design effective methods to identify what can be considered misogynous, on the other hand, it is even important to guarantee, both from a theoretical and a practical point of view, to develop fair models able to not be biased by the presence of specific elements tout court. Therefore, as main limitation, the proposed approach does not consider semantically similar terms and thus applies a partial debiasing that is strictly related (from a lexicographic perspective) to the identified biased terms. This means that when detecting *whore* as a biased term, the proposed strategy is currently not able to exploit words with analogous meaning like *prostitute* or *escort*.

- **Generalization of multimodal bias estimation and mitigation.** The proposed study has demonstrated that multimodal approaches are necessary to address the problem of misogyny recognition in online environments, with a main necessity to generalize to multiple modalities that can be exploited to express a hateful message towards women. This pushes the theoretical investigation towards multimodal settings where all the available sources of information should be considered together when estimating and mitigating any potential bias. In particular, the main findings of the paper relate to the identification of potential candidate elements that could lead to unintended and unfair predictions. One of the main limitations of the proposed approach is its assumption of conditional independence between elements that compose the meme (words and objects), which may not hold in most real-world memes. This paves the way for further studies related to multiple sources of bias as well as on works focusing on inter-dependencies between them.
- **Modeling the singularities of misogyny.** Misogyny can be manifested in a variety of ways, which unfortunately share a few common structures and thus denote the singularity of the phenomenon with respect to other targets of hate. The evidence given by the proposed investigation, where several archetypes of misogynous memes have been identified, has highlighted the current lack of knowledge when modeling this specific phenomenon from a computational point of view. Currently, the proposed model is characterized by the shortcoming of do not integrating semantic information regarding these specific aspects related to misogyny. Therefore, designing models able to address the identified archetypes, represents an open challenge that can be addressed only by exploiting the peculiarity of the phenomenon itself.

8. Conclusions

In this paper we focused on the problem of automatic detection of misogynous content on memes belonging to an in-the-wild collection investigating different modalities, analyzing the bias that could affect the classification models, and finally proposing a mitigation strategy. We demonstrated that, despite the significant contribution of textual components in differentiating between misogynous and not misogynous memes, a multimodal approach is necessary to address the task. Additionally, we identified a set of relevant elements that are part of the memes, both from a textual and a visual point of view, that can lead models to produce biased predictions. The proposed strategy allows not only the identification of those elements but also the measurement of the distortion of the predictive model. Moreover, we propose a small synthetic dataset built on the basis of the identified candidate biased elements, that can be used to measure bias and easily extended with more examples. The implemented mitigation strategy, based on Bayesian Optimization, appears to be appropriate for reducing bias on controversial memes while keeping good identification performance. Finally, we identified archetypes that represent open challenges for misogyny identification systems, highlighting the corresponding research directions to be taken from now onward. These archetypes should be further investigated by considering a variety of vision and language models, in order to provide a wide range of model predictions.

CRediT authorship contribution statement

Giulia Rizzi: Conceptualization, Methodology, Software, Writing – original draft, Data curation, Investigation, Validation. **Francesca Gasparini:** Conceptualization, Methodology, Writing – original draft. **Aurora Saibene:** Conceptualization, Methodology, Writing – original draft. **Paolo Rosso:** Conceptualization, Writing – original draft. **Elisabetta Fersini:** Conceptualization, Methodology, Writing – original draft, Investigation, Validation.

Data availability

Data will be made available on request.

Acknowledgments

The work of Paolo Rosso was done in the framework of the FairTransNLP-Stereotypes research project on Fairness and Transparency for equitable NLP applications in social media: Identifying stereotypes and prejudices and developing equitable systems (PID2021-124361OB-C31) funded by MCIN/AEI / 10.13039/501100011033 and by ERDF, EU A way of making Europe.

The work of Elisabetta Fersini and Francesca Gasparini has been partially funded by the European Union – NextGenerationEU under the National Research Centre For HPC, Big Data and Quantum Computing - Spoke 9 - Digital Society and Smart Cities (PNRR-MUR), and by REGAINS – Excellence Department Research Project.

Appendix A. Additional details on the model comparison

In this section, results referring to label comparison are analyzed and represented via a box plot. This graphical representation highlights locality, spread, and skewness groups of numerical data through their quartiles. Moreover, the *whiskers* represent the variability of the upper and lower quartiles.

A.1. AUC comparison of unimodal and multimodal models

The results presented in Section 4.3 were further evaluated to highlight differences in model performance. The box plot in Fig. A.13 represents the AUC measures computed on training data using a 10-fold cross-validation approach (Table 2). The medians of each box plot compared to the others suggest that there is a difference between the analyzed model and the others for Text-based, Image tag-based, Image Caption-based, MTC, and Visual-BERT. The suggested difference has been confirmed by the statistical pairwise t-test as shown in Table 2. Analyzing the Interquartile Ranges (IQR) (that is, the box lengths), all the distributions appear to be almost equally dispersed.

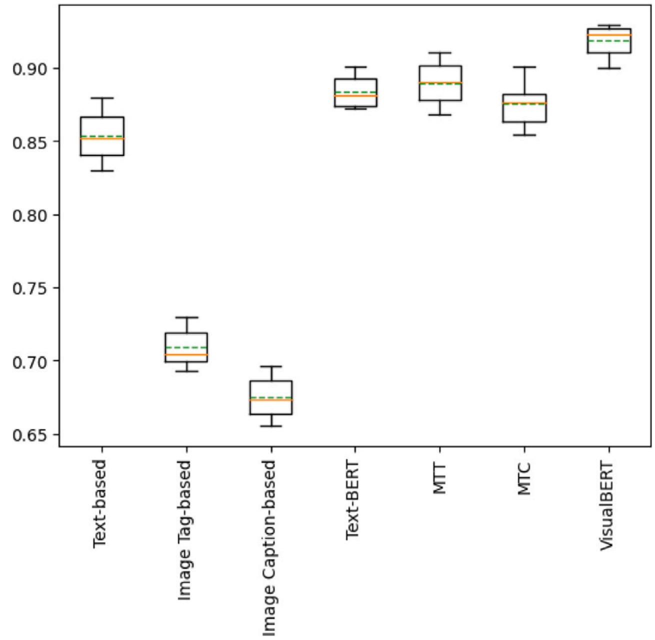


Fig. A.13. A box plot representing AUC measures computed on training data using 10-fold cross-validation.

3 A similar analysis has been performed for the AUC measures computed on the test data (Table 3) and reported in Fig. A.14. For Image tag-based, Image Caption-based, MTT, MTC, and Visual-BERT, the medians and the box plot suggest that there is a difference between the studied model and the others. The statistical paired t-test has validated the predicted difference, as displayed in Table

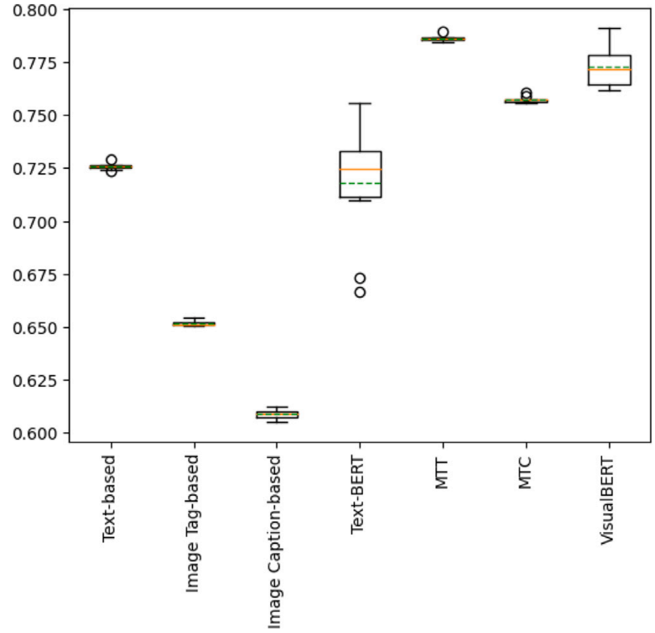


Fig. A.14. A box plot representing AUC measures computed on the test set.

The Interquartile Ranges (IQR) on the test data, differently from what has been observed in the training data, show a significantly higher dispersion for the transformer-based models (i.e., Text-BERT and Visual-BERT). Moreover, the same models report a wider distribution of the data with respect to the USE-based ones, as highlighted by the whiskers representing the score range.

Regarding the unimodal model, the achieved results indicate text as the most informative component. Furthermore, USE yields promising AUC with less dispersion (IQR and Range) on the test set. USE has thus been chosen as the embedding technique used in the multimodal models. Regarding the multimodal model, MTT seems to be the most promising, achieving good performances in terms of AUC measures while maintaining a small dispersion.

A.2. AUC comparison of original and mitigated models

Further analyses of the results reported in Table 7 have been performed in order to deepen the comparison between the original and mitigated models.

Fig. A.15 represents AUC measures reported in Table 7, computed on test data. The medians and the box plot suggest that there is likely to be a difference between all the proposed models. A statistical paired t-test has validated the predicted difference except for M-MTC and Visual-BERT. For all of the proposed models, the mitigated version appears to have more variability in terms of Interquartile Ranges and Ranges than the original one, still maintaining lower variability than Visual-BERT. Fig. A.16 represents

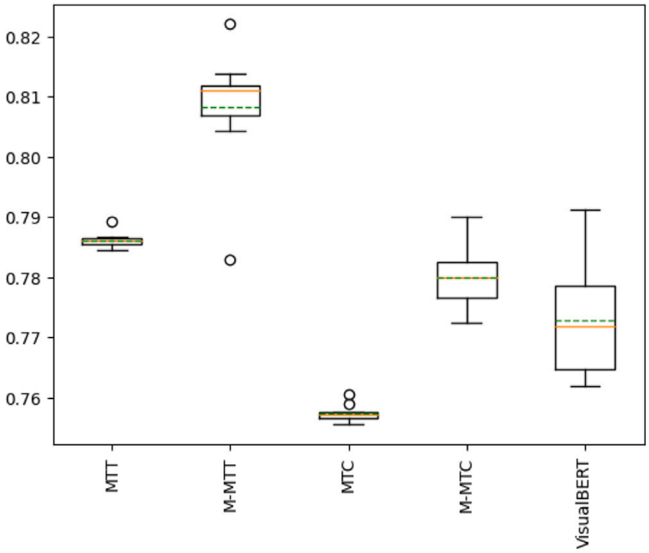


Fig. A.15. A box plot representing AUC measures computed on the test set.

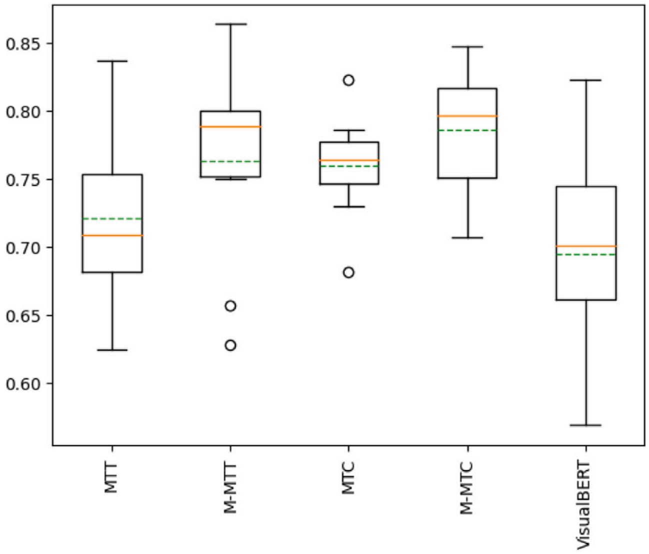


Fig. A.16. A box plot representing MBE measures.

the MBE scores reported in Table 7. The plot highlights a smaller variability of M-MTT with respect to MTT, and for MTC with respect to M-MTC. All the USE-based models achieved a smaller variability with respect to Visual-BERT. It can also be highlighted the presence of outliers both for M-MTT and MTC.

Appendix B. Tag mapping

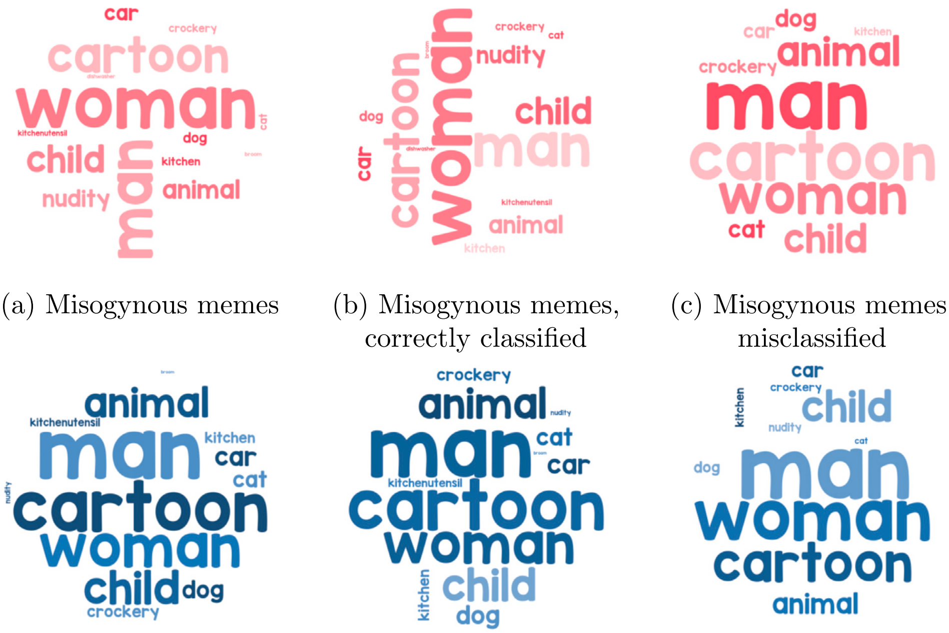
In this appendix, additional information about the process of tags mapping is reported. Tags provided by Clarifai through the usage of the general model have been mapped to the selected concepts (Table 1). In particular, each concept that appears also as a Clarifai tag has been included with a 1-to −1 match, while the other concepts have been reconducted to the most similar available tags, as shown in Table B.11.

Table B.11
Mapping table between the tags we defined and the ones proposed by Clarifai.

Tag	Clarifai mapping
Animal	Animal
Broom	Broom
Car	Car
Cartoon	Illustration
Cat	Cat
Child	Child
Crockery	Dishware, glass and flatware
Dishwasher	Dishwasher
Dog	Dog
KitchenUtensil	Kitchenware and cookware
Kitchen	Oven, stove, refrigerator and cabinet
Man	Man
Nudity	Nude, topless, nudist, bikini
Woman	Woman

Appendix C. Word clouds

An analysis of the frequencies analogous to the one shown in Section 4.3 has been performed also for tags and captions (Figs. C.17 and C.18, respectively).



(a) Misogynous memes

(b) Misogynous memes, correctly classified

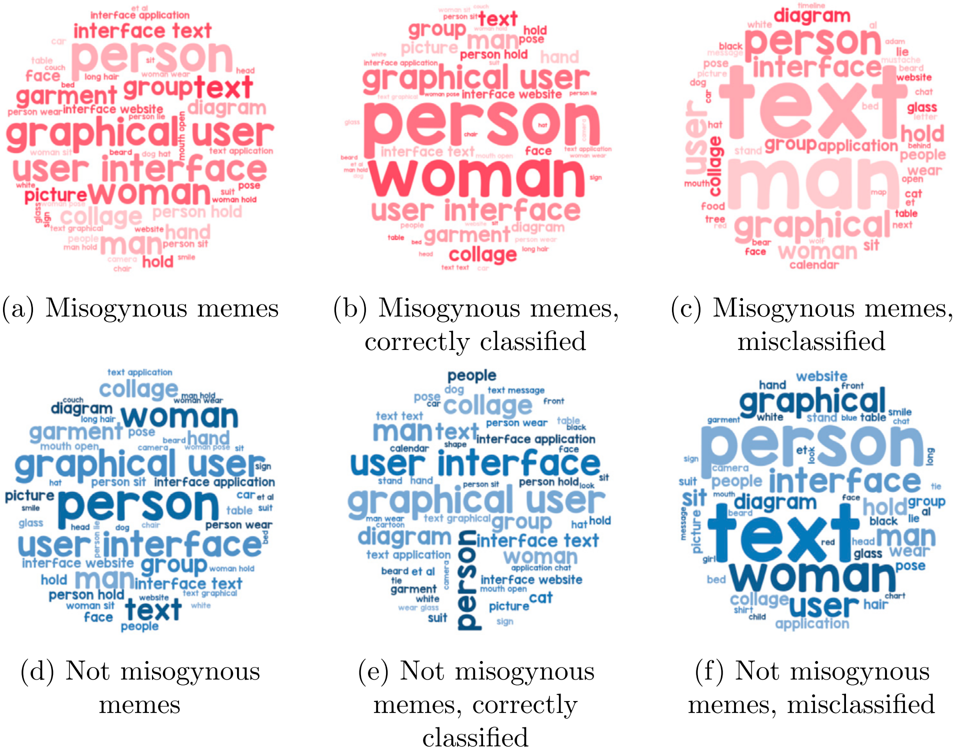
(c) Misogynous memes, misclassified

(d) Not misogynous memes

(e) Not misogynous memes, correctly classified

(f) Not misogynous memes, misclassified

Fig. C.17. Word Clouds to represent tags distribution.



(a) Misogynous memes

(b) Misogynous memes, correctly classified

(c) Misogynous memes, misclassified

(d) Not misogynous memes

(e) Not misogynous memes, correctly classified

(f) Not misogynous memes, misclassified

Fig. C.18. Word Clouds to represent terms distribution in caption.

Regarding the tag analysis (Fig. C.17), it emerges that the majority of the memes represent human or cartoon scenes. The tag *dishwasher* appears in the dataset only associated with the positive label and all the corresponding memes have been correctly classified as *misogynous*. Contrary to what has been observed for the text component, the word clouds associated with the tags seem to be more balanced. Finally, from the analysis performed on captions (Fig. C.18), it emerges that despite the terms woman and man being equally distributed in the dataset, referring to the misogynous label, the model seems to adopt an unintended behavior during the classification process. In fact, it emerges that memes whose caption contains the term *woman* are more likely to be associated with the positive label, while memes containing the term *man* are more likely to be associated with the not misogynous label.

Appendix D. Parameters exploration and selection

The hyperparameter space for Bayesian Optimization, described in Section 5.3, has been selected in order to have a suitable trade-off between time complexity and budget for the optimization steps, allowing a sufficient number of explorations of different hyperparameter configurations.

We report in Fig. D.19 a *2-Dimensional matrix of Partial Dependence* plots, which highlights the influence of each hyperparameter on the objective function. In particular, the reported plot refers to the selection of hyperparameters for the last iteration (10th Fold) for the M-MTC model. On the main diagonal, the relationships between each considered hyperparameter and the objective function are described, while the plots below the diagonal report how a hyperparameter relates to each other. The black dots represent the hyperparameter value evaluated during BO, while the red stars denote the optimal hyperparameter configuration. The plots on the main diagonal provide a few insights about the relationship between each individual feature and the target variable. The presence of flat or linear plots indicates a weak relationship between the predicted target variable (MBE) and the hyperparameters related to dropout, activation function and number of neurons. The plot showing non-linearity (i.e., the curve related to the pairs learning rate vs MBE) suggests that the impact of the hyperparameter on the target variable is not constant but changes as the hyperparameter configuration varies. The nonlinearity of these features is also reflected in the prevalence of lighter colors when indicating the relationship between them and the target variable. In our case, lr and epsilon result in a non-linear relationship with MBE.

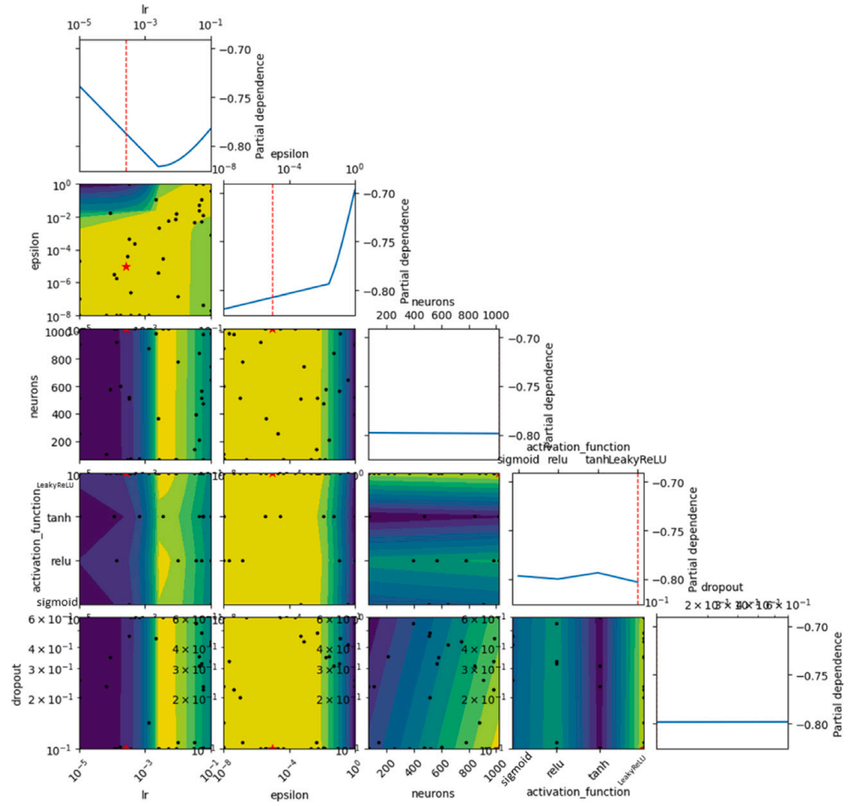


Fig. D.19. A 2-Dimensional matrix of Partial Dependence plot representing the selection of hyperparameters for the last iteration for the M-MTC model.

Appendix E. Debiasing methods

In this paper, we considered two main debiasing techniques as baselines to perform a comparison with the proposed approach, i.e., REPAIR (Li & Vasconcelos, 2019) and SAMPLING (Razo & Kübler, 2020).

REPresentAtion bIas Removal (REPAIR) (Li & Vasconcelos, 2019) is a debiasing method initially defined to deal with bias in image-based datasets. It computes a weight w_i for each sample based on its proportional loss contribution with respect to a reference model and resamples the original training dataset according to several strategies. In this paper, we considered REPAIR according to the sampling strategies proposed in the original paper (R- stands for REPAIR):

- R-model (Thresholding): Given a weight w_i for each meme i, it keeps only those instances such that $w_i >= t$, where $t = 0.5$ is the threshold;
- R-model (Ranking): Given a weight w_i for each meme i, it keeps p = 50% samples of largest w_i;
- R-model (Per-class Ranking): Given a weight w_i for each meme i, it keeps $p = 50\%$ examples with the largest weight w_i from each class;
- R-model (Sample): Given a weight w_i for each meme i, it disregards those memes with probability $1 - w_i$.
- R-model (Uniform): it keeps $p = 50\%$ examples uniformly at random.

SAMPLING (Razo & Kübler, 2020) is a debiasing method initially defined to deal with bias in (abusive) language. It removes a set of training instances to reduce the bias effect on a predictive model. In particular, given a set of narrow topics that in our case are denoted by the candidate biased terms and tags, any meme containing at least one of such candidates is removed from the training set. This strategy in the original paper is known as *Narrow Topic*.

References

AlDahoul, N., Abdul Karim, H., Lye Abdullah, M. H., Ahmad Fauzi, M. F., Ba Wazir, A. S., Mansor, S., et al. (2021). Transfer detection of YOLO to focus CNN's attention on nude regions for adult content detection. *Symmetry*, *13*(1).

Almenar, R. (2021). Cyberviolence against women and girls: Gender-based violence in the digital age and future challenges as a consequence of Covid-19. *Trento Student Law Review*, *3*(1), 167–230.

Andreasen, M. B. (2021). 'Rapeable'and 'unrapeable'women: the portrayal of sexual violence in Internet memes about #MeToo. *Journal of Gender Studies*, *30*(1), 102–113.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *International conference on applications of natural language to information systems* (pp. 57–64). Springer.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on world wide web companion* (pp. 759–760).

Bashar, M. A., Nayak, R., & Suzor, N. (2020). Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set. *Knowledge and Information Systems*, *62*, 4029–4054.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems*.

Butt, S., Ashraf, N., Sidorov, G., & Gelbukh, A. F. (2021). Sexism identification using BERT and data augmentation - EXIST2021. In *Iberian languages evaluation forum* (pp. 381–389).

Calderón-Suarez, R., Ortega-Mendoza, R. M., Montes-Y-Gómez, M., Toxqui-Quitl, C., & Márquez-Vera, M. A. (2023). Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases. *IEEE Access*, *11*, 13179–13190.

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., & Banaji, M. R. (2022). Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (pp. 156–170).

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., et al. (2018). Universal sentence encoder for english. In *Empirical methods in natural language processing (EMNLP): System demonstrations* (pp. 169–174).

Chaloner, K., & Maldonado, A. (2019). Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the first workshop on gender bias in natural language processing* (pp. 25–32).

Clarifai (2023). Clarifai guide. URL https://docs.clarifai.com/. (Accessed 01 February 2023).

Collett, C., Gomes, L. G., Neff, G., et al. (2022). *The effects of AI on the working lives of women*. UNESCO Publishing.

Delobelle, P., & Berendt, B. (2022). FairDistillation: Mitigating stereotyping in language models. In *European conference on machine learning and principles and practice of knowledge discovery in databases*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *17th Annual conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186).

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *AAAI/ACM conference on AI, ethics, and society* (pp. 67–73).

Djuraskovic, O. (2023). Cyberbullying statistics, facts, and trends (2023) with charts. URL https://firstsiteguide.com/cyberbullying-stats/. (Accessed 20 February 2023).

Dornaika, F., Moujahid, A., Wang, K., & Feng, X. (2020). Efficient deep discriminant embedding: qpplication to face beauty prediction and classification. *Engineering Applications of Artificial Intelligence*, *95*, Article 103831.

Dutta, S., Majumder, U., & Naskar, S. K. (2021). An efficient BERT based approach to detect aggression and misogyny. In *Proceedings of the 18th international conference on natural language processing* (pp. 493–498).

Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin*, *15*(4), 543–558.

Elsafoury, F., Wilson, S. R., Katsigiannis, S., & Ramzan, N. (2022). SOS: Systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th international conference on computational linguistics* (pp. 1263–1274). Gyeongju, Republic of Korea: International Committee on Computational Linguistics, URL https://aclanthology.org/2022.coling-1.108.

F.Cruz, A., Saleiro, P., Belém, C., Soares, C., & Bizarro, P. (2021). Promoting fairness through hyperparameter optimization. In *2021 IEEE international conference on data mining* (pp. 1036–1041). http://dx.doi.org/10.1109/ICDM51629.2021.00119.

Fersini, E., Gasparini, F., & Corchs, S. (2019). Detecting sexist MEME on the web: A study on textual and visual cues. In *8th International conference on affective computing and intelligent interaction workshops and demos* (pp. 226–231).

Fersini, E., Gasparini, F., Rizzi, G., Saibene, A., Chulvi, B., Rosso, P., et al. (2022). SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *16th International workshop on semantic evaluation*. Association for Computational Linguistics.

Fersini, E., Nozza, D., Rosso, P., et al. (2020). AMI@EVALITA2020: Automatic misogyny identification. In *7th Evaluation campaign of natural language processing and speech tools for Italian*.

Fersini, E., Rizzi, G., Saibene, A., & Gasparini, F. (2021). Misogynous MEME recognition: A preliminary study. In *International conference of the Italian association for artificial intelligence*. Springer.

Field, A., & Tsvetkov, Y. (2020). Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 596–608).

Fortuna, P., Domínguez, M., Wanner, L., & Talat, Z. (2022). Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 11794–11805).

Frenda, S., Ghanem, B., Guzmán-Falcón, E., Montes, M., Villaseñor-Pineda, L., & Villaseñor-Pineda, V. (2018). Automatic expansion of lexicons for multilingual misogyny detection.

Gandhi, S., Kokkula, S., Chaudhuri, A., Magnani, A., Stanley, T., Ahmadi, B., et al. (2020). Scalable detection of offensive and non-compliant content/logo in product images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2247–2256).

Gangwar, A., González-Castro, V., Alegre, E., & Fidalgo, E. (2021). AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images. *Neurocomputing*, *445*, 81–104.

García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., & Valencia-García, R. (2021). Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 506–518.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gasparini, F., Erba, I., Fersini, E., & Corchs, S. (2018). Multimodal classification of sexist advertisements. In *ICETE no. 1* (pp. 565–572).

Hee, M. S., Lee, R. K.-W., & Chong, W.-H. (2022). On explaining multimodal hateful meme detection models. In *ACM web conference* (pp. 3651–3655).

Hibberts, M., Burke Johnson, R., & Hudson, K. (2012). Common survey sampling techniques. *Handbook of Survey Methodology for the Social Sciences*, 53–74.

Hirota, Y., Nakashima, Y., & Garcia, N. (2022). Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13450–13459).

Hor, S. L., Karim, H. A., Abdullah, M. H. L., AlDahoul, N., Mansor, S., Fauzi, M. F. A., et al. (2021). An evaluation of state-of-the-art object detectors for pornography detection. In *IEEE international conference on signal and image processing applications* (pp. 191–196).

Hu, X., Fan, Z., Jiang, L., Xu, J., Li, G., Chen, W., et al. (2022). TOP-ALCM: A novel video analysis method for violence detection in crowded scenes. *Information Sciences*.

Hu, X., Yin, X., Lin, K., Zhang, L., Gao, J., Wang, L., et al. (2021). Vivo: Visual vocabulary pre-training for novel object captioning. In *AAAI conference on artificial intelligence, vol. 35, no. 2* (pp. 1575–1583).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Jatmiko, M. I., Syukron, M., & Mekarsari, Y. (2020). Covid-19, harassment and social media: A study of gender-based violence facilitated by technology during the pandemic. *The Journal of Society and Media*, *4*(2), 319–347.

Jiang, A., Yang, X., Liu, Y., & Zubiaga, A. (2022). SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, Article 100182.

Joniak, P., & Aizawa, A. (2022). Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. In *4th Workshop on gender bias in natural language processing* (pp. 67–73).

Kaneko, M., Imankulova, A., Bollegala, D., & Okazaki, N. (2022). Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 2740–2750).

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., et al. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in neural information processing systems* (pp. 2611–2624).

Kingma, D. P., & Ba, J. (2015). ADAM: A method for stochastic optimization. In *3rd International conference on learning representations*.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, *58*(5), Article 102643.

Kyriakou, K., Barlas, P., Kleanthous, S., & Otterbacher, J. (2019). Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the international AAAI conference on web and social media, vol. 13* (pp. 313–322).

Lebedeva, I., Guo, Y., & Ying, F. (2022). MEBeauty: a multi-ethnic facial beauty dataset in-the-wild. *Neural Computing and Applications*, *34*(17), 14169–14183.

Li, Y., & Vasconcelos, N. (2019). Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9572–9581).

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2020). What does BERT with vision look at? In *58th Annual meeting of the association for computational linguistics* (pp. 5265–5275).

Lin, X., Qin, F., Peng, Y., & Shao, Y. (2021). Fine-grained pornographic image recognition with multiple feature fusion transfer learning. *International Journal of Machine Learning and Cybernetics*, *12*(1), 73–86.

Ling, C., AbuHilal, I., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Dissecting the meme magic: Understanding indicators of virality in image memes. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–24.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in neural information processing systems*.

Manzini, T., Chong, L. Y., Black, A. W., & Tsvetkov, Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (Long and short papers)* (pp. 615–621).

Mao, H., Jin, L., & Du, M. (2009). Automatic classification of Chinese female facial beauty using support vector machine. In *IEEE international conference on systems, man and cybernetics* (pp. 4842–4846).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

Musso, M. G., Proietti, M., & Reynolds, R. R. (2020). Towards an integrated approach to violence against women: persistence, specificity and complexity. *International Review of Sociology*, *30*(2), 249–278.

Nevala, S. (2014). Violence against women: an EU-wide survey.

Ousidhoum, N., Song, Y., & Yeung, D.-Y. (2020). Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In *Empirical methods in natural language processing*.

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny detection in Twitter: a multilingual and cross-domain study. *Information Processing & Management*, *57*(6), Article 102360.

Pamungkas, E., Basile, V., & Patti, V. (2022). Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*.

Pang, W.-F., He, Q.-H., Hu, Y.-j., & Li, Y.-X. (2021). Violence detection in videos based on fusing visual and audio information. In *International conference on acoustics, speech and signal processing* (pp. 2260–2264).

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. In *Empirical methods in natural language processing* (pp. 2799–2804).

Poland, B. (2016). *Haters: Harassment, abuse, and violence online*. U of Nebraska Press.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*(2), 477–523.

Razo, D., & Kübler, S. (2020). Investigating sampling bias in abusive language detection. In *Proceedings of the fourth workshop on online abuse and harms* (pp. 70–78). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.alw-1.9, Online, URL https://aclanthology.org/2020.alw-1.9.

Rights, H. (2014). Violence against women: an eu-wide survey. Luxembourg: Publications Office of the European Union.

Rodríguez, D. A., Díaz-Ramírez, A., Miranda-Vega, J. E., Trujillo, L., & Mejía-Alvarez, P. (2021). A systematic review of computer science solutions for addressing violence against women and children. *IEEE Access*.

Samghabadi, N. S., Patwa, P., Pykl, S., Mukherjee, P., Das, A., & Solorio, T. (2020). Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the second workshop on trolling, aggression and cyberbullying* (pp. 126–131).

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing gender bias in image recognition systems. *Socius*, *6*, Article 2378023120967171.

Shen, T., Li, J., Bouadjenek, M. R., Mai, Z., & Sanner, S. (2023). Towards understanding and mitigating unintended biases in language model-driven conversational recommendation. *Information Processing & Management*, *60*(1), Article 103139.

Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., et al. (2020). MMF: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.

Song, R., Giunchiglia, F., Li, Y., Shi, L., & Xu, H. (2023). Measuring and mitigating language model biases in abusive language detection. *Information Processing & Management*, *60*(3), Article 103277.

Spinde, T., Rudnitckaia, L., Mitrović, J., Hamborg, F., Granitzer, M., Gipp, B., et al. (2021). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, *58*(3), Article 102505.

Szymanski, D. M., Moffitt, L. B., & Carr, E. R. (2011). Sexual objectification of women: advances to theory and research. *The Counseling Psychologist*, *39*(1), 6–38.

Ta, H. T., Rahman, A. B. S., Najjar, L., & Gelbukh, A. (2022). Transfer learning from multilingual DeBERTa for sexism identification. In *CEUR workshop proceedings, vol. 3202*. CEUR-WS.

Tabone, A., Camilleri, K., Bonnici, A., Cristina, S., Farrugia, R., & Borg, M. (2021). Pornographic content classification using deep-learning. In *21st ACM symposium on document engineering* (pp. 1–10).

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., et al. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF conference on computer vision and pattern recognition* (pp. 5238–5248).

Van Der Wilk, A., Natter, M., et al. (2018). *Cyber violence and hate speech online against women*. EPRS: European Parliamentary Research Service.

Van Royen, K., Poels, K., Vandebosch, H., & Walrave, M. (2018). Slut-Shaming 2.0. In *Sexting* (pp. 81–98).

Wang, P., Wang, P., & Fan, E. (2021). Violence detection and face recognition based on deep learning. *Pattern Recognition Letters*, *142*, 20–24.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. In *North American chapter of the association for computational linguistics: Human language technologies* (pp. 602–608).

Xu, B., Li, T., Zheng, J., Naseriparsa, M., Zhao, Z., Lin, H., et al. (2022). MET-Meme: A multimodal meme dataset rich in metaphors. In *45th International ACM SIGIR conference on research and development in information retrieval* (pp. 2887–2899).

Yalcin, E., & Bilge, A. (2022). Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Information Processing & Management*, *59*(6), Article 103100.

YPulse (2019). Report: Social media behavior. URL https://www.ypulse.com/report/2019/02/20/topline-social-media-behavior2/.

Zerroug, A., Vaishnav, M., Colin, J., Musslick, S., & Serre, T. (2022). A benchmark for compositional visual reasoning. In *Thirty-sixth conference on neural information processing systems datasets and benchmarks track*.

Zhang, D., Zhang, M., Zhang, H., Yang, L., & Lin, H. (2021). Multimet: A multimodal dataset for metaphor understanding. In *59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 3214–3225).

Zhao, D., Wang, A., & Russakovsky, O. (2021). Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14830–14840).

Zhou, K., Lai, E., & Jiang, J. (2022). VLStereoSet: A study of stereotypical bias in pre-trained vision-language models. In *2nd Conference of the Asia-Pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing* (pp. 527–538).