Unintended Bias Detection and Mitigation in Misogynous Memes

Gitanjali Kumari[1] **Anubhav Sinha** [1] **Asif Ekbal**[1]
[1]Department of Computer Science and Engineering,
[1]Indian Institute of Technology Patna, India
{gitanjali_2021cs03,anubhav_2111mc04,asif}@iitp.ac.in

Abstract

Online sexism has become a concerning issue in recent years, especially conveyed through memes. Although this alarming phenomenon has triggered many studies from computational linguistic and natural language processing points of view, less effort has been spent analyzing if those misogyny detection models are affected by an unintended bias. Such biases can lead models to incorrectly label non-misogynous memes misogynous due to specific identity terms, perpetuating harmful stereotypes and reinforcing negative attitudes. This paper presents the first and most comprehensive approach to measure and mitigate unintentional bias in the misogynous memes detection model, aiming to develop effective strategies to counter their harmful impact. Our proposed model, the **C**on**t**e**x**tualized **S**cene **G**raph-based **M**ultimodal **Net**work (CTXSGMNet), is an integrated architecture that combines VisualBERT, a CLIP-LSTM-based memory network, and an unbiased scene graph module with supervised contrastive loss, achieves state-of-the-art performance in mitigating unintentional bias in misogynous memes. Empirical evaluation, including both qualitative and quantitative analysis, demonstrates the effectiveness of our CTXSGMNet framework on the SemEval-2022 Task 5 (**MAMI** task) dataset, showcasing its promising performance in terms of Equity of Odds and F1 score. Additionally, we assess the generalizability of the proposed model by evaluating their performance on a few benchmark meme datasets, providing a comprehensive understanding of our approach's efficacy across diverse datasets [1].

1 Introduction

In recent years, the proliferation of memes on social media platforms like Facebook, Twitter, and Instagram has gained significant attention due to their widespread influence and potential to shape public discourse. Many memes, despite being humorous, use extremism and dark humor to promote societal harm (Kiela et al., 2020a; Kirk et al., 2021; Kumari et al., 2021; Bandyopadhyay et al., 2023). Among the various types of memes, misogynous memes hold a unique place, which exhibits and promote hatred, sexism, derogatory attitudes, harmful stereotypes, and objectification of women, and has become a concerning issue (Attanasio et al., 2022; Zhou et al., 2022; Arango et al., 2022; Zhang and Wang, 2022; Zhou et al., 2022; Chen and Chou, 2022; Fersini et al., 2022). While prior research has mostly focused on developing robust deep-learning models to identify such memes (Rijhwani et al., 2017; Sharma et al., 2020a; Kiela et al., 2020a; Suryawanshi et al., 2020; Pramanick et al., 2021a; Hossain et al., 2022; Sharma et al., 2022), the effect of the presence of unintentional biases within these models remains a difficult and understudied problem. In this context, Arango et al. (2022) highlights this particular error due to training data when a model misclassifies a meme as misogynous solely based on the presence of a certain image (like a woman) or inclusion of a specific word called **identity terms**[2] like *"kitchen," "woman," "dishwasher," "bitch" etc.*. These images or identity terms unintentionally introduce bias into the classification task (Attanasio et al., 2022; Nobata et al., 2016). More precisely, "a model exhibits **unintended bias** when its performance is better for samples that include specific identity terms compared to samples that contain others (Dixon et al., 2018)". (c.f. Figure 1 for such unintended bias in misogynous meme identification task). The identification and mitigation of such unintentional bias in misogynous meme classification are of paramount importance, which has compelled researchers to

[1]Codes are available at this link: https://www.iitp.ac.in/~ai-nlp-ml/resources.html as-well-as at our GitHub repository: https://github.com/Gitanjali1801/Gender_bias

[2]WARNING: This paper contains meme samples that are offensive in nature.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2719–2733

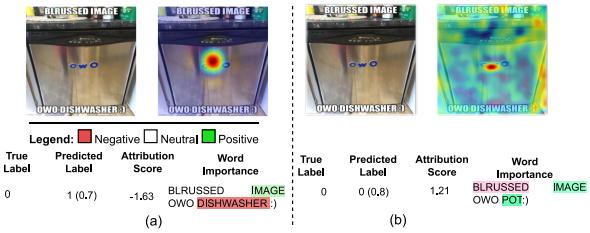

Figure 1: Analysis of the biased prediction made by a classifier on a *MAMI* dataset sample. (a) The baseline classifier's incorrect labels are influenced by an identity term "DISHWASHER." (b) Substituting the identity term with a neutral term "POT" alters the model's prediction.

explore diverse techniques that leverage both textual and visual components of memes (Nozza et al., 2019; Rao and Rao, 2022; Srivastava, 2022; Zhou et al., 2022; Godoy and Tommasel, 2021; Li et al., 2023b). However, a notable research gap remains in effectively leveraging contextual understanding to combat this bias as Zhou et al. (2022) highlights the reason for such unintentional bias as the lack of context and external knowledge during classifying misogynous memes.

Our work addresses this gap by introducing *CTXSGMNet* (Contextualized Scene Graph-based Multimodal Network), a novel architecture designed to enhance contextual analysis and mitigate unintended bias from misogynous meme classifiers. CTXSGMNet integrates VisualBERT, a CLIP-LSTM-based memory network, and an unbiased scene graph (Tang et al., 2020) modules. The CLIP-LSTM-based memory network retains contextual information, considering the sequential nature of textual content. Simultaneously, the unbiased scene graph module captures the unbiased semantic relationships between objects in memes, facilitating the identification of contextual cues. On top of it, CTXSGMNet improves multimodal representations while training by incorporating supervised contrastive learning (SCL) and cross-entropy loss jointly. SCL brings instances of the same class closer in the semantic space, promoting fair separation based on misogynous class labels while reducing the impact of identity terms.

2 Related Work

Detection of Misogynous memes. Though most of the existing prior research on misogynous content detection has primarily focused on the unimodal data (mainly on text only), incorporating multimodality (text with image), on the other hand, is still a work in progress (Srivastava, 2022; Chen and Chou, 2022; Rao and Rao, 2022; Zhang and Wang, 2022; Zhou et al., 2022; Zhi et al., 2022; Arango et al., 2022). However, their error analysis has uncovered limitations in the contextual understanding of memes, highlighting the importance of enhancing context comprehension to mitigate bias.

Detection and mitigation of Bias. Despite the importance of detecting and mitigating bias in machine learning models, there has been relatively little research focused on this area (Dixon et al., 2018; Park et al., 2018; Davidson et al., 2019; Sharma et al., 2020b; Aksenov et al., 2021; Xia et al., 2020). Park et al. (2018) proposed masking specific words or phrases to mitigate gender bias in text classification. Similarly, Godoy and Tommasel (2021) introduced an entropy-attention-based approach to reduce bias. Several research studies have also addressed the issue of gender bias in misogyny detection tasks (Nozza et al., 2019; Li et al., 2023b; Nadeem et al., 2021; Arango et al., 2022). Furthermore, Nozza et al. (2019); Hee et al. (2022) identified a lack of contextual understanding in a classification model as a contributing factor to bias, emphasizing the need for improved contextual representation. In this order, scene graphs have also gained significant attention to obtain a better representation of visual modality. Various approaches have been proposed to generate scene graphs from images, videos, and 3D scenes (Li et al., 2021b; Tang et al., 2020; Johnson et al., 2015; Li et al., 2023a; Wang et al., 2019; Liu et al., 2021; Garg et al., 2021; Nag et al., 2023; Dhamo et al., 2021). This enables a more advanced and context-aware visual data analysis, opening up opportunities for diverse applications. While their research did not offer a solution to address the unintended bias issues, our study addresses this gap by proposing a novel technique to mitigate unintended bias in multimodal classifiers, particularly in identifying misogynous classifiers. Our method distinguishes itself from existing debiasing techniques in two key ways. First, it achieves fairer representations, leveraging an unbiased scene graph alongside a memory network and employing contrastive learning instead of data manipulation. Second, our method is jointly trained with the baseline classifier rather than relying on post-processing to remove any identity term information.

3 Dataset

We use MAMI dataset (SemEval2022 Task 5, Subtask A) (Fersini et al., 2022) for conducting all the experiments. Furthermore, to show the generalizability of our bias mitigation technique, we perform experiments on three benchmark meme datasets: Hateful Memes (Kiela et al., 2020b), Memotion2 (Ramamoorthy et al., 2022) and Harmful Memes (Sharma et al., 2022). By considering these diverse datasets (Refer Table 1 for Data Statistics), we aimed to assess the effectiveness of our approach in addressing any unintentional bias across different contexts and content (Refer Appendix Table 5 for class-wise distribution of these datasets.)

Dataset	Train set	Test set	Task
MAMI	10000	1000	Misogynous Meme Detection
Hateful Meme	8500	1000	Hateful Meme Detection
Memotion2	7500	1500	Offensive Meme Detection
Harmful Meme	3013	354	Harmful Meme Detection

Table 1: Dataset Statistics

4 Measuring Unintended Bias

Baldini et al. (2022) discovered that prior research emphasized accuracy over fairness, resulting in inconsistent model fairness outcomes. In domains such as offensiveness detection, it is imperative to consider fairness metrics beyond mere accuracy. So, in order to achieve fairness in our misogynous meme classification model, we adopt the Equality of Odds (EO) principle proposed by Hardt et al. (2016). This principle emphasizes equalizing the false positive rates (FPR) and false negative rates (FNR) across samples containing different identity terms. To measure the extent of unintended bias, we analyze the model's fairness by calculating FPR and FNR for the entire test set, and each subset containing a specific identity term denoted as FPRt and FNRt, respectively. Let us say $T = \{t_1, t_2, ..., t_n\}$ is the list of all the identity terms in the MAMI dataset. Ideally, a fair model exhibits similar values across all terms and approaches to equality of odds (EO) ($FPR = FPRt_i$ and $FNR = FNRt_i$ for all terms $t_i \in T$). Conversely, wide variation among these values indicates a higher degree of unintended bias.

$$\text{False Positive Rate Equality Difference (FPRED)} = \sum_{t \in T} |\text{FPR} - \text{FPR}_t| \quad (1)$$

$$\text{False Negative Rate Equality Difference (FNRED)} = \sum_{t \in T} |\text{FNR} - \text{FNR}_t| \quad (2)$$

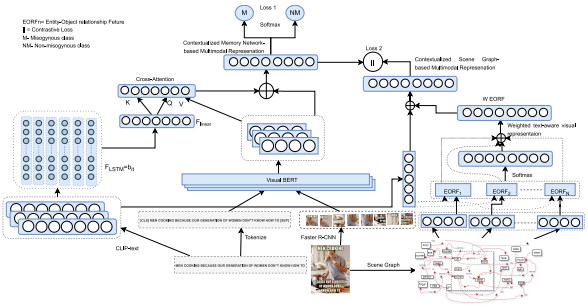

Figure 2: Illustration of our proposed architecture CTXSGMNet for mitigation of unintended biases in misogynous meme classifier.

High FPRED can suppress underrepresented opinions by mislabeling them. In contrast, high FNRED can perpetuate harmful stereotypes by misclassifying content. So, EO is achieved when FPRED = FNRED.

Identity Term List: To mitigate unintended biases, we begin by identifying potentially biased identity words. As shown in Figure 1, these terms significantly impact the final predictions. Our approach involves assessing their importance using post-training SHAP (Qian et al., 2021) values, which provide word-level contributions to false positives (Lundberg and Lee, 2017; Kokalj et al., 2021). We further select potential identity terms by calculating tf-idf word distributions within the misogynous class and ranking them. We ultimately identify the top 50 such terms (Refer Table 2 for such identity terms in the MAMI dataset).

Top frequent words	% Frequency	
	whole data	misogynous class
woman	1.8136	3.0710
girl	0.7565	1.3309
rape	0.1756	0.3440
bitch	0.1599	0.2812
hooker	0.2443	0.1080
kitchen	0.3906	0.4445
man	0.4221	0.5047
sandwich	0.1362	0.2787
dishwasher	0.1396	0.2837
feminist	0.4547	0.8889

Table 2: Frequency of top identity terms in overall train samples and misogynous samples in the MAMI dataset

5 Methodology

5.1 Problem Statement

In this section, we illustrate our proposed **CTXSGMNet** model to measure and mitigate unintended biases in the misogynous meme detection task.

5.2 Task Formulation

Let $\mathcal{D}$ denote the dataset of misogynous memes, with input space $\mathcal{X}$ and output space $\mathcal{Y}$. Each sample in the dataset is represented as $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where x denotes a meme (text and images) and $y \in \{0, 1\}$ indicates its misogyny label ($y = 1$ for misogynous and $y = 0$ for non-misogynous). We aim to train a classifier $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ with parameters θ to minimize the unintended biases by achieving the EO to increase the model's fairness across all the identity terms. The overall workflow of our proposed CTXSGMNet model is shown in Figure 2, and its components are discussed below.

5.3 Contextualized Scene Graph-based Multimodal Network (CTXSGMNet)

It is an end-to-end architecture integrating VisualBERT, CLIP-LSTM-based memory network, and unbiased scene graph analysis for state-of-the-art unintended bias mitigation in memes.

5.3.1 Encoding of meme

A meme m_i comprises text (T_i) and image (I_i). At first, we extract visual patches $v_i = \{v_{i_1}, v_{i_2}, \ldots, v_{i_N}\}$; for $v_{i_j} \in R^N$, where N is the number of regions of I_i via Faster R-CNN (Ren et al., 2017) and tokenized the corresponding meme text T_i into sub-word units and projected into high-dimensional feature vectors $t_i = (t_{i_1}, t_{i_2}, \ldots, t_{i_K})$; for $t_{i_j} \in R^K$, where K is the number of tokens in T_i. These t_i, v_i are then fed into the VisualBERT (Li et al., 2019) module to obtain an encoded multimodal representation M_i of the meme m_i.

$$M_i = VisualBERT(t_i, v_i) \tag{3}$$

Note that VisualBERT being a multimodal model, both t_i and v_i are encoded by it. M_i has a dimension with $1 \times k$, where k refers to the combined length of the input sequence (for both text tokens and image patches and $k > q$).

5.3.2 CLIP-LSTM-based Memory Network

To leverage the sequential and contextual information within the textual part of the meme $T_i = \{w_1, w_2, ..., w_k\}$, we introduce a CLIP-LSTM-based memory network. The motivation behind using a memory network is to enhance the model's ability to capture and retain important information across the text. First, we utilize a CLIP-based pre-trained model, specifically designed for understanding text and images at a semantic level (Radford et al., 2021a), to extract the textual features t_{it} from T_i:

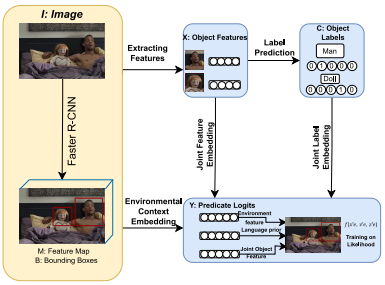

Figure 3: Illustration of the architecture of model used for scene graph.

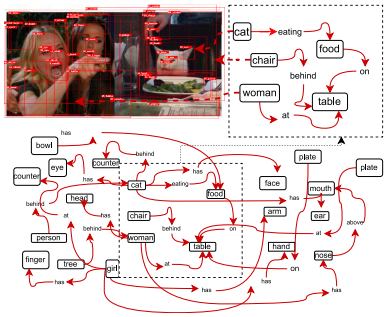

Figure 4: Illustration of scene graph for an image I.

$$t_{it} = \text{CLIP}(T_i) \tag{4}$$

Next, these textual features are fed into a Bidirectional LSTM (BiLSTM) layer (Graves et al., 2005), which serves as the memory component of our network. By considering the sequential dependencies and contextual relationships between words, the BiLSTM enables the model to comprehend the underlying patterns and connections in the text.

$$\overrightarrow{h_{it}} = \text{LSTM}(t_{it}), \quad \overleftarrow{h_{it}} = \text{LSTM}(t_{it}), \quad h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}] \tag{5}$$

To further refine and extract higher-level representations of the textual features, we apply a fully connected network (FCN) layer to the hidden states h_{it} obtained from the BiLSTM layer. The transformation, parameterized by weights W_i and biases b_{it}, allows the model to capture more abstract and comprehensive representations of the textual content, leveraging the memory-like properties of the BiLSTM:

$$b_{it} = \text{FCN}(h_{it}W_i + b_{it}) \tag{6}$$

5.3.3 Contextualized Memory Network-based Multimodal Fusion Module

A multi-headed attention module equipped with cross-attention is used to infuse the encoded meme

(M_i) in Equation 3 and encoded textual knowledge (b_{it}) calculated in Equation 6. By infusing textual awareness captured by the memory network with the multimodal representation, we enable the model to make more informed and contextually grounded predictions by capturing the synergistic effects of textual and multimodal information, enhancing its overall performance and effectiveness. We obtain a memory-enhanced representation of the meme, which is done by calculating the cross-attention (o_i) between M_i and b_{it}, where

$$o_{it} = \text{softmax}(M_i.{b_{it}}^T/sqrt(d_k)).b_{it} \quad (7)$$

Here, M_i acts as the Query (Q), and b_{it} acts as both the Value (V) and Key (K), respectively (c.f. Figure 2). The final representation is obtained by passing o_{it} through a layer-normalization layer (Ba et al., 2016) and then adding it with the query M_i:

$$o_{it} = \text{LayerNorm}(o_i) + M_{it} \quad (8)$$

The output of the cross attention o_{it}, along with the multimodal representation from VisualBERT M_{it}, is concatenated. The concatenated representation combines the complementary information captured through cross-attention and multimodal representation.

$$C_i = \text{concat}(o_{it}, M_{it}) \quad (9)$$

We use a singular feed-forward neural net (FFN) with softmax activation, which takes the concatenated representation (C_i) in Equation 9 as input and outputs class for misogynous meme identification, shown in the following Equation 10:

$$\hat{y}_t = P(Y_i|C_i, W, b) = \text{softmax}(C_i W_i + b_i) \quad (10)$$

The proposed classifier is trained using cross-entropy loss:

$$\mathcal{L}_1 = -\sum[y_t \log \hat{y}_t + (1 - y_t) \log(1 - \hat{y}_t)] \quad (11)$$

5.3.4 Generation of Scene graph

Improving the representation of textual and visual modalities in memes is crucial for bridging the semantic gap and enhancing their analysis. To get a better visual representation, we employ an unbiased scene graph proposed by Krishna et al. (2016). This scene graph model (Refer Figure 3) utilizes Faster RCNN with joint contextual feature embedding to extract unbiased *Entity Object Relationship (EOR)* information (Refer Figure 4) from the visual component of each meme. Given an image I_i of a meme m_i, the scene graph is represented as $T_I \subseteq (E_I * R_I * E_I)$, where T_I is the set of visual triples, E_I represents the entity set, and R_I represents the relation set. Here, $R_I \subseteq R$. Each entity $e_{I,k} = (e_{t,I,k}, A_{I,k}, b_{I,k}) \subseteq E_I$ consists of the entity type $e_{t,I,k} \subseteq E_t$, where E_t represents the set of entity types (c.f. Figure 4). The extraction of entity-object-relation triplets from the scene graph for meme m_i is denoted as EOR_i. EOR_i shows the top k (here k=20) entity object relations (EOR) extracted from the scene graph of the image I_i of a meme m_i as $\text{EOR}_i = (\text{EOR}_i^1, \text{EOR}_i^2, ..., \text{EOR}_i^k)$ By integrating this unbiased visual comprehension into our model, our objective is to attain a fairer visual representation of memes, ultimately enhancing the overall multimodal representation.

5.3.5 Encoding of Scene graph

Further, we utilize a CLIP-based pre-trained model to extract the scene graph features $EORF_i$ from EOR_i:

$$\begin{aligned}\text{EORF}_i = (\text{EORF}_i^1, \text{EORF}_i^2, ..., \text{EORF}_i^k) = \text{CLIP}(\text{EOR}_i) \\ = \text{CLIP}(\text{EOR}_i^1, \text{EOR}_i^2, ..., \text{EOR}_i^k)\end{aligned} \quad (12)$$

5.3.6 Contextualized Scene graph-based Multimodal Fusion Module

To get an unbiased multimodal representation, it is important that these unbiased EOR triplets ($EORF_i$) are aligned with the textual feature (t_{it} (c.f. Equation 4)) of meme m_i in an equitable manner. To establish a robust multimodal representation, we aim to evaluate the alignment between each $EORF_i$ and t_{it}. Within the set of k EORF triplets ($EORF_i$), our aim is to rank those that closely correspond to the textual features t_{it} of meme m_i. This ranking is achieved through a similarity score, which assigns weights to the EORFs based on their relevance to (t_{it}). This approach ensures a fair and meaningful alignment between visual and textual elements, contributing to unbiased multimodal representations. Let ${W_{sim}}_i^l$ denote the resulting weighted similarity scores where $l \in R^k$.

$${W_{sim}}_i^l = \frac{\exp(\text{Cos\_Sim}(\text{EORF}_i^l, t_{it}))}{\sum_{j=1}^{k} \exp(\text{Cos\_Sim}(\text{EORF}_i^j, t_{it}))} \quad (13)$$

Here, $Cos\_Sim(EORF_i^l, t_{it})$ represents the cosine similarity between the EORF triplet $EORF_i^l$ and the textual feature t_{it}, which is calculated as:

$$\text{Cos\_Sim}(\text{EORF}_i^l, t_{it})) = \frac{\text{EORF}_i^l \cdot t_{it}}{|\text{EORF}_i^l| \cdot |t_{it}|} \quad (14)$$

Models	Modality		Metrics			
	T	I	F1↑	FPRED↓	FNRED↓	EO↓
FasterRCNN		✓	64.5	19.54	10.62	8.92
BERT	✓		54.2	32.58	17.42	15.13
LaBSE	✓		59.6	26.45	15.13	11.32
$VGG-19$		✓	51.7	24.19	15.36	8.83
ViT		✓	61.3	22.87	12.81	10.06
$BERT+VGG$	✓	✓	59.9	20.27	10.05	10.23
$BERT+ViT$	✓	✓	63.7	12.35	6.33	6.02
$LXMERT$	✓	✓	65.9	27.18	12.22	14.95
$CLIP$	✓	✓	72.9	17.62	7.49	10.13
$BLIP$	✓	✓	70.8	17.85	10.62	7.23
$ALBEF$	✓	✓	65.9	19.74	**5.74**	13.99
$MisoM^{*}$	✓	✓	**71.4**	**11.78**	7.89	**3.88**
$^{\Phi}DM^{CTXMN}$	✓	✓	77.0	11.59	5.87	5.72
$^{\Phi}DM_{SGN}$	✓	✓	76.98	9.59	4.88	4.71
$^{\Phi}DM^{SCL-}$	✓	✓	73.84	12.07	6.17	5.9
DM^{CTXMN}_{SGN}	✓	✓	**79.59**	**6.98**	**3.63**	**3.35**

Table 3: Results from the debiased model and the various baselines on the MAMI dataset. Here, the bolded values indicate maximum scores. Here,T: Text, I: Image F1 is macro F1-score, FPRED: False Positive Rate Equility Difference, FNRED: False Negative Rate Equility Difference, EO: Equality of Odds. * represents the best-performing baseline model, whereas Φ presents the variants of proposed model. We observe that the performance gains are statistically significant with p-values (<0.0431) using a t-test, which signifies a 95% confidence interval.

Next, we obtain the weighted Entity Object Relations ($WEORF_i^l$) by element-wise multiplication of the EOR triplets ($EORF_i^l$) with their corresponding weights (W_{siml_i}). This operation combines the importance of the EOR triplets based on their similarity to the textual feature.

$$\text{WEORF}_i^l = W_{sim_i}{}^l \cdot \text{EORF}_i^l \tag{15}$$

The resulting weighted EORFs (WEORF_{k_l}) are then at first concatenated with each other resulting $WEORF_i = (\text{WEORF}_i^1, \text{WEORF}_i^2, ..., \text{WEORF}_i^k)$ and finally with the textual representation (t_{it}) (Refer Equation 4), resultant a fairer, more fine-grained and robust multimodal representation, denoted by $C_i^{'}$.

$$C_i^{'} = concat(t_{it}, WEORF_i) \tag{16}$$

5.3.7 Network Training

In addition to cross-entropy loss, we incorporate supervised contrastive loss (SCL) to enhance fair supervised learning and provide empirical evidence of its effectiveness in learning unbiased representations and fair classifiers (Li et al., 2023b; Shen et al., 2021). This loss component encourages well-separated representations for the misogynous meme identification task, creating equitable representations and unbiased predictions. Our context-aware multimodal representations,i.e., (C_i,$C_i^{'}$ in Equation 9 and 16), are assumed to capture similar contexts for a given meme m_i. During training, these representations are aligned within the same semantic space, enabling effective utilization of both the CLIP-LSTM-based memory network and scene-graph modules through contrastive learning.

$$\mathcal{L}_{SCL} = -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{C_i}, \boldsymbol{C'}_i\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k\neq i]} \exp\left(\text{sim}\left(\boldsymbol{C_i}, \boldsymbol{C'}_k\right)/\tau\right)} \tag{17}$$

where N is the batch size, sim() is the cosine similarity, and τ is the temperature to scale the logits. During training time, for the given ith meme sample, $\mathcal{L}_{MM}$ tries to find the similarity between both the modalities of a given meme (T_i, V_i) by minimizing the distance between the context-aware multimodal representation (C_i,$C_i^{'}$).

Now, to minimize the overall loss of the proposed model, $\mathcal{L}_{SCL}$ is combined along with categorical cross-entropy loss defined in Equation 11. The weights (α and β) control the relative importance of each loss.:

$$\mathcal{L}_1^{'} = \alpha \cdot \mathcal{L}_1 + \beta \cdot \mathcal{L}_{SCL} \tag{18}$$

6 Baseline Models

6.1 Unimodal Systems

For the baseline model, we implement BERT(Pires et al., 2019), LaBSE (Feng et al., 2020), VGG-19 (Simonyan and Zisserman, 2015), ViT (Dosovitskiy et al., 2020).

6.2 Multimodal Systems

Early Fusion: For this category, we extracted textual and visual features from different pre-trained models and then applied early fusion to get a multimodal representation. By doing so, we have developed the following baseline models: BERT+VGG, BERT+ViT.

Pre-trained Models: For the pre-trained multimodal system, we used the following pre-trained models to extract the multimodal features: *LXMERT* (Tan and Bansal, 2019), *VisualBERT* , *MMBT: Supervised Multimodal Bitransformers (Kiela et al., 2019)*. Further, we used another three multi-modal feature extractors (*CLIP*(Radford et al., 2021b), *BLIP*(Li et al., 2022), and *ALBEF*(Li et al., 2021a) Each of their features is passed through a projection layer to make the final predictions.

7 Experimental setups

All models, including baselines, were developed using Huggingface Transformer[3] with a default random seed of 42 to ensure applicability. The model is trained for 60 epochs with a batch size of 64 and Adam optimizer (Kingma and Ba, 2015) with a learning rate of 3e-5. The models were trained on a single NVIDIA-GTX-1080Ti with 16-bit mixed precision. During the training of the proposed model, through grid search, we fixed the α and β hyperparameters (Eqn. 18) in the $\mathcal{L}_1^{'}$ as 0.5, and 0.4, respectively.

8 Result Analysis

8.1 Model Results and Comparisons

In this section, we present the results of our comparative analysis, which examines the presence of unintended biases in baseline models, our proposed model, and their respective variations. To measure the unintended biases, we use False Positive Rate Equility Difference (FPRED), False Negative Rate Equility Difference (FNRED), Equality of Odds (EO), and the macro-F1 score (F1) score as the preferred metrics.

Models: $MisoM$: VisualBERT-based baseline model, DM_{SGN}^{CTXMN}: Proposed debiased model with both context-aware memory and scene graph network based module. DM^{CTXMN}: This model is trained solely with a CLIP-LSTM-based memory network, excluding a scene graph, DM_{SGN}: This model is trained solely with a scene graph and lacks a memory network, DM^{SCL^-}: This model is trained with both context-aware memory and scene graph network but without $\mathcal{L}_{SCL}$ loss.

Mitigation of unintended biases: In Table 3, we show the results of baseline models for the misogynous meme identification task. Notably, our VisualBERT-based baseline classifier (*MisoM*) outperforms other baselines with a 73.84% F1 score (indicated by low FPRED, FNRED, and EO values), forming the foundation of our proposed method. Furthermore, it is noteworthy that the multimodal baselines consistently outperform their unimodal counterparts, achieving a substantial 15%-17% increase in F1 score. By applying an in-processing debiasing method, our proposed model DM_{SGN}^{CTXMN} effectively reduces unintended biases and improves fairness in identifying misogynous memes. It achieves notable reductions in FPRED (from 12.35 to 6.98) and FNRED (4.33 to 3.63) when comparing it to MisoM, indicating significant improvements. The equality of odds (EO) deduction with -4.67% further supports the efficacy of our approach in mitigating biases and promoting fair representation. These findings underscore the importance of our method in addressing bias and fostering inclusivity in misogynous meme identification tasks.

Ablation. To test the proposed architecture, we develop multimodal variants of our proposed model, as shown in Table 3. These variants allowed us to evaluate the impact of each debiasing component on our model performance. Comparing the performance of the ablation models, DM_{SGN}^{CTXMN} stands out as the most effective in achieving bias mitigation by -2.37 and -1.36% in terms of EO, respectively. This can be attributed to its utilization of unbiased textual and visual information and its contextual understanding through scene graph integration. Our proposed model, the first of its kind, is pioneering in extracting *unbiased Entity Object Relationship* (EOR) information from the visual component of memes. This information helps identify misogynous content, as it can reveal the unbiased semantic relationships between objects in the meme. Our results indicate that by incorporating a scene graph, the model gains a better understanding of the contextual cues, making it more adept at mitigating biases.

8.2 Detailed Result Analysis

8.2.1 Error Rates

Figures 6 display the false positive and negative rates per identity term for the best-performing baseline *MisoM* and proposed debiased model DM_{SGN}^{CTXMN}. The proposed model DM_{SGN}^{CTXMN} exhibits improved performance uniformity across terms, indicating the effectiveness of the bias-mitigation technique in reducing unintended bias. Although some variations in performance persist, there is still potential for further enhancement. Figure 6 (b) specifically examines the false negative rates per term, revealing that the bias mitigation technique successfully reduces bias in false positives without introducing false negatives for the measured terms.

8.2.2 Result Analysis with Case Study

In Figure 5, we present three randomly chosen memes where *MisoM* misclassified the labels due

[3] https://huggingface.co/docs/transformers/index

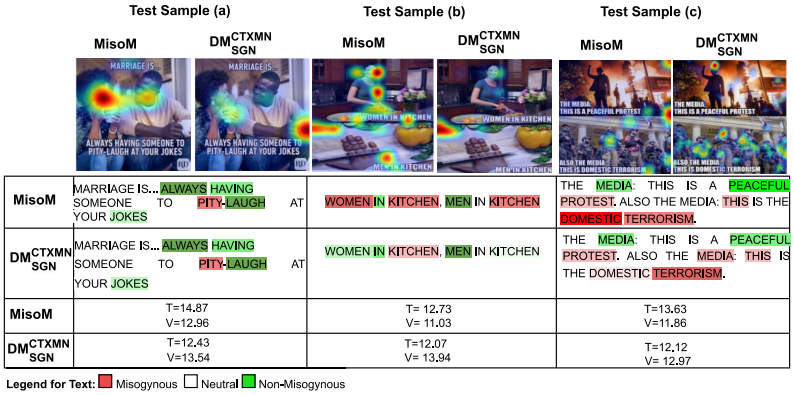

Figure 5: Case studies comparing the attention-maps for the baseline $MisoM$ and the proposed model DM^{CTXMN}_{SGN} using Grad-CAM, LIME (Ribeiro et al., 2016), and Integrated Gradient (Sundararajan et al., 2017) on test samples. Here, T and V are the normalized textual and visual contribution scores in the final prediction using Integrated Gradient.

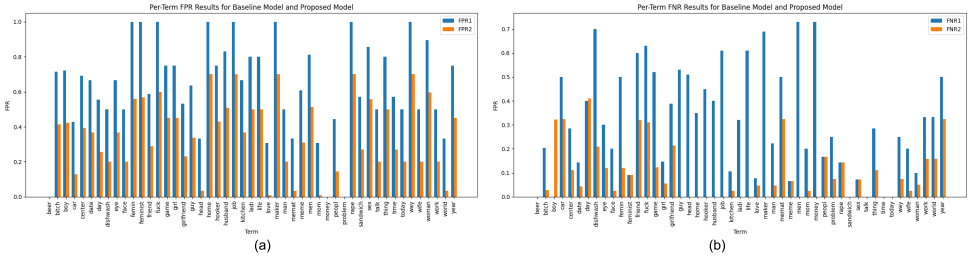

Figure 6: Per-term (a) false positive rates and (b) false negative rates for the MisoM (denoted with blue) and proposed model DM^{CTXMN}_{SGN} (denoted with orange) on the MAMI dataset.

to overgeneralizing identity terms, but our proposed debiased model (DM^{CTXMN}_{SGN}) made correct predictions. These samples have the gold label "non-misogynous." For sample (a), *MisoM* focused on the word "marriage" and the presence of a woman's image, leading to a misogynous prediction. In contrast, our debiased model, DM^{CTXMN}_{SGN}, utilizes fairer contextual CLIP-LSTM-based memory and a scene graph module, enabling a fairer representation of the textual and visual components. As a result, it correctly predicts the meme as non-misogynous. In sample (b), *MisoM* associates both "woman" and "kitchen" with misogynous class prediction. However, DM^{CTXMN}_{SGN} assigns equal importance to both modalities, as shown by the normalized integrated gradient scores, and reduces the influence of the term "woman" in the prediction. Similarly, for sample (c), *MisoM* exhibits bias towards a misogynistic prediction because the word "DOMESTIC" disproportionately influences the textual modality. In contrast, our debiased model, DM^{CTXMN}_{SGN}, captures the sarcastic context related to "Media" and correctly predicts the meme as non-misogynous. These results demonstrate the effectiveness of our proposed model in addressing and mitigating biases, leading to a more balanced and fair representation in the classification process.

	F1↑	FPRED↓	FNRED↓	EO↓
SRCB (Zhang and Wang, 2022)	77.6	15.81	5.41	10.40
MMBT(Kiela et al., 2019)	74.8	18.97	9.29	9.68
DisMultiHate (Lee et al., 2021)	67.24	12.15	4.07	8.07
$^{\Psi}$Wang et al. (2021)	65.91	22.12	13.43	8.69
Momenta (Pramanick et al., 2021b)	72.81	23.78	29.37	5.60
PromptHate (Cao et al., 2022)	**79.98**	10.61	4.83	5.77
DM^{CTXMN}_{SGN}	79.59	**6.99**	**3.63**	**3.36**

Table 4: Comparison of our proposed model with the existing SOTA models, Ψ is the SOTA model on MAMI Dataset

8.3 Comparison with State-of-the-Art (SOTA) Models

Table 4 provides a comprehensive comparison between our proposed model, DM^{CTXMN}_{SGN}, and several existing SOTA models. Notably, in the MAMI task, DM^{CTXMN}_{SGN} outperforms the current SOTA models. It is worth mentioning that PromptHate represents the latest SOTA model for the hateful meme dataset among all the existing models. Although PromptHate achieves high accuracy on the MAMI dataset, it falls short due to its lack of

contextual knowledge, which introduces inherent modality-specific biases, resulting in elevated error rates and a higher EO. Momenta, despite its efforts to leverage augmented image entities, still possesses visually biased representations. In contrast, our DM_{SGN}^{CTXMN} model outperforms most of the SOTA due to its incorporation of unbiased scene graphs with a contextual memory network. Another contributing factor is the improved training through Supervised Contrastive Learning (SCL), which yields well-separated representations for misogynous and non-misogynous classes. Together, these components result in fairer multimodal representations, effectively reducing errors associated with individual identity terms.

8.4 Generalibilty of the proposed architecture

To demonstrate the versatility of our proposed architecture DM_{SGN}^{CTXMN}, we evaluate its performance on three benchmark datasets in English (Hateful Memes, Memotion2, and Harmful dataset). These experiments validate the generalizability of our architecture, not only in misogynous tasks but also in addressing unintentional biases arising from different types of datasets and tasks. In the Appendix Section A, we have done a detailed discussion of the impact of the proposed debiasing method for all the datasets.

9 Error Analysis

Our proposed model, despite its high performance, occasionally makes misclassifications. We categorize them as follows:

- **Contextual challenges due to wrong visual attention:** Our model sometimes struggles to classify the misogynous class due to incorrect visual cues. This can be attributed to the complex nature of contextual nuances, cultural references, and subtle expressions found in memes (Refer to Appendix Figure 7 (a)).
- **Linguistic ambiguity:** Some memes contain textual elements with linguistic constructs or wordplay that introduce ambiguity in their interpretation and hinder the model's ability to discern the intended meaning accurately. (See Appendix Figure 7 (b)).
- **Model overcompensation by hallucinations:** Certain situations where the model overcompensates and misclassifies misogynous memes as non-misogynous due to cautious predictions (See Appendix Figure 7 (c))

10 Conclusion

In summary, in this work, we introduce *CTXSGMNet*, an advanced framework for mitigating unintended bias in misogynous meme identification tasks. CTXSGMNet achieves state-of-the-art performance in addressing unintended bias and promoting inclusivity in meme classification by leveraging contextualized text modeling, scene graph analysis, and multimodal fusion. The empirical evaluation of diverse datasets demonstrates its effectiveness in capturing complex relationships between textual and visual components. This research contributes to combating unintended bias and fostering a more equitable digital environment while inspiring future research in bias mitigation.

Limitations

In section 9, we discussed the limitations of our proposed work. Our baseline models face challenges in detecting subtle or implicit elements in memes, particularly context-dependent and culturally referenced misogynous memes. Analyzing these errors provides valuable insights into the limitations and challenges of our model, guiding future improvements in unintended bias mitigation for the identification of misogynous memes.

Ethics Statement

Broader Impact: The broader impact of this work lies in its potential to mitigate bias in meme analysis and classification. We promote fairness and inclusivity in online content moderation by developing techniques to address bias in detecting misogynous memes. **Intended Use** Our research is presented to encourage research into studying the detection and mitigation of bias from a classifier. We believe that it has the potential to positively impact the experiences of social media users, content moderators, and the overall online community when used appropriately. **Misuse Potential** Identity words contain many vulgar words. It is to be noted that we have used those keywords only for subsequent understanding of the dataset, and it is not in our intention to harm an individual or any group.

Acknowledgements

The research reported in this paper is an outcome of the project "**HELIOS: Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System,**" sponsored by Wipro AI Labs, India.

References

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in German news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131, Online. Association for Computational Linguistics.

Ayme Arango, Jesus Perez-Martin, and Arniel Labrada. 2022. HateU at SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 581–584, Seattle, United States. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. 2022. MilaNLP at SemEval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 654–662, Seattle, United States. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland. Association for Computational Linguistics.

Dibyanayan Bandyopadhyay, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and Vinutha BN. 2023. A knowledge infusion based multitasking system for sarcasm detection in meme. In *Advances in Information Retrieval*, pages 101–117, Cham. Springer Nature Switzerland.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lei Chen 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)*, pages 636–641, Seattle, United States. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *CoRR*, abs/1905.12516.

Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. 2021. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16352–16361.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

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

Sarthak Garg, Helisa Dhamo, Azade Farshad, Sabrina Musatian, Nassir Navab, and Federico Tombari. 2021. Unconditional scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16362–16371.

Daniela Godoy and Antonela Tommasel. 2021. Is my model biased? exploring unintended bias in misogyny detection tasks. In *AIofAI 2021: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies*, volume 2942 of *CEUR Workshop Proceedings*, pages 97–11, Montreal, Canada.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. pages 799–804.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the*

Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 32–39, Online. Association for Computational Linguistics.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020a. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 26–35, Online. Association for Computational Linguistics.

Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21, Online. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332.

Gitanjali Kumari, Amitava Das, and Asif Ekbal. 2021. Co-attention based multimodal factorized bilinear pooling for Internet memes analysis. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 261–270, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5138–5147, New York, NY, USA. Association for Computing Machinery.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation.

Lin Li, Jun Xiao, Hanrong Shi, Wenxiao Wang, Jian Shao, An-An Liu, Yi Yang, and Long Chen. 2023a. Label semantic knowledge distillation for unbiased scene graph generation. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. 2021b. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11109–11119.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023b. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14254–14267, Toronto, Canada. Association for Computational Linguistics.

Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. 2021. Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11546–11556.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K. Roy-Chowdhury. 2023. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22803–22813.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

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

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, Online. Association for Computational Linguistics.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual models from natural language supervision.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, Suryavardan S, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes.

Ailneni Rakshitha Rao and Arjun Rao. 2022. ASRtrans at SemEval-2022 task 5: Transformer-based models for meme classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 597–604, Seattle, United States. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020a. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

Vidish Sharma, Aditya Bendapudi, Tarun Trehan, Ashutosh Sharma, and Adwitiya Sinha. 2020b. Analysing political bias in social media. In *2020*

Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH), pages 241–246.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Harshvardhan Srivastava. 2022. Poirot at SemEval-2022 task 5: Leveraging graph network for misogynistic meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 793–801, Seattle, United States. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

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

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. *CoRR*, abs/2002.11949.

Jialu Wang, Yang Liu, and Xin Wang. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1995–2008, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2019. Exploring context and visual pattern of relationship for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mengzhou Xia, 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. Association for Computational Linguistics.

Jing Zhang and Yujin Wang. 2022. SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, Seattle, United States. Association for Computational Linguistics.

Jin Zhi, Zhou Mengyuan, Mengfei Yuan, Dou Hu, Xiyang Du, Lianxin Jiang, Yang Mo, and XiaoFeng Shi. 2022. PAIC at SemEval-2022 task 5: Multimodal misogynous detection in MEMES with multi-task learning and multi-model fusion. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 555–562, Seattle, United States. Association for Computational Linguistics.

Ziming Zhou, 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 (SemEval-2022)*, pages 563–570, Seattle, United States. Association for Computational Linguistics.

A Result analysis on Hateful meme, Memotion2 and Harmful meme dataset

A.1 Results on Hateful meme dataset

The evaluation results on the Hateful meme dataset, as outlined in Table 7, provide valuable insights into the performance of our novel DM^{CTXMN}_{SGN} model in comparison to various baseline and SOTA models. It's important to note that the Hateful meme dataset is distinctively created through the synthetic generation of hateful memes, which replaces specific keywords and images within template-based memes. Despite its synthetic nature, this dataset is specifically designed for a highly task-specific challenge, similar to the MAMI dataset. Hence, we obtained remarkably similar results when applying the Hateful meme dataset for mitigating unintended bias in the context of the hateful classification task, as observed in the MAMI dataset. In our analysis, while PromptHate demonstrates higher accuracy on this dataset, a deeper examination reveals the superiority of our DM^{CTXMN}_{SGN} model in effectively mitigating unintended biases due to the identity terms (Refer to Ablation Table 10 for such terms in this dataset.). It's also worth noting that even though PromptHate gets higher accuracy, the primary objective of our DM^{CTXMN}_{SGN} model is to provide a more equitable and fair classification, prioritizing the mitigation

of biases introduced by identity terms. This differentiation in focus is reflected in the substantial improvements observed in the FPERD, FNERD, and EO matrices, which are essential for the unbiased classification of hateful memes.

Dataset	Split	Label	#Memes
MAMI	Train	Misogynous	5000
		Non-Misogynous	5000
	Test	Misogynous	500
		Non-Misogynous	500
Memotion2	Train	Offensive	1933
		Non-Offensive	5567
	Test	Offensive	557
		Non-Offensive	943
Hateful meme	Train	Offensive	3050
		Non-Offensive	5450
	Test	Offensive	500
		Non-Offensive	500
Harmful meme	Train	Harmful	1,064
		Non-harmful	1,949
	Test	Harmful	124
		Non-Harmful	230

Table 5: Classwise distribution of (MAMI, Memotion2, Hateful meme, and HarmMeme dataset) distribution in Train Set and Test Set

A.2 Results on Memotion2 dataset

In Table 6, we outlined the results obtained from our proposed model, DM_{SGN}^{CTXMN}, alongside various baseline models and state-of-the-art (SOTA) models on the Memotion2 dataset. This dataset (we primarily focused on Task 2, i.e., the Offensive meme detection task only) offers a unique challenge due to its diversity and generic nature. It's important to note that while our model effectively reduces unintended biases when compared to baseline models, there are still variations in the results. This variation can be attributed to the dataset's diverse and non-patterned nature, which reflects real-world memes that often deviate from any specific trend. Unlike the other datasets, which are created for dedicated single tasks, Memotion2 is created for multiple correlated tasks, making its patterns more generic and wide-ranging. Consequently, our model encounters a broader spectrum of meme content, which can lead to a wider range of results. In Table 6, it's noteworthy that our DM_{SGN}^{CTXMN} model outperforms most baseline and SOTA models in terms of F1 score, FPRED, FNRED, and EO. This underlines the efficacy of our model in mitigating biases, even in the face of the dataset's inherent diversity. Overall, our model showcases its ability to handle diverse and real-world meme content effectively, providing valuable contributions to mitigating unintended biases in this context.

A.3 Results on Harmful meme dataset

Similarly, in Table 8, we present the results from our proposed model, DM_{SGN}^{CTXMN}, alongside various baseline and state-of-the-art (SOTA) models on the harmful meme dataset. Much like the MAMI and Hateful meme data results, we encounter similar trends here. Since this dataset primarily focuses on the COVID-19 domain, most state-of-the-art models exhibit unintended bias due to the identity terms.

Models	Modality		Metrics			
	Text	Image	F1	FPRD	FNRD	EO
$FasterRCNN$		✓	48.9	29.7	13.6	16.1
BERT	✓		50.01	34.7	16.3	18.4
ViT		✓	51.17	22.4	12.3	10.1
Late-Fusion	✓	✓	51.4	25.6	10.1	15.5
$BERT + ViT$	✓	✓	51.9	22.4	10.2	12.2
$UNITER$	✓	✓	52.7	18.9	10.5	8.4
$LXMERT$	✓	✓	52.3	28.1	13.4	14.7
$MMBT$	✓	✓	52.1	26.4	19.6	6.8
$CLIP$	✓	✓	48.4	24.8	17.8	7.0
$ALBEF$	✓	✓	50.8	25.7	16.8	8.9
$^{\Psi}$Ramamoorthy et al. (2022)	✓	✓	55.17	18.3	7.2	11.1
MisoM	✓	✓	51.06	14.6	10.8	3.8
DisMultiHate	✓	✓	50.57	18.2	14.5	3.7
Momenta	✓	✓	50.9	18.1	9.8	8.3
PromptHate	✓	✓	50.89	19.4	12.5	6.9
DM_{SGN}^{CTXMN}	✓	✓	**56.73**	**17.9**	14.7	**3.2**

Table 6: Results from the debiased model, various baselines, and SOTA on the Memotion2 dataset, Here, the bolded values indicate maximum scores. Here, F1 is macro F1-score, FPRED: False Positive Rate Equility Difference, FNRED: False Negative Rate Equility Difference, EO: Equality of Odds. We observe that the performance gains are statistically significant with p-values ($<$0.05) using a t-test, which signifies a 95% confidence interval. $^{\Psi}$ is the SOTA model on Memotion2 Dataset

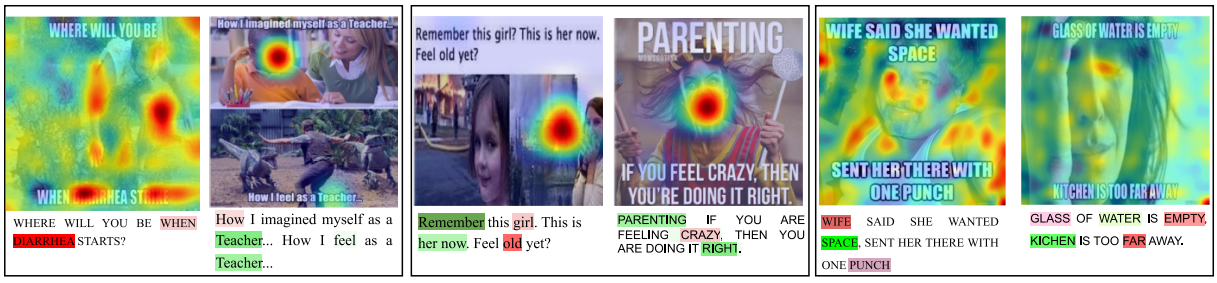

Figure 7: **Error Analysis:** Test cases where proposed multimodal model (DM_{SGN}^{CTXMN}) fails

.

Model	Modality		Metrics					
	Text	Image	F1↑	Acc↑	AUROC↑	FPRED	FNRED	EO
FasterRCNN		✓	38.81	58.20	59.97	42.6	19.9	22.70
BERT	✓		58.41	65.80	67.92	32.5	17.4	15.10
Late-Fusion	✓	✓	64.40	57.55	72.51	25.5	12.8	12.70
UNITER	✓	✓	61.66	60.6	60.02	19.5	10.6	8.90
LXMERT	✓	✓	69.45	70.6	76.15	27.2	12.2	15.00
MMBT	✓	✓	58.29	69.80	76.77	19.0	9.3	9.70
CLIP	✓	✓	53.22	70.40	75.98	17.6	7.5	10.10
ViLBERT	✓	✓	52.60	70.80	76.32	17.9	10.6	7.30
MisoM	✓	✓	67.46	69.6	74.63	10.9	5.7	5.20
DisMultiHate	✓	✓	66.71	68.6	73.43	12.4	4.3	8.10
Momenta	✓	✓	66.71	68.6	73.43	15.6	9.43	6.17
PromptHate	✓	✓	**71.22**	**72.60**	77.07	10.1	4.9	5.20
DM_{SGN}^{CTXMN}	✓	✓	68.93	71.0	74.89	8.5	3.4	**5.10**

Table 7: Results from the debiased model, various baselines, and SOTA on the Hateful meme dataset, Here, the bolded values indicate maximum scores. Here, F1 is macro F1-score, FPRED: False Positive Rate Equility Difference, FNRED: False Negative Rate Equility Difference, EO: Equality of Odds. We observe that the performance gains are statistically significant with p-values (<0.05) using a t-test, which signifies a 95% confidence interval.

Models	Modality		Metrics			
	Text	Image	F1	FPRD	FNRD	EO
FasterRCNN		✓	65.9	22.69	7.05	15.64
BERT	✓		77.92	25.79	11.93	13.32
ViT		✓	67.88	18.88	10.35	8.53
Late Fusion	✓	✓	78.50	22.07	11.83	10.24
MMBT	✓	✓	80.2	21.69	11.69	10.00
Visual BERT COCO	✓	✓	86.1	12.87	6.70	6.17
CLIP	✓	✓	82.9	31.07	15.98	15.09
ALBEF	✓	✓	87.5	17.8	10.9	6.9
MisoM	✓	✓	85.83	16.78	10.62	6.16
DisMultiHate	✓	✓	84.57	18.69	12.9	5.81
Momenta	✓	✓	88.3	16.12	11.19	4.93
PromptHate	✓	✓	**89.0**	14.59	9.83	**4.76**
DM_{SGN}^{CTXMN}	✓	✓	88.76	**12.59**	**9.8**	4.8

Table 8: Results from the debiased model, various baselines, and SOTA on the Harmful meme dataset, Here, the bolded values indicate maximum scores. Here, F1 is macro F1-score, FPRED: False Positive Rate Equility Difference, FNRED: False Negative Rate Equility Difference, EO: Equality of Odds. We observe that the performance gains are statistically significant with p-values (<0.05) using a t-test, which signifies a 95% confidence interval.

Top frequent words	% Frequency	
	whole data	offensive class
man	0.1377	0.1944
mom	0.1871	0.1604
friends	0.1166	0.1203
meme	0.1085	0.1234
parents	0.0753	0.0987
shit	0.0866	0.0987
woman	0.0388	0.0678
twitter	0.0599	0.0678
son	0.0445	0.0555
fuck	0.0656	0.0704

Table 9: Frequency of top textual attributes in overall train samples and offensive samples in the Memotion2 dataset

Top frequent words	% Frequency	
	whole data	offensive class
people	0.5690	0.6827
women	0.1391	0.2372
men	0.1212	0.1909
muslims	0.1964	0.3558
girl	0.1223	0.1475
dishwasher	0.1021	0.1243
fuck	0.2155	0.2487
wife	0.0718	0.0723
religion	0.0583	0.0838
children	0.0628	0.0723

Table 10: Frequency of top textual attributes in overall train samples and offensive samples in the Hateful meme dataset