Misogynous Memes Recognition: Training vs Inference Bias Mitigation Strategies

Gianmaria Balducci*
Università di Milano-Bicocca
PMI Reboot S.r.l.

Giulia Rizzi**
Università di Milano-Bicocca
Universitat Politècnica de València

Elisabetta Fersini†
Università di Milano-Bicocca

Warning: This paper contains examples of language and images which may be offensive. *In this paper, we address the problem of automatic misogynous meme recognition by dealing with potentially biased elements that could lead to unfair models. In particular, a bias estimation technique is used to identify those textual and visual elements that unintendedly affect the model prediction, and a few bias mitigation methods are proposed, investigating two different types of debiasing strategies, i.e., at training time and at inference time. The proposed approaches achieve remarkable results both in terms of prediction and generalization capabilities.*

1. Introduction

Memes have become an increasingly popular form of communication, representing humor, culture, and shared experiences in a compact, easily shareable format. Their viral nature makes them an effective medium for spreading messages online, and they are especially popular on social media. According to the Digital 2022 Global Overview Report (We Are Social and Hootsuite 2022), more than 37% of internet users between 16 and 64 years old watch memes and viral videos. However, these memes can also perpetuate discriminatory behaviors towards certain groups and minorities. Memes are, in fact, sometimes being used as a vehicle for hate speech, which is defined as any form of derogatory communication that targets individuals or groups based on characteristics such as race, religion, ethnicity, sexual orientation, disability, or gender. The task of identifying hate in memes is challenging due to the need to understand textual content and visual cues, as well as the subtle interplay between the two. Additionally, in many cases, cultural context is required since the conveyed message might refer to specific persons or events. One form of hate frequently depicted in memes is misogyny. As social networks have become increasingly prevalent, new modes of communication and social interaction have risen, enabling hateful messages towards women to be expressed (Fontanella et al. 2024). In particular, the multimodal nature of memes allows hatred towards women to be represented by leveraging different aspects such as female stereotyping, shaming, objectification, and violence. This phenomenon is particularly

* University of Milano-Bicocca, Milan, Italy; PMI Reboot S.r.l., Milan, Italy.
E-mail: g.balducci1@campus.unimib.it
** University of Milano-Bicocca, Milan, Italy; Universitat Politècnica de València, Valencia, Spain .
E-mail: g.rizzi10@campus.unimib.it
† University of Milano-Bicocca, Milan, Italy. E-mail: elisabetta.fersini@unimib.it

serious considering that more than 46% of Social Media users are female (We Are Social and Hootsuite 2022). While misogyny recognition mechanisms have been widely investigated focusing on textual sources (i.e., tweets) (Anzovino, Fersini, and Rosso 2018; Bashar, Nayak, and Suzor 2020; Ta et al. 2022; Calderón-Suarez et al. 2023), less attention has been paid to misogynous identification in multimodal settings, and in particular on memes. A preliminary insight is represented by (Fersini, Gasparini, and Corchs 2019) in which simple unimodal and multimodal approaches are compared to investigate the contribution of the two modalities that compose a meme (i.e., textual and visual). Further investigations from the same authors (Fersini et al. 2021) presented a multimodal approach that considers the visual component through image captioning and the textual one, with the textual transcription, to distinguish between misogynous and non-misogynous memes. Recently, pre-trained and trained-from-scratch models were compared to see if domain-specific pre-training could improve recognition performance (Singh, Haridasan, and Mooney 2023), highlighting the importance of domain-specific pretraining in identifying multimodal misogyny. Several authors (Song et al. 2023; Shen et al. 2023) have pointed out that, despite the high results obtained on the recognition task, there may be some potential *bias* affecting the models. The majority of works aim at quantifying and minimizing the bias at the dataset or model level by focusing on a fixed set of seed words to propose bias estimation metrics and related mitigation strategies. However, in the field of misogynous meme recognition, consolidated metrics to estimate the bias and techniques to mitigate it are still missing. To this purpose, in this work, we provide the following main contributions related to mitigation strategies at inference time:

- *Bayesian Averaging Model - Class-based Mitigation (BMA-CM)*, which mitigates the predictions of each model considering the respective bias towards the class of the elements that appear in the meme;
- *Bayesian Averaging Model - Element-based Mitigation (BMA-EM)*, which mitigates the predictions of each model considering the respective bias towards the candidate biased elements that appear in the meme;

The rest of the paper is organized as follows. In Section 2 a summary of the state of the art is reported. The adopted dataset is presented in Section 3. The Bias Estimation strategy is introduced in Section 4. The bias estimation phase is articulated in the identification of candidate bias elements (Section 4.1), the creation of a synthetic dataset (Sections 4.2), and the description of a multimodal Bias Estimation metric (Section 4.3). In Section 5, the proposed debiasing strategies are defined. In Section 6, the experimental results are discussed. In Section 7, conclusions are reported.

2. Background

The ease with which information may be shared on social networks, as well as their new modes of communication and social interaction, have fostered the spread of numerous sorts of material in which users express their beliefs, ideologies, and opinions. As a consequence, even more deeply rooted ideologies and beliefs with historical origins, such as various types of hatred, for example, towards women, have evolved new modes of representation (Fontanella et al. 2024). Likewise, research in the area of hate identification has adjusted to the phenomenon: the detection of hate speech directed at various minorities and subgroups was first restricted to Tweets (Anzovino, Fersini,

and Rosso 2018; Bashar, Nayak, and Suzor 2020; Ta et al. 2022; Calderón-Suarez et al. 2023); only in recent years has it expanded to encompass multimodal content, including memes. One of the areas that are gaining popularity refers to the identification of hateful content towards women, which is articulated into the identification of misogynistic or sexist content, also considering the different ways in which this type of hate can be expressed (e.g., by means of stereotypes, objectification, etc.). It is crucial to address this issue, given that almost half of the global population, and more than 46% of social media users, are female (We Are Social and Hootsuite 2022). However, detecting misogyny is a complex area of research due to the various ways in which hate towards women can be expressed. This includes targeting women for different reasons. A first insight to counter sexist memes has been proposed in (Fersini, Gasparini, and Corchs 2019). This approach aims to address the issue of memes that can convey sexist messages by investigating both unimodal and multimodal approaches. The study examines the contribution of textual and visual cues in order to understand the various ways that hate towards women is expressed, ranging from stereotyping women to shaming, objectification, and violence. Simultaneously, other approaches have focused on evaluating the information content introduced by the two modalities that make up memes (i.e., Visual and textual components). An example is represented by (Sabat, Ferrer, and i Nieto 2019), in which the authors identify the visual component as more informative for detecting hate speech in memes. More recently, two benchmark datasets have been proposed to facilitate the investigation related to misogynous meme detection. The first benchmark presented in (Gasparini et al. 2022) contains 800 memes from the most popular social media platforms. All the memes that compose the dataset have been labeled by three experts and by three annotators from a crowdsourcing platform, involving a total of 60 annotators. More recently, a similar benchmark has been collected for the *MAMI* shared task at SemEval 2022 (Fersini et al. 2022). It contains 10.000 memes for training and additional 1.000 memes for testing, allowing both to (i) identify misogynistic memes and (ii) recognise the misogynistic type among potentially overlapping types (i.e., Shaming, Stereotype, Objectification, and Violence). The majority of the participants (Zhou et al. 2022; Chen and Chou 2022; Hakimov, Cheema, and Ewerth 2022; Zhi et al. 2022) presented pre-trained models-based approaches and/or investigated ensemble strategies. However, as highlighted by the challenge organizers, due to the presence of specific terms or elements within the images, most of the systems tend to be biased towards the misogyny category. While a lot of researchers have investigated the potential bias that the models could inherit from the training dataset, from an unimodal perspective (Nozza, Volpetti, and Fersini 2019; Zueva, Kabirova, and Kalaidin 2020; Nascimento, Cavalcanti, and Da Costa-Abreu 2022; Fersini, Candelieri, and Pastore 2023), less attention has been paid to investigate the bias in a multimodal settings. Particular attention has been paid to studying the distributions of the terms in the datasets in order to identify specific terms, called *identity terms*, frequently related to hateful expressions referring to a specific target. The proposed works demonstrated how the models evolved undesired behaviours based on biased implicit associations between such terms and the provided class label, resulting in unfair predictions. The state-of-the-art also proposes several mitigation strategies to counteract this phenomenon. Among the proposed solutions, the most widely adopted strategy is related to data augmentation (Calderón-Suarez et al. 2023; Zmigrod et al. 2019; Guo et al. 2023). This strategy consists of the inclusion of additional instances with specific characteristics, aiming to adjust the unbalanced distribution that characterized the identity terms. In hate-related domains, typically, additional non-toxic comments that report the selected identity terms are collected. While this technique aims at mitigating bias directly from

the datasets, more complex alternatives directly mitigate the models through the definition of specific objective functions (Xia, Field, and Tsvetkov 2020; Sridhar and Yang 2022) or optimization strategies (Perrone et al. 2021; Sikdar, Lemmerich, and Strohmaier 2022). Although the above-mentioned strategies represent a fundamental step towards bias mitigation, they are defined for unimodal settings. Less attention has been paid to investigating bias estimation and mitigation for a multimodal perspective, especially for misogynous meme identification. A first insight in analyzing the bias that could affect the classification models for misogynous meme recognition and defining a mitigation strategy is represented by (Rizzi et al. 2023). The authors propose a strategy to identify 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 and a metric to measure the distortion of the predictive model. Moreover, they propose a mitigation strategy based on Bayesian Optimization. A more recent approach is represented by (Balducci, Rizzi, and Fersini 2023) in which the authors propose a mitigation strategy at training time, named Masking Mitigation, that masks the candidate biased elements to reduce the distortion introduced by their presence. This work represents an extension of (Balducci, Rizzi, and Fersini 2023), which proposes a few sophisticated debiasing techniques. In particular, while the former approaches have been defined to mitigate the bias at training time, making the learning process more complex from a computational point of view, the proposed approaches work at inference time with the advantage of keeping the training phase simple and reducing the bias when processing unseen memes.

3. Dataset

The proposed method has been evaluated on the Multimedia Automatic Misogyny Identification (MAMI) Dataset (Fersini et al. 2022), consisting of 10.000 memes for training and 1.000 memes for testing. As shown in Figure 1, the dataset contains memes representing different types of misogyny, including:

- *Shaming*: The practice of criticising women who violate expectations of behaviour and appearance regarding issues related to gender typology (such as "slut shaming") or related to physical appearance (such as "body shaming") (Van Royen et al. 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 and 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 behaviours (gender stereotyping).
- *Objectification*: A practice of seeing and/or treating a woman like an object (Szymanski, Moffitt, and Carr 2011).
- *Violence*: A meme that indicates physical and/or a call to violence against women (Andreasen 2021).

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

Figure 1
Examples of misogynous memes.

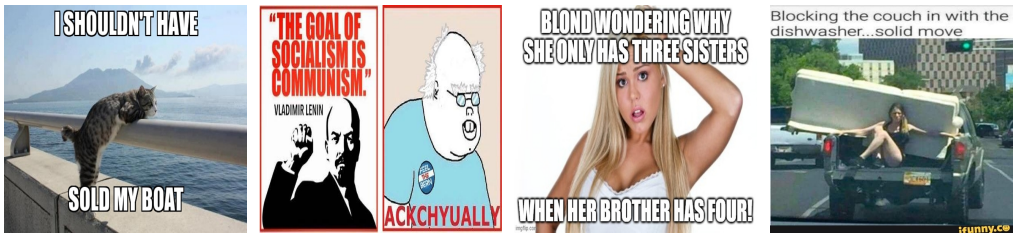
4. Bias Estimation

In order to evaluate whether a given model for misogyny identification is biased, we adopt the approach proposed in (Rizzi et al. 2023). The Bias Estimation strategy is composed of the following main steps:

1. **Identification of Candidate Biased Elements**, which allows us to identify specific elements related to the different modalities that compose the sample (e.g. visual or textual) that, due to an unbalanced distribution in the training dataset, could lead a model to unfair predictions,
2. **Creation of a Synthetic Dataset** with specific characteristics linked with the candidate elements identified accordingly with the previous step. The proposed synthetic dataset allows evaluating model behaviors in challenging examples,
3. **Estimation of the Model Bias** to quantify how a model could be biased from such elements. In order to comprehensively evaluate the model's ability to solve the classification task, a metric to evaluate model prediction both on a test set, conform with respect to the training data, and on a synthetic dataset with challenging samples will be implemented.

4.1 Candidate Bias Elements Estimation

Unbalanced distributions of specific elements in the dataset might lead classification models to unwanted behaviors, especially in the presence of those specific elements. As highlighted by the literature (Rizzi et al. 2023; Balducci, Rizzi, and Fersini 2023), in a multimodal setting, those elements can manifest in different modalities. In the case of memes, both specific terms or visual elements strongly associated with a given class label can result in a distortion of data-derived models. Those *candidate biased elements* can, in fact, be detected in the text that composes the memes - *candidate biased terms* - or within the objects that describe the visual scene - *candidate biased tags*. In this work, the estimation strategy proposed in (Rizzi et al. 2023) is exploited to recognize candidate-biased elements linked to both modalities. This estimation technique allows the estimation of a score, bounded in the interval between -1 and 1, for each element that appears in the training dataset keeping into account the context in which the



(a) Non-misogynous memes with the negative candidate tag *cat*.

(b) Non-misogynous memes with the negative candidate term *communism*.

(c) Misogynous memes with the positive candidate tag *woman*.

(d) Misogynous memes with the positive candidate term *dishwasher*.

Figure 2
Examples of memes in the training dataset.

elements appear. This estimation strategy, on the one hand, overcomes the Polarized Weirdness Index (PWI) (Poletto et al. 2021) limitations, and, on the other hand, extends the estimation to consider different modalities.

According to the (Rizzi et al. 2023) estimation strategy, given a multimodal dataset $\mathcal{D}$ and a visual or textual element e 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:

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

Where $\mathcal{M}_e$ 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. $P(c^+ \mid T_m)$ represents the probability of a meme m of being associated with the misogynous label, given its terms and tags T_m, and, similarly, $P(c^+ \mid T_m - \{e\})$ denotes the probability of a meme m of being associated with the misogynous label c^+, given its terms and tags, excluding the element e in analysis. The achieved score indicates how likely a given element would induce bias towards the positive class (high positive scores) and towards the negative class (low negative scores). Intuitively, terms with scores close to zero are considered neutral with respect to a given label. Examples of memes included in the training dataset are reported in Figure 2.

We report in Tables 1 and 2 the set of biased terms and biased tags identified on the MAMI training dataset. From Table 1, it is possible to see how the set of candidate-biased terms with the highest score for the misogynous class include words, like *dishwasher* and *chick*, typically associated with specific misogyny categories, and others, like *whore* to identification, confirming the ability of the approach to identify elements linked with the different types of misogyny present in the dataset (see Section 3). Additionally, tokens representing websites that have been used to collect memes in the dataset creation phase also appear in the list, suggesting that some websites are more prone in sharing misogynistic memes. Moreover, few terms that achieve positive scores correspond to seed words used for memes' collection (e.g., *whore*); confirming the ability of the proposed approach to capture the *Selection Bias* (i.e., bias introduced in the dataset-creation phase). It is however, easy to notice the presence of other terms

Table 1
Top-10 candidate biased terms.

Candidate Biased Terms			
Misogynous		**Not Misogynous**	
Term	**Score**	**Term**	**Score**
demotivational	0.39	mcdonald	-0.26
dishwasher	0.38	ambulance	-0.24
promotion	0.35	communism	-0.23
whore	0.35	anti	-0.21
chick	0.34	valentine	-0.20
motivate	0.33	developer	-0.20
chloroform	0.30	template	-0.20
blond	0.30	weak	-0.19
diy	0.30	zipmeme	-0.18
belong	0.28	identify	-0.17

Table 2
Top-10 candidate biased tags.

Candidate Biased Tags			
Misogynous		**Not Misogynous**	
Tag	**Score**	**Tag**	**Score**
Woman	0.11	Penguin	-0.27
Earring	0.11	Cat	-0.26
Lip	0.11	Whisker	-0.23
Strap	0.11	Beak	-0.18
Tire	0.10	Gun	-0.17
Eyebrow	0.10	Dog	-0.16
Girl	0.09	Toy	-0.15
Teeth	0.08	Paw	-0.15
Short	0.08	Animal	-0.14
Dress	0.08	Bear	-0.14

(e.g., *chloroform*), demonstrating the ability of the proposed approach to generalize with respect to the dataset creation process and include elements that may induce bias due to their unintended unbalanced distribution.
For what concerns tokens linked with the negative label (not misogynous class), the candidate biased terms include very general words commonly used in several popular memes. Analogous considerations can be drawn for the visual component.

4.2 Synthetic Dataset

Analogously to what has been performed in (Rizzi et al. 2023), a *synthetic dataset* has been created collecting memes with specific characteristics linked with the presence

of the identified candidate elements. The evaluation of classification models in such memes highlights the bias of the models. For the data collection, the following procedure has been followed:

- Considering all the biased candidate elements with a positive (tags E_o^+ and terms E_t^+) and negative (tags E_o^- and terms E_t^-) score, memes have been collected for both classes such that:
  - not misogynous memes containing e_t^+ (or e_o^+) does not contain any other biased candidate terms (or tags) with a negative score. This is to evaluate the impact of the selected biased element in introducing a bias towards the misogynous class in not misogynous memes;
  - misogynous memes containing e_t^+ (or e_o^+) does not contain any other biased candidate terms (or tags) with a positive score. This is to verify if the model, given the presence of biased element in introducing a bias towards the non misogynous class, is able to perform well on misogynous memes.
- Analogously, misogynous and not misogynous memes according to the candidate biased terms and tags with a negative score, have been collected following a similar procedure.

As previously mentioned, this paper represent an extension of a previous work, therefore the same synthetic dataset as presented in (Balducci, Rizzi, and Fersini 2023) will be adopted and later recalled as *synt*. Examples of memes included in the Synthtetic dataset are reported in Figure 3.

(a) Non-misogynous memes with the positive candidate term "dishwasher".

(b) Misogynous memes with the positive candidate term "dishwasher".

(c) Misogynous memes with the negative candidate term "mcdonald".

(d) Non-misogynous memes with the negative candidate term "mcdonald".

Figure 3
Examples of memes in the synthetic dataset.

4.3 Multimodal Bias Estimation (MBE)

In order to measure if a given model is affected by bias, the **Multimodal Bias Estimation** (MBE) metric introduced in (Rizzi et al. 2023) has been adopted. The MBE metric is a combination of two AUC-based measure that measure model performance both on the official MAMI test set (AUC_{raw}), and on the synthetic dataset (AUC_{synt}).

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

As shown in Equation 3, the AUC_{synt} measure capture different aspect of the synthetic dataset including the following:

- $AUC_{Subgroup}(\cdot)$, 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.

$$\begin{aligned} AUC_{synt} = & \frac{1}{2}\frac{\sum_{t\in T} AUC_{\text{Subgroup}}(\mathcal{M}_t) + \sum_{t\in T} AUC_{BPSN}(\mathcal{M}_t) + \sum_{t\in T} AUC_{BNSP}(\mathcal{M}_t)}{|T|} \\ & + \frac{1}{2}\frac{\sum_{i\in I} AUC_{\text{Subgroup}}(\mathcal{M}_i) + \sum_{i\in I} AUC_{BPSN}(\mathcal{M}_i) + \sum_{i\in I} AUC_{BNSP}(\mathcal{M}_i)}{|I|} \end{aligned} \tag{3}$$

In particular, $\mathcal{M}_t$ represents the subgroup of memes identified by the presence of a term t from the subset T of selected biased terms. $\mathcal{M}_i$ denotes the subgroup of memes identified by the presence of a tag i from the subset I of selected biased tags.
The MBE metric ranges in the interval $[0, 1]$, allowing a robust comparison among different models evaluating both the ability to perform a good prediction on the raw test data and simultaneously the potential significant performance on memes that, opportunely selected to include specific characteristics that can lead models to a biased prediction.

5. Debiasing Strategies

5.1 Baseline Models

In order to detail the proposed mitigation strategies, which are based on an Ensemble method called Bayesian Model Averaging (BMA) (Fersini, Messina, and Pozzi 2014), several baseline models have been considered for distinguishing between misogynous and non-misogynous memes. In particular, the following models have been taken into account:

- **Support Vector Machine** (SVM) excels in classification tasks by finding the optimal hyperplane that maximizes the margin between different classes, making it robust in high-dimensional spaces. This approach has been effectively used in the domain of hate speech detection, for instance by (Asogwa et al. 2022; MacAvaney et al. 2019).
- **K-Nearest Neighbors** (KNN) utilizes the proximity of data points to classify new instances, making it effective for classification tasks where

local patterns are important and computation cost is not a significant concern. This approach has been proven to be effective for several classification tasks, including hate speech detection (Cahyana et al. 2022; Prasetyo and Samudra 2021).

- **Naive Bayes** classifier calculates the probability of each class based on the features of the data, assuming independence between features, making it efficient and suitable for text classification and other tasks with high-dimensional feature spaces. From the experiments carried out by (Ruwandika and Weerasinghe 2018), Naive Bayes classifier with Tf-idf features performed best in comparison with several supervised and unsupervised models.
- **Decision Tree** recursively splits the data based on features, creating a tree-like structure that makes decisions by following paths from the root to the leaf nodes, making it interpretable and suitable for tasks where understanding the decision-making process is important. Decision trees have shown promising results in dealing with highly unstructured data because they do not require data scaling and are usually adopted for several classification tasks (Ruwandika and Weerasinghe 2018).
- **Multi-layer Perception** (MLP) is a type of artificial neural network composed of multiple layers of nodes (neurons), which can learn complex patterns in data and perform well in classification tasks with large datasets, given sufficient computational resources for training. This approach, eventually embedded within the BERT-based models in a feed-forward artificial neural network, like in (Anjum 2023), is proven to be effective for hate speech detection.

All the models have been trained independently on each unimodal representation of the memes, i.e., using the textual and visual sources separately. In particular, for what concerns the textual component, each textual transcription obtained with Optical Character Recognition techniques has been embedded with the Universal Sentence Encoder (USE) (Cer et al. 2018).
For what regards the visual component, the image that composes the memes has been processed to identify the objects that compose it (*object tags*) by the Scene Graph Generation method (Han et al. 2021). For each sample, an n-dimensional vector containing the probabilities that the given memes contain one or more pre-defined objects is derived.
Finally, the Bayesian Model Averaging (BMA) (Fersini, Messina, and Pozzi 2014) ensemble paradigm has been adopted to combine the selected classifiers. Three different BMA ensembles have been derived, i.e., (a) BMA on Visual Component, (b) BMA on Textual Component, and (c) BMA on Multimodal Components (both Visual and Textual).

5.2 Mitigation Strategies

Bias mitigation is adopted in both unimodal and multi-modal contexts. In the unimodal setting, only the considered modality is mitigated. In a multi-modal scenario, all the models based on visual and textual components that compose the ensemble are mitigated. In particular, two different types of debiasing strategies are experimented, i.e., at training time and at inference time.

5.2.1 Debiasing at training time

This approach requires retraining the baseline models to obtain a mitigated prediction. In order to mitigate each model at training time, we follow the approach presented in (Balducci, Rizzi, and Fersini 2023).

Masking Mitigation (MM). is proposed. In particular, for what concerns the textual component, each biased term is masked according to the class label that they affect more (see Table 1). Any given biased term, estimated using the strategy presented in section 4, is masked in the training dataset according to the class towards they induce bias. In particular, if a candidate biased term induces a bias towards the misogynous label, then it is replaced with a positive mask [POS-MASK] in misogynous memes. On the contrary, if a candidate biased term induces a bias towards the not misogynous label, then it is replaced with a negative mask [NEG-MASK] in not misogynous memes. An example is reported in the following.

`Original Text`: *When you can't afford a new* ***dishwasher*** *so you...*

`Masked Text`: *When you can't afford a new [POS-MASK] so you...*

Regarding the visual component, when a candidate biased tag is present, the probability value of that tag is set equal to 0, and a new feature indicating the presence of the masking is added to the original n-dimensional vector. A toy example is reported in Figure 4.

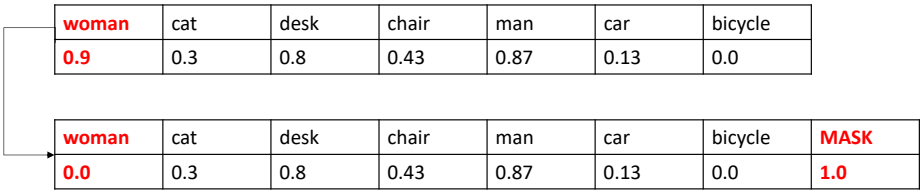


Figure 4
Visual Masking

5.2.2 Debiasing at inference time

This strategy is applied once the baseline models' output probabilities are estimated at inference time, not requiring any re-training step. We propose two novel debiasing strategies at inference time: (1) Bayesian Averaging Model - Class-based Mitigation and (2) Bayesian Model Averaging Element-based Mitigation.

Bayesian Averaging Model Class-based Mitigation (BMA-CM). The proposed Model-based mitigation is based on a smoothed estimation of the posterior probability of the Bayesian Model Averaging ensemble. In particular, the ensemble model is debiased according to the presence of a given biased element. In particular, if a meme contains a candidate-biased element that introduces a bias towards the misogynous label (e.g., *dishwasher*), then the positive probability resulting from the BMA ensemble strategy, $P(c_m^+)$, is penalized by the MBE^+ (i.e., the MBE measure computed as shown in Equation 2, on the subset defined by the misogynous candidate elements), while the

posterior probability that refers to negative class, $P(c_m^-)$, is not mitigated. This implies that if a given positive candidate elements e^+ appears within a meme m, $e^+ \in m$, then:

$$\begin{aligned} P(c_m^+) &= MBE^+ \sum_{g \in G} P(c^+(m) \mid g, \mathcal{D}) F_1^+(g) \\ P(c_m^-) &= \sum_{g \in G} P(c^-(m) \mid g, \mathcal{D}) F_1^-(g) \end{aligned} \tag{4}$$

where $\mathcal{D}$ is the considered multimodal dataset, c^+ and c^- represent the misogynous and non-misogynous labels respectively, and $\mathcal{G}$ represents the set of models included in the BMA.

Similarly, if a meme contains a candidate-biased element e^- that introduces a bias towards the not misogynous label (e.g., *ambulance*), then the negative probability resulting from the BMA ensemble strategy is penalized by the MBE^- (i.e., the MBE measure computed on the subset defined by the not misogynous candidate elements), while the posterior probability that refers to positive class is not mitigated. This implies that if $e^- \in m$, then:

$$\begin{aligned} P(c_m^+) &= \sum_{g \in G} P(c^+(m) \mid g, \mathcal{D}) F_1^+(g) \\ P(c_m^-) &= MBE^- \sum_{g \in G} P(c^-(m) \mid g, \mathcal{D}) F_1^-(g) \end{aligned} \tag{5}$$

Bayesian Model Averaging Element-based Mitigation (BMA-EM). This mitigation strategy, for both textual and visual components, is based on a smoothed estimation of the posterior probability of each unimodal classifier enclosed in the Bayesian Model Averaging ensemble but only debiasing the models according to the presence of a given biased element. In particular, if a meme contains a candidate-biased element e^+ that introduces a bias towards the misogynous label (e.g. *dishwasher*), then the contribution that each model gives to the posterior probability to the positive class is penalized by the MBE_e^+ (i.e. the MBE measure computed on the subset defined by that specific candidate element), while the posterior probability that refers to negative class is not mitigated. MBE_e^+ in this case is obtained considering $AUC_{BNSP}(\cdot)$ $AUC_{BPSN}(\cdot)$ $AUC_{BNSP}(\cdot)$ where the subgroup is composed by only by the element e^+. For each model, bias introduced by an element is represented by MBE_e^+. This implies that if $e^+ \in m$, then:

$$\begin{aligned} P(c_m^+) &= \sum_{g \in G} P(c^+(m) \mid g, \mathcal{D}) F_1^+(g) MBE_e^+ \\ P(c_m^-) &= \sum_{g \in G} P(c^-(m) \mid g, \mathcal{D}) F_1^-(g) \end{aligned} \tag{6}$$

Similarly, if a meme contains a candidate-biased element e^- that introduces a bias towards the not misogynous label (e.g. *ambulance*), then the contribution that each model gives to the posterior probability to the negative class is penalized by the MBE_e^- (i.e. the MBE measure computed on the subset defined by e^-), while the posterior probability

that refers to positive class is not mitigated. This implies that if $e^- \in m$, then:

$$\begin{aligned} P(c_m^+) &= \sum_{g \in G} P(c^-(m) \mid g, \mathcal{D}) F_1^+(g) \\ P(c_m^-) &= \sum_{g \in G} P(c^-(m) \mid g, \mathcal{D}) F_1^-(g) * MBE_e^- \end{aligned} \tag{7}$$

6. Experimental Results

We report in this section the results of the proposed mitigation strategies, comparing their performance with several baseline approaches. In particular, we report AUC_{raw}, AUC_{synt} and MBE related to each model enclosed in the ensemble, i.e., Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naive Bayes (NB), Decision Tree (DT), and Multi-layer Perception (MLP), together with their Bayesian Model Averaging (BMA). Since the currently available transformer-based models can be used as baselines, we also included BERT (Devlin et al. 2019) as a benchmark model for the textual component, ViT (Su et al. 2019) for the visual component, and CLIP (Radford et al. 2021) as a multimodal model that considers both the textual and the visual sources. We also show the performance of the Masking Mitigation on BMA (BMA-MM)(Balducci, Rizzi, and Fersini 2023), Bayesian Averaging Model Class-based Mitigation (BMA-CM) and Bayesian Averaging Model Element-based Mitigation (BMA-EM). All the models enclosed within the BMA-based ensemble share a uniform input representation (i.e., the embedding representation for the extracted text for the unimodal text-based models, the vector containing the probabilities of the selected tag within a given sample for the unimodal image-based models, and their concatenation for the multimodal models). On the other hand, the pre-trained large language models adopt peculiar feature extraction, leading therefore to a different input representation, which results to be incoherent with respect to one adopted for bias evaluation. Therefore, the state-of-the-art pre-trained models have not been enclosed in the BMA ensembles and the corresponding mitigated version. Finally, we report, as baseline debiasing technique available in the state of the art REPAIR (Li and Vasconcelos 2019) as a benchmark mitigation model. REPAIR 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 particular, given a weight w_i for each meme i, it keeps $p = 50\%$ examples with the largest weight w_i from each class.

We show in Tables 3-5, the comparison between all the considered models, distinguished according to the modalities used to perform the training and the corresponding mitigation phase. A few considerations can be derived from Table 3, where the models have been trained using the textual component only: (1) training on the textual component only leads all the models to obtain good results on both raw and $synt$ test sets, (2) BMA is able to achieve remarkable results compared with the baselines, (3) the proposed Masking Mitigation strategy (BMA-MM) significantly outperforms all the baseline models and the original BMA, but also the REPAIR strategy. BMA-CM and BMA-EM also outperform all the baselines and the original BMA without the need to re-train models. Regarding the inclusion of pre-trained transformer-based LM, we reported state-of-the-art approaches to provide an additional comparison, in particular, BERT model has been fine-tuned on the task. In this case, all the models outperformed the fine-tuned BERT. The three proposed strategies are able to maintain good recognition performance on the raw test set, still improving significantly the

Table 3
Model performance using the textual component only. Underline denotes the best result while **Bold** reflects that the mitigated model outperforms the best non-mitigated approach (BMA)

Textual Component Only			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.7183	0.7508	0.7346
KNN	0.7145	0.7172	0.7158
NB	0.7011	0.7398	0.7204
DT	0.6352	0.7370	0.6861
MLP	0.7240	0.7440	0.7340
BERT	0.6547	0.6590	0.6568
BMA	**0.7312**	0.7606	0.7459
REPAIR	0.6679	0.6918	0.6798
BMA-CM	**0.7312**	**0.7699**	**0.7506**
BMA-EM	**0.7314**	**0.7744**	**0.7529**
BMA-MM	0.7308	**0.7921**	**0.7615**

generalization capabilities on the controversial memes available in the $synt$ test set. For

Table 4
Model performance using the visual component only. Underline denotes the best result while **Bold** reflects that the mitigated model outperforms the best non-mitigated approach (BMA)

Visual Component Only			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.6806	0.5964	0.6385
KNN	0.6612	0.5627	0.6119
NB	0.6634	0.5511	0.6072
DT	0.6491	0.6077	0.6284
MLP	0.6693	0.6093	0.6393
ViT	0.5330	0.5260	0.51997
BMA	**0.6877**	0.6085	**0.6481**
REPAIR	0.6632	0.5814	0.6223
BMA-CM	**0.688**	0.5970	0.6425
BMA-EM	0.6875	0.5930	0.6402
BMA-MM	0.6625	**0.6218**	0.6419

what concerns Table 4, where the models have been trained using the visual component only, the considerations are a bit different. As demonstrated in other state-of-the-art studies (Rizzi et al. 2023), the visual component is less impactful on the recognition capabilities than the textual one. We hypothesize that the reduced contribution of the pictorial component is mainly due to conceptualization issues in relating a given object to an abstract concept (e.g. dishwasher). However, also in this case, BMA is able to achieve better results than the baselines and BMA-MM, BMA-CM, BMA-EM, is still able to significantly outperform the baseline models and REPAIR but not the original BMA. Also in this case BMA, BMA-MM, BMA-CM, BMA-EM, outperform Vision-based

Transformer model (ViT) fine-tuned on the binary task. Regarding the performance of

Table 5
Model performance using the multimodal components. Underline denotes the best result while **Bold** reflects that the mitigated model outperforms the best non-mitigated approach (BMA)

Multimodal Components			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.7620	0.7615	0.7618
KNN	0.7567	0.6937	0.7252
NB	0.7319	0.7411	0.7365
DT	0.7501	0.7361	0.7184
MLP	0.7501	0.7378	0.7440
CLIP	0.6209	0.6843	0.6526
BMA	**0.7794**	0.7991	0.7892
REPAIR	0.7247	0.6792	0.7020
BMA-CM	0.7726	**0.8154**	**0.7940**
BMA-EM	0.7722	**0.8104**	**0.7913**
BMA-MM	0.7648	**0.8187**	**0.7917**

the multimodal settings reported in Table 5, we can assert that not only the proposed mitigation strategies significantly outperform all the other configurations presented above, but they are also able to achieve a very promising compromise between raw and $synt$ samples that facilitate the adoption of the BMA-MM, BMA-CM, and BMA-EM in a real setting. In order to provide an additional comparison from a multimodal perspective, the CLIP model has been fine-tuned. Also, in this setting, BMA and the corresponding mitigated version outperformed the state-of-the-art multimodal transformer-based model.

7. Conclusions

This paper addressed the problem of mitigating misogynous meme detection. In particular, a candidate biased element estimation and corresponding mitigation strategies are proposed to perform fair prediction in a real setting. The proposed approach at training time (BMA-MM) was validated on a benchmark dataset and achieved remarkable results both in terms of prediction and generalization capabilities, reducing the bias in a significant way. Also, the proposed approaches at inference time, BMA-CM, and BMA-EM achieved notable results, reducing bias without the need to re-train models on the dataset and reducing computational resources used. This facilitates the adoption of these strategies on pre-trained models, which are increasingly used today, at inference time.

Acknowledgments

We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. The work of Elisabetta Fersini has been partially funded by MUR under the grant ReGAInS, *Dipartimenti di Eccellenza 2023-2027* of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca.

References

Andreasen, Maja Brandt. 2021. 'rapeable'and 'unrapeable'women: the portrayal of sexual violence in internet memes about #metoo. *Journal of Gender Studies*, 30(1):102–113.

Anjum, Rahul Katarya. 2023. Hatedetector: Multilingual technique for the analysis and detection of online hate speech in social networks. *Multimedia Tools and Applications*, 83(16):48021–48048.

Anzovino, Maria, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64, Paris, France, June. Springer.

Asogwa, Doris Chinedu, Chiamaka Ijeoma Chukwuneke, Chigozie Chidimma Ngene, and Nkiru Gloria Anigbogu. 2022. Hate speech classification using svm and naive bayes. *arXiv preprint arXiv:2204.07057*.

Balducci, Gianmaria, Giulia Rizzi, and Elisabetta Fersini. 2023. Bias mitigation in misogynous meme recognition: A preliminary study. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, Venice, Italy, November 30th - December 2nd. CEUR Workshop Proceedings.

Bashar, Md Abul, Richi Nayak, and Nicolas Suzor. 2020. Regularising lstm classifier by transfer learning for detecting misogynistic tweets with small training set. *Knowledge and Information Systems*, 62:4029–4054.

Cahyana, Nur Heri, Shoffan Saifullah, Yuli Fauziah, Agus Sasmito Aribowo, and Rafal Drezewski. 2022. Semi-supervised text annotation for hate speech detection using k-nearest neighbors and term frequency-inverse document frequency. *International Journal of Advanced Computer Science and Applications*, 13(10).

Calderón-Suarez, Ricardo, Rosa M. Ortega-Mendoza, Manuel Montes-Y-Gómez, Carina Toxqui-Quitl, and Marco A. Márquez-Vera. 2023. Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases. *IEEE Access*, 11:13179–13190.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November.

Chen, Lei and Hou Wei Chou. 2022. RIT boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States, 14th-15th July. Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, volume 1, page 2, Minneapolis, Minnesota, June.

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

Fersini, Elisabetta, Antonio Candelieri, and Lorenzo Pastore. 2023. On the generalization of projection-based gender debiasing in word embedding. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 336–343, Varna, Bulgaria, September.

Fersini, Elisabetta, Francesca Gasparini, and Silvia Corchs. 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 (ACIIW)*, pages 226–231, Cambridge, UK, 3rd-6th September.

Fersini, Elisabetta, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States, 14th-15th July. Association for Computational Linguistics.

Fersini, Elisabetta, Enza Messina, and Federico Alberto Pozzi. 2014. Sentiment analysis: Bayesian Ensemble Learning. *Decision Support Systems*, 68:26–38.

Fersini, Elisabetta, Giulia Rizzi, Aurora Saibene, and Francesca Gasparini. 2021. Misogynous meme recognition: A preliminary study. In *20th International Conference of the Italian Association for Artificial Intelligence, AIxIA 2021*, Online, December. Springer.

Fontanella, Lara, Berta Chulvi, Elisa Ignazzi, Annalina Sarra, and Alice Tontodimamma. 2024. How do we study misogyny in the digital age? a systematic literature review using a computational linguistic approach. *Humanities and Social Sciences Communications*, 11(1):1–15.

Gasparini, Francesca, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526.

Guo, Kangshuai, Ruipeng Ma, Shichao Luo, and Yan Wang. 2023. Coco at semeval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 469–476, Toronto, Canada, 13th-14th July.

Hakimov, Sherzod, Gullal Singh Cheema, and Ralph Ewerth. 2022. TIB-VA at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. In *The 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States, 14th-15th July. Association for Computational Linguistics.

Han, Xiaotian, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. 2021. Image scene graph generation (sgg) benchmark. *arXiv preprint arXiv:2107.12604*.

Li, Yi and Nuno Vasconcelos. 2019. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9572–9581, Long Beach, CA, USA, June.

MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Nascimento, Francimaria R.S., George D.C. Cavalcanti, and Márjory Da Costa-Abreu. 2022. Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, 201:117032.

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

Perrone, Valerio, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. 2021. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 854–863, Online, May.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Prasetyo, Vincentius Riandaru and Anton Hendrik Samudra. 2021. Hate speech content detection system on twitter using k-nearest neighbor method. In *International Conference on Informatics, Technology, and Engineering 2021 (InCITE 2021): Leveraging Smart Engineering*, volume 2470, Online, August. American Institute of Physics.

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

Rizzi, Giulia, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(5):103474.

Ruwandika, N.D.T. and Ruvan Weerasinghe. 2018. Identification of hate speech in social media. In *2018 18th international conference on advances in ICT for emerging regions (ICTer)*, pages 273–278, Colombo, Sri Lanka, 27th-28th September. IEEE.

Sabat, Benet Oriol, Cristian Canton Ferrer, and Xavier Giro i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv: 1910.02334*.

Shen, Tianshu, Jiaru Li, Mohamed Reda Bouadjenek, Zheda Mai, and Scott Sanner. 2023. Towards understanding and mitigating unintended biases in language model-driven conversational recommendation. *Information Processing & Management*, 60(1):103139.

Sikdar, Sandipan, Florian Lemmerich, and Markus Strohmaier. 2022. Getfair: Generalized fairness tuning of classification models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 289–299, Seoul, South Korea, June.

Singh, Smriti, Amritha Haridasan, and Raymond Mooney. 2023. "female astronaut: Because sandwiches won't make themselves up there": Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, Toronto, Canada, 13th July.

Song, Rui, Fausto Giunchiglia, Yingji Li, Lida Shi, and Hao Xu. 2023. Measuring and mitigating language model biases in abusive language detection. *Information Processing & Management*, 60(3):103277.

Sridhar, Rohit and Diyi Yang. 2022. Explaining toxic text via knowledge enhanced text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–826, Seattle, United States, July. Association for Computational Linguistics.

Su, Weijie, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: pre-training of generic visual-linguistic representations. *CoRR*, abs/1908.08530.

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

Ta, Hoang Thang, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and Alexander Gelbukh. 2022. Transfer learning from multilingual deberta for sexism identification. In *Proceedings of the IberLEF 2022 Workshop co-located with the 38th Intenational Conference of the Spanish Society for Natural Language Processing (SEPLN 2022)*, volume 3202, A Coruña, Spain, September. CEUR-WS.

Van Royen, Kathleen, Karolien Poels, Heidi Vandebosch, and Michel Walrave. 2018. Slut-shaming 2.0. In Michel Walrave, Joris Van Ouytsel, Koen Ponnet, and Jeff R. Temple, editors, *Sexting : motives and risk in online sexual self-presentation*. Palgrave Macmillan Cham, pages 81–98.

We Are Social and Hootsuite. 2022. Digital 2022 global overview report. https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/.

Xia, Mengzhou, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online, July.

Zhi, Jin Mei, Zhou Mengyuan, Mengfei Yuan, Dou Hu, Xiyang Du, Lianxin Jiang, Yang Mo, and XiaoFeng Shi. 2022. PAIC at semeval-2022 task 5: Multi-modal misogynous detection in MEMES with multi-task learning and multi-model fusion. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States, 14th-15th July. Association for Computational Linguistics.

Zhou, Ziming, Han Zhao, Jingjing Dong, Ning Ding, Xiaolong Liu, and Kangli Zhang. 2022. DD-TIG at semeval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (Sem-Eval-2022)*, Seattle, United States, 14th-15th July. Association for Computational Linguistics.

Zmigrod, Ran, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 1651–1661, Florence, Italy, July-August.

Zueva, Nadezhda, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in russian hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69, Online, November.