Bias Mitigation in Misogynous Meme Recognition: A Preliminary Study

Gianmaria Balducci[1,2], Giulia Rizzi[1,3] and Elisabetta Fersini[1,*]

[1]University of Milano-Bicocca, Milan, Italy
[2]PMI Reboot S.r.l., Milan, Italy
[3]Universitat Politècnica de València, Valencia, Spain

Abstract

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 proposed to identify those textual and visual elements that unintendedly affect the model prediction, together with a naive bias mitigation strategy. The proposed approach is able to achieve good recognition performance characterized by promising generalization capabilities.

Keywords

Bias Mitigation, Bias Estimation, Misogyny Identification, Meme

1. Introduction

In the context of social media, memes have become popular as a means of expressing irony or opinions on various topics. However, these memes can also perpetuate discriminatory behaviours towards certain groups and minorities. Misogyny, in particular, has gained attention as a form of ***hateful language*** conveyed through memes in various ways, such as female stereotyping, shaming, objectification, and violence. While misogyny recognition mechanisms have been widely investigated focusing on textual sources (i.e., tweets) [1, 2, 3, 4], misogynous identification in multimodal settings, and in particular on memes, is still in its infancy. In [5], a few naive unimodal and multimodal approaches have been investigated to understand the contribution of textual and visual cues. Further investigations from the same authors [6] have introduced a multimodal approach that considers both visual (in the form of captioning) and textual information to distinguish between misogynous and non-misogynous memes. Recently, the performance of multiple pre-trained and trained from scratch models have been compared to verify if domain-specific pre-training could help to improve the recognition performance [7].

Independently on the textual, visual or multimodal sources, several authors highlighted how the classification models may be subject to ***bias*** that could affect the real performance of the models [8, 9] in a real setting. Most of the investigations propose a few bias estimation metrics and related mitigation policies that are based on a fixed set of seed words to quantify and minimize the bias at the dataset or model level. When dealing with misogynous memes recognition, metrics to estimate the bias and techniques to mitigate it are still missing.

To this purpose, we provide the following main contributions:

- a *candidate biased elements identification* in a multi-modal setting, focusing on both textual and visual constituents of a meme;
- a mitigation strategy at training time, named *Masking Mitigation*, that masks the candidate biased elements to reduce the distortion introduced by their presence.

The rest of the paper is organized as follows. In Section 2 a summary of the state of the art is reported. In Section 3 the candidate biased element identification strategy is detailed. In Section 4 the proposed mitigation strategy is presented. In Section 5 the experimental results are reported. In Section 6 conclusion are reported.

2. Related work

The majority of works on hate content detection focus on tweets, while, only in recent years, they have started to address multimodal content such as memes. For instance, the approach proposed in [5] aims to counter the phenomenon of memes that can convey sexist messages ranging from stereotyping women to shaming, objectification, and violence, investigating both unimodal and multimodal approaches to understand the contribution of textual and visual cues. In [10], the authors indicate how the visual mode may be much more informative

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.
g.balducci1@campus.unimib.it (G. Balducci); g.rizzi10@campus.unimib.it (G. Rizzi); elisabetta.fersini@unimib.it (E. Fersini)
0000-0002-0619-0760 (G. Rizzi); 0000-0002-8987-100X (E. Fersini)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

for detecting hate speech than the linguistic mode in memes. More recently, two benchmark datasets have been proposed to facilitate the investigation related to misogynous meme detection. The first benchmark presented in [11] is composed of 800 memes from the most popular social media platforms. The dataset has been labelled through a crowdsourcing platform, involving 60 subjects, in order to collect three evaluations for each instance. Each instance, labelled according to misogyny, aggressiveness and irony, has been labelled by three annotators from the crowd and three expert labellers. A more recent benchmark has been collected for *MAMI* shared task at SemEval 2022 [12]. The dataset, composed of 10.000 memes for training and 1.000 memes for testing, allowed to approach: (i) the identification of misogynistic memes, and (ii) the recognition of the type of misogyny among potential overlapping categories. For the MAMI challenge, most of the participants [13, 14, 15, 16] exploited pre-trained models and ensemble strategies.

Regarding the potential bias that the models could inherit from the training dataset, most of the investigations focus only on a unimodal setting and more precisely on the textual component [17, 18, 19]. In particular, special attention has been devoted to *identity terms*, i.e. those terms frequently associated with hateful expressions in the dataset referred to a specific target (e.g., woman, wife, girlfriend, etc...). It has been demonstrated that such identity terms lead the models to biased implicit associations between such terms and a given class label, finally originating unfair predictions. In order to counteract the potential bias, several mitigation strategies have been proposed in the literature. One of the most widely used strategies is data augmentation [4, 20, 21], which consists in adding data containing examples of non-toxic comments that bring back those identity terms that have the most disproportionate distribution in the dataset. Alternative solutions are focused on mitigating directly the models by means of specific objective functions [22, 23] or optimization strategies [24, 25, 26]. Although the above-mentioned strategies represent a fundamental step towards bias mitigation, they are defined for unimodal settings. Bias estimation and mitigation for multimodal perspective are still missing for misogynous meme identification.

3. Bias Estimation

In order to understand if a given misogyny identification model is biased, three main steps are performed: (i) Candidate Biased Elements Estimation, which allows us to identify specific textual or visual elements that could lead a model to unfair predictions, (ii) the creation of a Synthetic Dataset with specific characteristics that allow evaluating models behaviours in challenging examples, and (iii) the definition of a metric to quantify how a model could be biased from such elements. The proposed method has been evaluated on the MAMI Dataset [12] consisting of 10.000 memes for training and 1.000 memes for testing. The MAMI test set will later be referred to as *raw*.

3.1. Candidate Bias Elements Estimation

As highlighted in the literature, classification models may be affected by bias: the presence of specific elements can lead the model to an erroneous behaviour by predicting a specific label due to the presence of such elements. This distortion in the investigated data-derived models can be in fact caused by an imbalance distribution, in relation to the prediction label, of specific terms or visual elements strongly associated with a given class label. Those *candidate biased elements* can be distinguished in *candidate biased terms*, which are related to the superimposed text of a meme, and *candidate biased tags*, which are concerned with the objects that describe the scene of a meme. We exploit a novel estimation for identifying candidate biased elements [26] that overcomes the limitations of the Polarized Weirdness Index (PWI) [27], which is unbounded and does not consider the context in which the elements appear, and extended the estimation process to address more than one modality.

Given a multimodal dataset $\mathcal{D}$, 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:

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

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. $P(c^+ \mid T_m)$ represents the probability of a meme m of being associated with the misogynous label, given the terms and tags T_m within the meme itself, and, analogously, $P(c^+ \mid \{T_m - e\})$ denotes the probability of a meme m of being associated with the misogynous label c^+, given the text (tags) present in the instance (meme), excluding the evaluated element e except for the term (tag) in analysis. 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 and tags with a score close to zero, are considered neutral with respect to a given label.

We report in Tables 1 and 2 the set of biased terms and biased tags identified on the MAMI training dataset. As we can see, the set of candidate biased terms with the highest score for the misogynous class is composed of words 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. A few terms identified as convey potential bias are related to the seed words used to collect the dataset (e.g. *whore*), confirming the ability of the proposed approach to capture the bias introduced in the dataset-creation phase (*Selection Bias*). On the other hand, the presence of other terms (e.g. *chloroform*) highlights 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. Concerning the set of terms with the highest negative bias score for the not misogynous class, it is composed of words that are very general and commonly used in a variety of popular memes. An analogous consideration can be drawn for the candidate biased tags.

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 1
Top-10 candidate biased terms.

3.2. Synthetic Dataset

In order to measure the bias of the models when making predictions, a *synthetic dataset* has been created with specific characteristics that can effectively help to highlight the bias of the models given the presence of the candidate biased elements.

In particular, let E_t^+ and E_o^+ be respectively the set of all the biased candidate terms and tags with a positive score, which qualifies elements that are expected to introduce the bias towards the misogynous class. Also, let E_t^- and E_o^- be respectively the set of all the biased candidate terms and tags with a negative score, which qualifies elements that are expected to introduce the bias towards the not misogynous class. Given a specific element $e_t^+ \in E_t^+$ and $e_o^+ \in E_o^+$, we collected misogynous and not misogynous memes according to the following criteria:

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

Table 2
Top-10 candidate biased tags.

- a not misogynous meme is part of the synthetic dataset if it contains e_t^+ (or e_o^+) and it does not contain any biased candidate terms (or tags) with a negative score. This is to evaluate the impact of e_t^+ (or e_o^+) 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_t^+ (or e_o^+) and it does not contain any other element in E_t^+ (or E_o^+). This is to verify if the model, given the presence of e_t^+ (or e_o^+), is able to perform well on misogynous memes.

An analogous procedure has been adopted to create misogynous and not misogynous memes according to the candidate biased terms and tags with a negative score. The synthetic test set will later be recalled as *synt*.

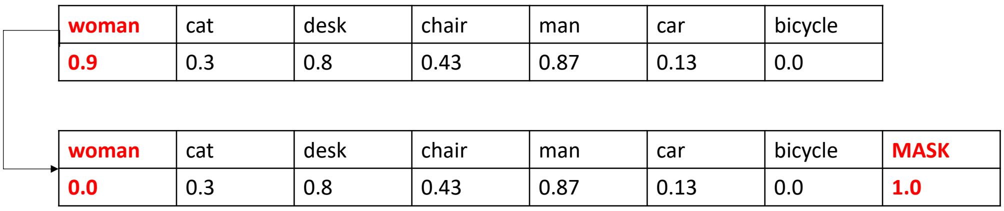
3.3. Multimodal Bias Estimation (MBE)

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

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

where AUC_{synt} is computed as reported in Equation 3. $\mathcal{M}_t$ represents the subgroup of memes identified by the presence of a biased term t, T is the subset of selected

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



woman	cat	desk	chair	man	car	bicycle
0.9	0.3	0.8	0.43	0.87	0.13	0.0

woman	cat	desk	chair	man	car	bicycle	MASK
0.0	0.3	0.8	0.43	0.87	0.13	0.0	1.0

Figure 1: Visual Masking

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 biased 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:

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

4. Debiasing Strategy

Several baseline models have been initially considered for distinguishing between misogynous and not misogynous memes. We trained SVM, KNN, Naive Bayes, Decision Tree, and Multi-layer Perception independently on each unimodal representation of the memes. In particular, the following modalities have been considered as (separate) input space:

- **textual component**, that is the transcription of the text contained within the meme (obtained with OCR) embedded through the Universal Sentence Encoder (USE) [28].
- **visual component**, expressed by the objects identified within the meme (*object tags*) by the Scene Graph Generation method [29] and represented through a n-dimensional vector that denotes if a given meme contains one or more predefined objects with the corresponding probabilities.

The classifiers have been combined, accordingly to each modality (e.g. visual or textual), through a Bayesian Model Averaging (BMA) [30] ensemble paradigm. BMA has been employed also for creating a multimodal ensemble that considers all the predictions provided by the above-mentioned models trained on each representation independently.

4.1. Mitigation Strategy

Bias mitigation is adopted in both unimodal and multi-modal contexts. In an 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 order to debias the model at training time (and inference time), a **Masking Mitigation** 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 to the strategy presented in section 3, 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 1.

5. Experimental Results

We report in this section the results of the proposed mitigation strategy, comparing the performance with several 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). We also show the performance of the proposed Masking Mitigation on BMA (BMA-MM). Finally, we report a baseline debiasing technique available in the state of the art. In particular, we used REPAIR [31] as a benchmark mitigation model. 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 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 T-test has been performed to compute the statistical equality with a pairwise analysis between the best-performing approach (BMA) against the compared mitigation strategies, i.e. BMA-MM and REPAIR.

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 lead 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-MM is able to maintain good recognition performance on the raw test set, still improving significantly the generalization capabilities on the controversial memes available in the $synt$ test set.

Textual Component Only			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.7202	0.7801	0.7501
KNN	0.7173	0.7041	0.7107
NB	0.7010	0.7687	0.7348
DT	0.6301	0.7475	0.6880
MLP	0.7257	0.7521	0.7389
BMA	**0.7326**	0.7841	0.7583
REPAIR	0.6775	0.6811	0.6793
BMA-MM	0.7325	**0.8052**	**0.7689***

Table 3
Model performance using the textual component only. **Bold** denotes the best result, while (*) reflects that the mitigated model outperforms the best non-mitigated approach (BMA) and the improvement is statistically significant.

Visual Component Only			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.6808	0.5918	0.6363
KNN	0.6623	0.5942	0.6283
NB	0.6635	0.5773	0.6204
DT	0.6499	0.5888	0.6194
MLP	**0.6912**	0.6047	0.6480
BMA	0.6870	0.5990	0.6430
REPAIR	0.6651	0.5922	0.6286
BMA-MM	0.6655	**0.6416**	**0.6535***

Table 4
Model performance using the visual component only. **Bold** denotes the best MBE, while (*) reflects that the mitigated model outperforms the best non-mitigated approach (BMA) and the improvement is statistically significant.

For 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 [26], 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 to relate a given object to a an abstract concept (e.g. dishwasher). However, also in this case, BMA is able to achieve better results than the baselines and BMA-MM is still able to significantly outperform the original BMA and REPAIR.

Multimodal Components			
Model	AUC_{raw}	AUC_{synt}	MBE
SVM	0.7632	0.7794	0.7713
KNN	0.7590	0.7277	0.7433
NB	0.7326	0.7794	0.7560
DT	0.7006	0.7483	0.7245
MLP	0.7690	0.7374	0.7532
BMA	**0.7802**	0.7908	0.7855
REPAIR	0.7360	0.6982	0.7171
BMA-MM	0.7676	**0.8306**	**0.7991***

Table 5
Model performance using the multimodal components. **Bold** denotes the best result, while (*) reflects that the mitigated model outperforms the best non-mitigated approach (BMA) and the improvement is statistically significant.

Regarding the performance of the multimodal settings reported in Table 5, we can assert that not only the proposed mitigation strategy significantly outperforms all the other configurations presented above, but it is also able to achieve a very promising compromise between raw and $synt$ samples that facilitate the adoption of the BMA-MM in a real setting.

6. Conclusions

This paper addressed the problem of mitigating misogynous meme detection. In particular, a candidate biased element estimation and a corresponding mitigation strategy is proposed to perform fair prediction in a real setting. The proposed approach, validated on a benchmark dataset, achieved remarkable results both in terms of prediction and generalization capabilities, reducing the bias in a significant way.

Acknowledgments

The work of Elisabetta Fersini 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 MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

References

[1] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2018, pp. 57–64.

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

[3] H. T. Ta, A. B. S. Rahman, L. Najjar, A. Gelbukh, Transfer learning from multilingual deberta for sexism identification, in: CEUR Workshop Proceedings, volume 3202, CEUR-WS, 2022.

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

[5] E. Fersini, F. Gasparini, S. Corchs, 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), 2019, pp. 226–231.

[6] E. Fersini, G. Rizzi, A. Saibene, F. Gasparini, Misogynous meme recognition: A preliminary study, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2021.

[7] S. Singh, A. Haridasan, R. Mooney, "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), 2023, pp. 150–159.

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

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

[10] B. O. Sabat, C. C. Ferrer, X. G. i Nieto, Hate speech in pixels: Detection of offensive memes towards automatic moderation, 2019. arXiv:1910.02334.

[11] F. Gasparini, G. Rizzi, A. Saibene, E. Fersini, Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content, Data in brief 44 (2022) 108526.

[12] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549.

[13] Z. Zhou, H. Zhao, J. Dong, N. Ding, X. Liu, K. Zhang,

DD-TIG at semeval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification, in: The 16th International Workshop on Semantic Evaluation, 2022.

[14] L. Chen, H. W. Chou, 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: The 16th International Workshop on Semantic Evaluation, 2022.

[15] S. Hakimov, G. S. Cheema, R. Ewerth, 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, 2022.

[16] J. M. ZHI, Z. Mengyuan, M. Yuan, D. Hu, X. Du, L. Jiang, Y. Mo, X. Shi, PAIC at semeval-2022 task 5: Multi-modal misogynous detection in MEMES with multi-task learning and multi-model fusion, in: The 16th International Workshop on Semantic Evaluation, 2022.

[17] D. Nozza, C. Volpetti, E. Fersini, Unintended bias in misogyny detection, in: IEEE/WIC/ACM International Conference on Web Intelligence, WI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 149–155. URL: https://doi.org/10.1145/3350546.3352512. doi:10.1145/3350546.3352512.

[18] N. Zueva, M. Kabirova, P. Kalaidin, Reducing unintended identity bias in russian hate speech detection, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020, pp. 65–69.

[19] F. R. Nascimento, G. D. Cavalcanti, M. Da Costa-Abreu, Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning, Expert Systems with Applications 201 (2022) 117032.

[20] R. Zmigrod, S. J. Mielke, H. Wallach, R. Cotterell, 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, 2019, pp. 1651–1661.

[21] K. Guo, R. Ma, S. Luo, Y. Wang, Coco at semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 469–476.

[22] M. Xia, A. Field, Y. Tsvetkov, Demoting racial bias in hate speech detection, in: Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, 2020, pp. 7–14.

[23] R. Sridhar, D. Yang, 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, Association for Computational Linguistics, Seattle, United States, 2022, pp. 811–826. URL: https://aclanthology.org/2022.naacl-main.59. doi:10.18653/v1/2022.naacl-main.59.

[24] V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, C. Archambeau, Fair bayesian optimization, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 854–863.

[25] S. Sikdar, F. Lemmerich, M. Strohmaier, Getfair: Generalized fairness tuning of classification models, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 289–299.

[26] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Information Processing & Management 60 (2023) 103474.

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

[28] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal Sentence Encoder for English, in: Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, 2018, pp. 169–174.

[29] X. Han, J. Yang, H. Hu, L. Zhang, J. Gao, P. Zhang, Image scene graph generation (sgg) benchmark, 2021. arXiv:2107.12604.

[30] E. Fersini, E. Messina, F. A. Pozzi, Sentiment analysis: Bayesian Ensemble Learning, Decision Support Systems 68 (2014) 26–38.

[31] Y. Li, N. Vasconcelos, Repair: Removing representation bias by dataset resampling, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9572–9581.