TFCD: Towards Multi-modal Sarcasm Detection via Training-Free Counterfactual Debiasing

Zhihong Zhu[1], **Xianwei Zhuang**[1], **Yunyan Zhang**[2], **Derong Xu**[3], **Guimin Hu**[4], **Xian Wu**[2*], **Yefeng Zheng**[2]
[1]Peking University
[2]Jarvis Research Center, Tencent YouTu Lab
[3]University of Science and Technology of China
[4]University of Copenhagen
{zhihongzhu, xwzhuang}@stu.pku.edu.cn, derongxu@mail.ustc.edu.cn
guhu@di.ku.dk, {yunyanzhang, kevinxwu, yefengzheng}@tencent.com

Abstract

Multi-modal sarcasm detection (MSD), which aims to identify whether a given sample with multi-modal information (*i.e.*, text and image) is sarcastic, has garnered widespread attention. Recent approaches focus on designing sophisticated architectures or mechanisms to extract sarcastic cues from entire or local image and text features. Nevertheless, a long-overlooked issue is that current MSD task invariably suffers from unintended dataset biases, especially the *statistical label bias* and *sarcasmless word bias*. Concretely, such harmful biases are confounders that may mislead existing models to learn spurious correlations, significantly limiting models' performance. To tackle this issue, this paper proposes a **T**raining-**F**ree **C**ounterfactual **D**ebiasing framework TFCD, which first formulates the causalities among variables in MSD via a tailored causal graph. Then, TFCD extracts biases from the conventionally-trained model by generating counterfactual utterances and contexts and mitigates them using element-wise subtraction. Extensive experiments on two benchmarks demonstrate the effectiveness of the proposed TFCD. Remarkably, TFCD requires neither data balancing nor model modifications, and thus can be seamlessly integrated into diverse state-of-the-art approaches and achieve considerable improvement margins.

1 Introduction

Due to the rise of social media platforms such as X and Facebook, multi-modal sarcasm detection (MSD) has garnered substantial research interest in recent years. Formally, MSD aims to recognize the sarcastic sentiment in multi-modal social posts [Cai *et al.*, 2019], which typically refer to textual paragraphs accompanying images. Unlike the traditional sarcasm detection which primarily focuses on textual cues [Riloff *et al.*, 2013; Poria *et al.*, 2016; Zhang *et al.*, 2016], the key objective of MSD is to effectively identify inter- and intra-modal inconsistencies in the expression of sentiment within a coupled image text pair.

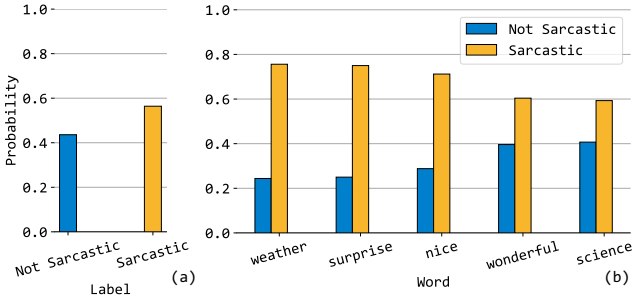

Figure 1: The probability distribution of (a) different labels and (b) sarcasmless words from one current MSD [Cai *et al.*, 2019] *training* dataset, confirms various biases in the MSD task.

Towards this goal, a series of models have been proposed for MSD. Therein, early approaches concentrated on leveraging entire image and text features for incongruity learning [Pan *et al.*, 2020; Xu *et al.*, 2020]. However, these methods neglected the fact that sarcastic information might be expressed in specific segments of the text and particular regions of the image. Motivated by this, recent studies tried to employ graph neural networks to explore the local semantic relationships within textual and visual modalities [Liang *et al.*, 2021; Liang *et al.*, 2022]. More recently, [Qin *et al.*, 2023] employed the powerful pre-trained CLIP model [Radford *et al.*, 2021] to detect different sarcasm cues captured from multiple perspectives, achieving state-of-the-art results.

Despite the significant advancements achieved by existing MSD works, we argue that they remain susceptible to capturing harmful dataset biases. These biases can mislead the models, leading to inaccurate predictions. Based on our observations, we group these biases into two main types: (1) *Statistical label bias*. As depicted in Figure 1(a), the MSD training set is dominated by the sarcastic class, which comprises 56.4%, compared to 43.6% for the non-sarcastic class. Numerous studies [Lin *et al.*, 2022; Pan *et al.*, 2023] have demonstrated that models trained on such imbalanced data

* Corresponding Author

Debiasing strategies	Data-manipulation-free	Model-balancing-free
Data-level debiasing [Qian *et al.*, 2020; Wang and Culotta, 2021; Jia *et al.*, 2024]	✗	✓
Model-level debiasing [Zhang *et al.*, 2020; Sun *et al.*, 2022; Chen *et al.*, 2023]	✓	✗
TFCD (Ours)	✓	✓

Table 1: Method comparison.

tend to be biased towards predicting the majority class. (2) *Sarcasmless word bias*, which refers to the strong associations that may develop between sarcasmless specific words and the sarcastic class. As shown in Figure 1(b), the word '`weather`' is highly likely to be correlated with the sarcastic class. In this context, the conventionally-trained models tend to unfairly categorize utterances containing these specific words, relying on biased statistical information rather than intrinsic semantics [Zhu *et al.*, 2022; Zhou *et al.*, 2023].

To mitigate biases, various debiasing strategies have been proposed, which can be mainly categorized into two types (*cf.* Table 1): (1) *Data-level* debiasing strategies (*e.g.*, resampling [Qian *et al.*, 2020] and generating counterfactual samples [Wang and Culotta, 2021]), aim to balance the training set but are often limited by the additional data handling [Zhang *et al.*, 2020] and extra training time. (2) *Model-level* debiasing ones (*e.g.*, reweighting [Zhang *et al.*, 2020] and counterfactual reasoning [Sun *et al.*, 2022; Chen *et al.*, 2023]), adjust category influence during training but require careful strategy selection and retraining from scratch.

Unlike current debiasing works that rely on data balancing manipulations or modifying then retraining the model, we propose a **T**raining-**F**ree **C**ounterfactual **D**ebiasing framework termed TFCD for MSD. Concretely, we first formulate the procedure of the MSD task via a proposed causal graph. Then, we employ a masking mechanism [Qian *et al.*, 2021] to generate counterfactual examples from the original input data to identify detrimental biases in the trained models. In this case, the biases (including *statistical label bias* and *sarcasmless word bias*) are essentially unintended confounders that mislead the models to learn the spurious correlations, and need to be eliminated. We mitigate these two extracted biases using conceptually simple yet empirically robust element-wise subtraction operations on prediction distributions, without any retraining. We comprehensively evaluate the effectiveness and superiority of the proposed TFCD on two MSD benchmarks. Extensive experiments and analyses demonstrate that TFCD can significantly and consistently improve existing baselines, achieving a new state-of-the-art.

Overall, the main contribution of this paper is three-fold:

- To our best knowledge, this is the first work to investigate debiasing in MSD. We identify that certain biases (*i.e.*, *statistical labels bias* and *sarcasmless word bias*) act as confounders, misleading models to learn the spurious correlations for prediction.
- TFCD, a training-free counterfactual debiasing framework for MSD that mitigates *statistical label bias* and *sarcasmless word bias*, to facilitate a fair contribution of diverse samples and contexts to sarcasm detection.
- Extensive experiments verify that the proposed TFCD can facilitate existing models to achieve unbiased predictions. More encouragingly, TFCD is model-agnostic and can be seamlessly integrated into any existing MSD approach to boost baseline performance.

2 Related Work

Multi-modal Sarcasm Detection. With the rapid popularization of social media platforms, multi-modal sarcasm detection (MSD) has garnered increasing research attention in recent years [Zhu *et al.*, 2024a]. [Schifanella *et al.*, 2016] first used both textual and visual information to tackle the MSD task. [Cai *et al.*, 2019] created an MSD benchmark based on *Twitter* and proposed a hierarchical fusion model. Thereafter, [Xu *et al.*, 2020] and [Pan *et al.*, 2020] captured both intra-modality and inter-modality incongruities based on global textual and visual features, respectively. [Liang *et al.*, 2022] and [Liang *et al.*, 2021] utilized cross-modal graph-based models [Zhu *et al.*, 2024b] for drawing incongruous relations across local multi-modal features. Most recently, [Qin *et al.*, 2023] leveraged the power of pre-trained CLIP [Radford *et al.*, 2021] to perform MSD and achieved cutting-edge results. Despite these advancements, unintended dataset biases have been neglected in previous MSD studies, which limits the performance of MSD models.

Debiasing Strategy. To address dataset bias, several debiasing strategies are proposed to enhance the robustness and reasoning ability of models [Xin *et al.*, 2023], which can be roughly divided into two groups: (1) *Data-level* debiasing strategies include data balancing, data resampling as well as data augmentation [Qian *et al.*, 2020; Wang and Culotta, 2021]. (2) *Model-level* debiasing strategies include using unbiased embeddings [Sun *et al.*, 2022], adjusting thresholds [Kang *et al.*, 2019], and reweighting [Zhang *et al.*, 2020] techniques. However, the former results in additional manual costs for data manipulations. On the other hand, the latter necessitates a meticulous selection of balancing techniques and retraining. Within the multi-modal research community, various debiasing efforts have emerged on different tasks such as multi-modal fake news detection [Chen *et al.*, 2023], and multi-modal sentiment analysis [Sun *et al.*, 2022]. Unfortunately, their methods still necessitate retraining. There is contemporaneous work [Jia *et al.*, 2024; Yang *et al.*, 2024a; Yang *et al.*, 2024b] that also proposed debiasing methods for MSA and MSD. Besides, several training-free methods have been proposed [Qian *et al.*, 2021; Tu *et al.*, 2023] and only perform debiasing on textual-modality data. In this work, we extend the existing training-free debiasing scheme to multi-modal setting and make the first attempt to achieve a training-free debiasing strategy for MSD, which has superiority over complex retraining modules employed in previous approaches.

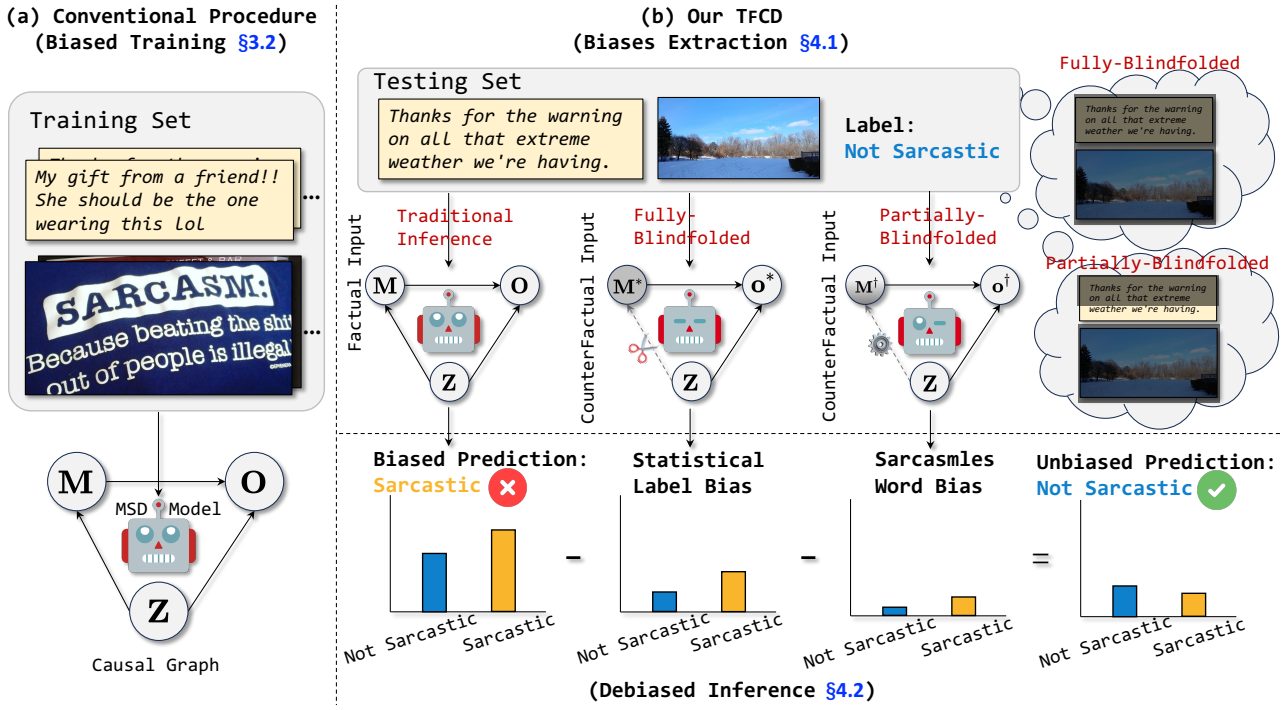

Figure 2: (a) Current MSD models follow the conventional procedure of biased training (§3.2). (b) The architecture of our proposed TFCD, which contains biases extraction (§4.1) and debiased inference (§4.2) on the same trained models.

3 Preliminaries

3.1 Task Formulation

Given a sample $\mathbf{x}_i$ from the training set, the objective of MSD task is to determine whether the sample implies any sarcasm by learning a model $f(\cdot)$ using the text $\mathbf{T}_i$ and corresponding image v_i. This conventional training procedure is represented as $\hat{y}_i = f(\mathbf{T}_i, v_i \,|\, \Theta) \in \{0, 1\}$, where $\hat{y}_i = 1$ indicates the sample is sarcastic and vice versa; Θ represents the learnable model parameters. For simplicity, we temporally omit the superscript i that indexes the training samples in §3.2 and §4.1.

3.2 Biased Training

Let $\mathbf{H}_t$ and $\mathbf{H}_v$ denote the encoded representations of the textual modality (T) and visual modality (V), respectively. As mentioned in §1, existing MSD models focus on designing sophisticated architecture or mechanisms to extract sarcastic cues from *entire* or *local* text and image features:

$$\mathbf{M} = f_m(T = \mathbf{H}_t, V = \mathbf{H}_v), \tag{1}$$

where $\mathbf{M}$ denotes the fused multi-modal feature and $f_m(\cdot)$ denotes the fusion strategy that depends on a certain MSD model. Then, we use feedforward propagation to predict examples and backward propagation to update the learnable parameters of the model in an end-to-end fashion as shown in Figure 2(a). Following previous works, we employ the cross-entropy loss to optimize the MSD task:

$$\mathcal{L}(\Theta) = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}). \tag{2}$$

4 Training-Free Counterfactual Debiasing

This section elaborates on the details of our proposed TFCD framework, whose architecture is shown in Figure 2(b).

4.1 Biases Extraction

During inference, the conventionally-trained models in §3.2 make predictions via the feedforward propagation to obtain the probability distribution. However, the prediction is easily affected by unplanned confounders [Pearl and Mackenzie, 2018], which may produce *statistical label bias* and *sarcasmless word bias*. Aiming to obtain unbiased prediction, **our objective** is to debias only during inference by blocking the spread of biases from training. Towards this goal, we propose training-free counterfactual debiasing to extract the two biases captured by the trained models and mitigate them.

In the following, we first formulate a tailored causal graph for MSD. Then, we elaborate on how to extract the two biases from the trained model based on the constructed graph.

MSD Causal Graph. The causal graph [Pearl *et al.*, 2016] is a probabilistic graphical model used to describe how variables interact with each other, expressed by a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ consisting of the sets of variables $\mathcal{N}$ and the cause-and-effect correlations $\mathcal{E}$ between two nodes (variables). For instance, $\mathbf{X} \rightarrow \mathbf{Y}$ indicates that $\mathbf{X}$ is the cause of the effect $\mathbf{Y}$, meaning that the value of $\mathbf{Y}$ is influenced by $\mathbf{X}$.

As illustrated in Figure 3(a), there are four variables involved in the MSD causal graph, which are textual feature $\mathbf{H}_t$, visual feature $\mathbf{H}_v$, multi-modal feature $\mathbf{M}$, and predictions $\mathbf{o}$. As such, we can obtain the factual casual graph for §3.2. Note that our causal graph applies to a variety of MSD

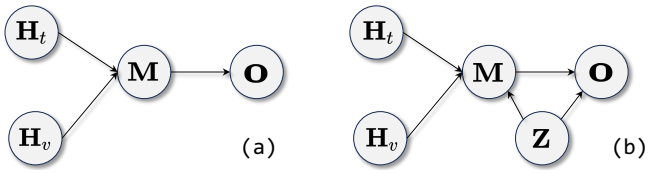

Figure 3: (a) Causal graph for factual MSD. (b) Causal graph with confounders $\mathbf{Z}$ for counterfactual MSD.

methods, since it is highly general, imposing no constraints on the detailed implementations.

Statistical Label Bias. As illustrated in Figure 3(b), there exist unplanned confounders $\mathbf{Z}$, which is the cause of the spurious correlations between the model input $\mathbf{H}_{\{t,v\}}$ and predictive logits $\mathbf{o}$. Such confounders may occur due to the unbalanced label distribution (*i.e.*, '`Sarcastic`' dominates the training data over "`Not Sarcastic`"). Therefore, the trained models tend to unfairly assign outputs to '`Sarcastic`' category based on biased statistical information, which is not unreliable for MSD. To decouple the spurious correlation, we use the backdoor adjustments [Pearl *et al.*, 2016] with do-calculus operation to calculate the corresponding intervention distribution:

$$\begin{aligned} \mathbf{P}(\mathbf{o}^*|\,\mathrm{do}(\mathbf{M})) = \mathbf{P}(\mathbf{o}^*|\mathbf{M}^*) = f(\mathbf{M}^*), \\ \mathbf{M}^* = f_m(T = \mathbf{H}_t^*, V = \mathbf{H}_v^*), \end{aligned} \tag{3}$$

where $\mathbf{M}^*$ can take any form as long as they are no longer influenced by $\mathbf{Z}$, effectively breaking the connection between $\mathbf{H}_{\{t,v\}}$ and $\mathbf{o}$. To extract statistical label biases, in the causal intervention operation, $\mathbf{M}^*$ is intervened under no-treatment condition where T and V have not been accessible. Note that neural models cannot deal with no-treatment conditions where the inputs are void. Therefore, we assume that the model will randomly guess with equal probability (*i.e.*, uniform distribution assumption) [Niu *et al.*, 2021] to guarantee a safe estimation. As MSD models cannot observe any words and images after the intervention, the predictive logits $\mathbf{o}^*$ only reflect the adverse impact of statistical label bias.

Sarcasmless Word Bias. We further utilize a partially-blindfolded counterfactual input where some words are masked to distill the sarcasmless word bias from the trained model. Specifically, our goal is to retain only *sarcasmless words* (*e.g.*, the spurious '`weather`'-to-'`sarcastic`' mapping) in the utterance to measure their potentially negative influence. We first utilize pysentiment library[1] to construct a main-content words dictionary $\mathcal{D}_s$. By masking *main-content words*, we deliberately expose any spurious correlations between sarcastic categories and *sarcasmless words*:

$$\begin{aligned} \mathbf{P}(\mathbf{o}^\dagger|\,\mathrm{do}(\mathbf{M})) = \mathbf{P}(\mathbf{o}^\dagger|\mathbf{M}^\dagger) = f(\mathbf{M}^\dagger), \\ \mathbf{M}^\dagger = f_m(T = \mathbf{H}_t^\dagger, V = \mathbf{H}_v^\dagger). \end{aligned} \tag{4}$$

Here, $\mathbf{H}_t^\dagger$ denotes counterfactual word embedding where the *main-content words* are masked. To be specific, the mask operation process can be formulated as follows:

$$\mathbf{H}_t^\dagger = \{\mathbf{h}_{t,1}, [mask], \cdots, \mathbf{h}_{t,n}\}, \forall\, \mathbf{h}_{t,j} \leftarrow [mask] \in \mathcal{D}_s, \tag{5}$$

where $[mask]$ is used to hide a single token belonging to $\mathcal{D}_s$ in the input sentence $\mathbf{H}_t$. Meanwhile, $\mathbf{H}_v^\dagger$ should similarly represent an unseen feature. In this work, we use all-zero vectors to initialize it [Sun *et al.*, 2022]. In this case, the predictive logits $\mathbf{o}^\dagger$ reflect the pure influence of *sarcasmless words* to the trained biased model.

[1]https://pypi.python.org/pypi/pysentiment

MMSD/MMSD2.0	Train	Validation	Test
Sentences	19,816/19,816	2,410/2,410	2,409/2,409
Positive	8,642/9,572	959/1,042	959/1,037
Negative	11,174/10,240	1,451/1,368	1,450/1,372

Table 2: Statistics of two experimental datasets.

4.2 Debiased Inference

After obtaining two biases, the final objective is to leverage the direct effect from multi-modal representation $\mathbf{M}_i$ to predictive logits $\mathbf{o}_i$ for each sample i to facilitate unbiased prediction. This process can be formalized using the conceptually simple yet empirically robust element-wise subtraction:

$$\hat{\mathbf{o}_i} = \mathbf{o}_i - \alpha\mathbf{o}_i^* - \beta\mathbf{o}_i^\dagger, \tag{6}$$

where α and β are two hyper-parameters. Note that we ascertain two adjustable scaling factors which are optimized to the model's performance on the validation set, since different biases have diverse impacts on the final prediction.

5 Experiments

5.1 Experiment Setup

Datasets. We conduct experiments on two datasets: MMSD [Cai *et al.*, 2019] and MMSD2.0 [Qin *et al.*, 2023]. MMSD is derived from English tweets. Thereinto, tweets with some special hashtags are positive examples and those without such hashtags are negative examples. MMSD2.0 is upgraded from MMSD. The providers removed misleading cues such as hashtags and emoji words, while it does not address the inherent label bias and word bias present in the dataset, which still influence the model's learning process and potentially lead to biased predictions. The statistics of these two datasets are shown in Table 2.

Evaluation Metrics. Following previous works [Liu *et al.*, 2022; Qin *et al.*, 2023], we adopt accuracy (Acc.), precision (P), recall (R) and micro-average F1 score (F1) to evaluate the model performance. Note that higher numerical values signify better performance across all metrics.

Comparison Models. In our experiments, we choose seven representative MSD models: D&R Net [Xu *et al.*, 2020], InCrossMGs [Liang *et al.*, 2021], HFM [Cai *et al.*, 2019], Att-BERT [Pan *et al.*, 2020], CMGCN [Liang *et al.*, 2022], HKE [Liu *et al.*, 2022] and Multi-view CLIP [Qin *et al.*, 2023]. In detail, to demonstrate the effectiveness of the proposed TFCD, we compare the performance of five *reproducible* models with and without the TFCD framework.

Model	MMSD				MMSD2.0			
	Acc. (%)	P (%)	R (%)	F1 (%)	Acc. (%)	P (%)	R (%)	F1 (%)
D&R Net* [Xu *et al.*, 2020]	84.02	77.97	83.42	80.60	–	–	–	–
InCrossMGs* [Liang *et al.*, 2021]	86.10	81.38	84.36	82.84	–	–	–	–
HFM† [Cai *et al.*, 2019]	83.58	76.79	84.03	80.35	71.04	64.92	69.63	67.01
HFM† + TFCD	84.63 (↑1.05)	77.95 (↑1.16)	85.08 (↑1.05)	81.48 (↑1.13)	72.01 (↑0.97)	65.83 (↑0.91)	70.44 (↑0.81)	67.86 (↑0.85)
Att-BERT† [Pan *et al.*, 2020]	86.21	78.79	83.42	80.93	80.10	76.35	77.76	77.14
Att-BERT† + TFCD	87.48 (↑1.27)	79.83 (↑1.04)	84.83 (↑1.41)	82.22 (↑1.29)	80.98 (↑0.88)	77.28 (↑0.93)	78.86 (↑1.10)	78.20 (↑1.06)
CMGCN† [Liang *et al.*, 2022]	86.63	82.14	83.95	83.49	79.92	75.84	78.10	76.86
CMGCN† + TFCD	87.85 (↑1.22)	83.24 (↑1.10)	85.29 (↑1.34)	84.75 (↑1.26)	80.86 (↑0.94)	76.87 (↑1.03)	79.06 (↑0.96)	77.87 (↑1.01)
HKE† [Liu *et al.*, 2022]	87.34	82.36	86.53	84.38	76.47	73.51	71.62	72.40
HKE† + TFCD	88.67 (↑1.33)	83.43 (↑1.07)	87.78 (↑1.25)	85.52 (↑1.14)	77.51 (↑1.04)	74.38 (↑0.87)	72.54 (↑0.92)	73.28 (↑0.88)
Multi-view CLIP† [Qin *et al.*, 2023]	88.29	83.51	88.32	86.84	85.35	81.37	87.05	83.28
Multi-view CLIP† + TFCD	89.57 (↑1.28)	84.83 (↑1.32)	89.43 (↑1.11)	88.13 (↑1.29)	86.54 (↑1.19)	82.46 (↑1.09)	87.95 (↑0.90)	84.31 (↑1.03)

Table 3: Main results. Results with '∗' denote that the code is not released. Results with '†' stand for the model we re-implemented.

Model	MMSD				MMSD2.0			
	Acc. (%)	P (%)	R (%)	F1 (%)	Acc. (%)	P (%)	R (%)	F1 (%)
HFM† + TFCD	84.63 (-)	77.95 (-)	85.08 (-)	81.48 (-)	72.01 (-)	65.83 (-)	70.44 (-)	67.86 (-)
w/o Label Debiasing	84.19 (↓0.44)	77.33 (↓0.62)	84.50 (↓0.58)	80.95 (↓0.53)	71.49 (↓0.52)	65.35 (↓0.48)	70.07 (↓0.37)	67.45 (↓0.41)
w/o Word Debiasing	83.96 (↓0.67)	77.24 (↓0.71)	84.48 (↓0.60)	80.91 (↓0.57)	71.48 (↓0.53)	65.32 (↓0.50)	70.03 (↓0.42)	67.38 (↓0.48)
HFM†	83.58 (↓1.05)	76.79 (↓1.16)	84.03 (↓1.05)	80.35 (↓1.13)	71.04 (↓0.97)	64.92 (↓0.91)	69.63 (↓0.81)	67.01 (↓0.85)
Multi-view CLIP† + TFCD	89.57 (-)	84.83 (-)	89.43 (-)	88.13 (-)	86.54 (-)	82.46 (-)	87.95 (-)	84.31 (-)
w/o Label Debiasing	88.92 (↓0.65)	84.07 (↓0.76)	88.84 (↓0.59)	87.40 (↓0.73)	85.97 (↓0.57)	81.92 (↓0.54)	87.47 (↓0.48)	83.72 (↓0.59)
w/o Word Debiasing	88.91 (↓0.66)	83.99 (↓0.84)	88.80 (↓0.63)	87.36 (↓0.77)	85.91 (↓0.63)	81.88 (↓0.58)	87.47 (↓0.48)	83.70 (↓0.61)
Multi-view CLIP†	88.29 (↓1.28)	83.51 (↓1.32)	88.32 (↓1.11)	86.84 (↓1.29)	85.35 (↓1.19)	81.37 (↓1.09)	87.05 (↓0.90)	83.28 (↓1.03)

Table 4: Experiment results of ablation study across different datasets. '*w/o*' is short for 'without'.

Implementation Details. The proposed TFCD and five reproducible models are implemented on PyTorch [Paszke *et al.*, 2017]. All experiments are conducted on Nvidia Tesla V100 GPUs. We have utilized grid search to determine the optimal values for the parameters α and β on the validation set. The grid search is performed with a step size of 0.1 and a range spanning from 0 to 1. For a fair comparison, the training settings (*e.g.*, loss function, batch size, learning rate strategy, etc) of these models are consistent with the details reported in their original papers. The results reported in our experiments are the average scores from five random runs on the test set. Please refer to the Appendix for more details.

5.2 Main Results

The main results of TFCD and baselines are shown in Table 3, from which we have the following observations:

(1) The baselines with our TFCD significantly outperform their original counterparts across all evaluation metrics on both datasets. This validates the superior *generalizability* ability of our framework over existing methods. On the other hand, the better results show that these biases are ignored during previous MSD studies, which further supports our claims and motivation. (2) The improvements on MMSD are much sharper than MMSD2.0. We hypothesize that this is due to the specific characteristics between datasets: MMSD2.0, by correcting erroneous labels in MMSD, has somewhat mitigated the issue of label imbalance. Nonetheless, our TFCD consistently achieves gains, which illustrates the *robustness* of our debiasing framework across varying datasets.

5.3 Ablation Study

We select two representative models, *i.e.*, HFM and Multi-view CLIP, to evaluate the contribution of each component in the proposed TFCD framework. The ablation studies on both datasets are reported in Table 4, where all the improvements are statistically significant, as evidenced by the paired t-tests with a p-value < 0.05. And we have the following takeaways:

(1) Removing any component results in a decrease across all metrics on both datasets, which verifies the effectiveness of the proposed label and word debiasing. This is because label debiasing introduces a global offset, while word debiasing contributes to a local one to 'move' in the predicted space, which renders the trained models 'blind' to potentially harmful biases present in the observed data, allowing them to focus solely on the core content of each sample for inference. (2) The improvements in word debiasing are more pronounced. This could be attributed to the fact that trained models typically utilize word-level information for inference, which may inevitably utilize sarcasmless words that are potential biases. Thanks to word debiasing, TFCD addresses spurious correlations introduced by these words to some extent.

Effect of Label Debiasing. One of the core contributions of our work is to achieve label debiasing, where the uniform distribution assumption is introduced under the no-treatment conditions to obtain the intervened outcomes. To verify its effectiveness, we substitute it with two candidate assumptions: `Random` denotes $\mathbf{M}^*$ is learned without any constraint and `Prior` denotes that $\mathbf{M}^*$ obeys the prior distribution of the training set. As shown in Table 5, we find that both `Random`

Model	Metric	Uniform	Random	Prior
HFM† + TFCD	Acc. (%)	84.63	83.47	83.81
	F1 (%)	81.48	80.23	80.66
Mul-CLIP† + TFCD	Acc. (%)	89.57	88.25	88.62
	F1 (%)	88.13	86.78	87.23

Table 5: Ablation study about label debiasing on MMSD. 'Mul-CLIP' represents 'Multi-view CLIP' [Qin *et al.*, 2023].

Model	Metric	Full	*w/o* M	*w/* All M	*w/* Rand M
HFM† + TFCD	Acc. (%)	84.63	84.05	84.33	84.17
	F1 (%)	81.48	80.88	81.14	80.95
Mul-CLIP† + TFCD	Acc. (%)	89.57	89.02	89.26	88.89
	F1 (%)	88.13	87.53	87.85	87.41

Table 6: Ablation study about word debiasing on MMSD. 'M' means the mask operation in Eq. (5). 'Rand' is short for 'Random'.

and `Prior` even perform worse than the baseline counterparts. We attribute this to the fact that the uniform distribution assumption guarantees a safe estimation for the label biases.

Effect of Word Debiasing. To further investigate the effectiveness of word debiasing, we devised three distinct variants: word non-masking (`w/o M`), all masking (`w/ All M`), and random masking (`w/ Random M`). From the results in Table 6, the decreased performance confirms the effect of selectively masking the main-content words. The proposed TFCD enables the baseline to more effectively identify and eliminate spurious correlations introduced by sarcasmless words, while other masking strategies tend to intertwine with the statistical shortcuts of main-content words to varying degrees.

5.4 Method Analysis

Applicability across Pre-trained Backbones. A natural question that arises is whether our model is effective for diverse pre-trained backbones. To answer the question, we conduct experiments with four variants of CMGCN with and without our proposed TFCD by using different text and image encoders. From the results in Table 7, we observe that the gain from TFCD increases as more advanced pre-trained backbone networks are used. This proves the performance of TFCD doesn't rely on a specially chosen backbone.

Sensitivity of Hyper-parameters. The hyper-parameters α and β in Eq.(6) indicate the degree of subtraction of label-debiased predictive logits $\mathbf{o}^{*}$ and word-debiased predictive logits $\mathbf{o}^{\dagger}$ from biased predictive logits $\mathbf{o}$, respectively. According to §5.1, we evaluate the scale range setting $\alpha, \beta \in [0.0, 1.0]$ with a step size of 0.1 as shown in Figure 4. We find that as the coefficient incrementally increases, there is a corresponding increase in accuracy. Optimal performance for each debiasing strategy is achieved at distinct points (0.6 and 0.4 for label and word debiasing in practice, respectively), beyond which a downward trend emerges. Furthermore, we find that label debiasing exhibits relatively mild, whereas word debiasing demonstrates heightened sensitivity to coefficient changes, and requires more careful tuning.

Comparison with Other Debiasing Strategies. We also conduct comparison experiments to verify the proposed debiasing framework against other debiasing strategies, and the results are reported in Table 8. We can find that our framework outperforms typical debiasing strategies, from data manipulation to model balancing. Note that these two debiasing approaches require extra manual or training costs. Whereas, the proposed framework works only in inference and can thus be employed on the previously already trained models, which can serve as a powerful, 'data-manipulation-free' and 'model-balancing-free' weapon to enhance current MSD baselines.

T	V	MMSD		MMSD2.0	
		w/o TFCD	*w/* TFCD	*w/o* TFCD	*w/* TFCD
GloVe	ResNet	85.46	86.69	78.78	79.73
GloVe	ViT	85.65	86.86	79.04	80.07
BERT	ResNet	86.31	87.53	79.59	80.48
BERT	ViT	86.63	87.85	79.92	80.86

Table 7: Accuracy performance of using different pre-trained backbones for CMGCN [Liang *et al.*, 2022] on two datasets. "T" and "V" denote the textual and visual modality, respectively.

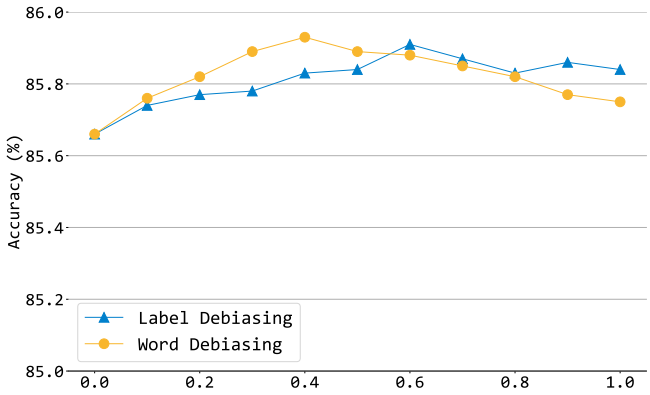

Figure 4: Accuracy performance of Multi-view CLIP [Qin *et al.*, 2023] with the two proposed debiasing strategies across different debiasing coefficients on the MMSD2.0 *validation* set.

Robustness under Low-resource Scenario. To further explore the effectiveness of TFCD under the low-resource scenario [Chen *et al.*, 2024b], we conduct experiments following [Qin *et al.*, 2023] to utilize various quantities of training samples including 10%, 20% and 50%. From Figure 6, we observe that Multi-view CLIP with our TFCD, consistently surpasses its baseline counterpart under low-resource scenarios. We attribute this to our proposed debiasing strategies (especially label debiasing), the model can remove spurious correlations even with varying proportions of positive and negative training samples. This demonstrates the robustness of our TFCD against inconsistent distributions of training and test samples, which achieves considerable improvements.

Generalizability on Sentiment Analysis. To verify the generalizability of our TFCD on other tasks, we conduct preliminary experiments on multi-modal sentiment analysis, where we select two representative models (*i.e.*, MulT [Tsai *et al.*, 2019] and DMD [Li *et al.*, 2023]). The results are reported in Table 9. We find that models with TFCD consistently boost performance by approximately 1% over their vanilla counterparts, which demonstrates the generalizability of our TFCD. We refer readers to Appendix for more details.

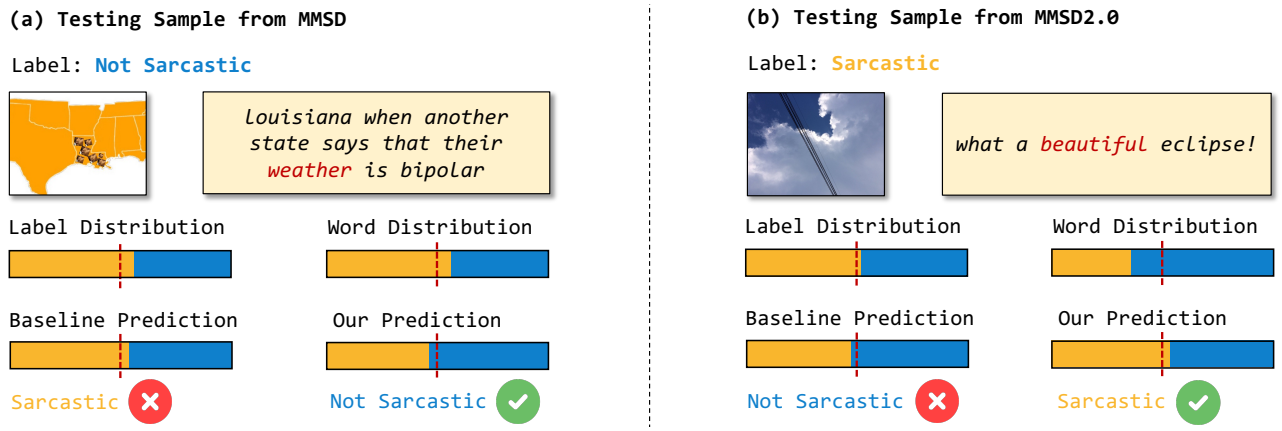

Figure 5: Case study on MMSD (a) and MMSD2.0 (b). "Label/word distribution" denote the distribution of label and sarcasmless word (marked red within the sentence) coming from the training set, where the blue and yellow colors denote probabilities of the non-sarcastic and sarcastic labels, respectively. The bars at the bottom demonstrate the prediction distribution of the best baseline Multi-view CLIP with and without the proposed TFCDframework. Better view in color.

Model	Metric	Base	*w/* DM	*w/* MB	*w/* TFCD
HFM†	Acc. (%)	83.58	83.95	84.20	84.63
	F1 (%)	80.35	80.64	81.06	81.48
Mul-CLIP†	Acc. (%)	88.29	88.72	89.04	89.57
	F1 (%)	86.84	87.17	87.65	88.13

Table 8: Ablation study about debiasing strategies on MMSD. 'DM' and 'MB' denote typical data manipulation [Qian *et al.*, 2020] and model balancing [Zhang *et al.*, 2020] method, respectively.

Model	MOSEI		MOSI	
	Acc-2 (%)	Acc-7 (%)	Acc-2 (%)	Acc-7 (%)
MulT†	81.84	52.76	83.95	41.47
MulT† + TFCD	82.79	53.84	85.01	42.88
DMD†	86.31	54.35	85.92	45.87
DMD† + TFCD	87.70	55.63	86.87	46.94

Table 9: Performance on two multi-modal sentiment analysis datasets. Results with '†' stand for the model we implemented.

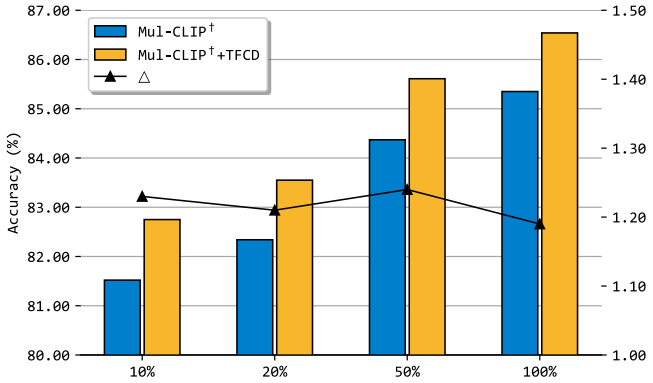

Figure 6: Low-resource performance of Multi-view CLIP [Qin *et al.*, 2023] (denoted as Mul-CLIP in the plot) on MMSD2.0. '△' denotes relative improvement achieved by TFCD upon the baseline.

5.5 Case Study

In Figure 5, we select one representative example from each dataset to show the performance of the model with and without the TFCD. For instance, in Figure 5(a), the vanilla baseline is misled to predict sarcastic because the MMSD is dominated by the sarcastic class (*label* bias) and the '`weather`' word within the sentence is mostly associated with the sarcastic class (*word* bias). Applying counterfactual debiasing, TFCD corrects the *label* and *word* biases in the model's prediction. In Figure 5(b), the vanilla baseline is misled to predict not sarcastic due to the serious '`beautiful`' word distribution in not sarcastic class. Thanks to the proposed TFCD, the model now results in the correct prediction.

6 Conclusion

This paper proposed a training-free debiasing strategy termed TFCD to reduce the harmful bias of *statistical label* and *sarcasmless words* for multi-modal sarcasm detection (MSD). Concretely, TFCD disentangled the causalities among variables via a tailored causal graph and presented a biases extraction module to extract the adverse effect caused by the two biases. These biases were then mitigated by element-wise subtraction to achieve debiased inference. Numerous experiments proved that TFCD could consistently improve existing baselines. The model-agnostic and training-free TFCD undoubtedly has superiority over complex retraining modules employed in previous approaches.

Limitations and Future Work. Although our TFCD has demonstrated promising outcomes, it can still benefit from the following two aspects: (1) Further exploration of the debiasing technique on visual component [Chen *et al.*, 2024a]. (2) In scenarios involving samples with highly balanced classes, the impact of label bias becomes negligible. Future work will extend our framework under more challenging scenarios (*e.g.*, out-of-distribution and long document) for MSD.

References

[Cai *et al.*, 2019] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *ACL*, 2019.

[Chen *et al.*, 2023] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *ACL*, 2023.

[Chen *et al.*, 2024a] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. *ICML*, 2024.

[Chen *et al.*, 2024b] Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. AutoPRM: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *NAACL*, 2024.

[Jia *et al.*, 2024] Mengzhao Jia, Can Xie, and Liqiang Jing. Debiasing multimodal sarcasm detection with contrastive learning. In *AAAI*, 2024.

[Kang *et al.*, 2019] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv*, 2019.

[Li *et al.*, 2023] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *CVPR*, 2023.

[Liang *et al.*, 2021] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *ACM MM*, 2021.

[Liang *et al.*, 2022] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *ACL*, 2022.

[Lin *et al.*, 2022] Zhiyu Lin, Yifei Gao, and Jitao Sang. Investigating and explaining the frequency bias in image classification. In *IJCAI*, 2022.

[Liu *et al.*, 2022] Hui Liu, Wenya Wang, and Haoliang Li. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *EMNLP*, 2022.

[Niu *et al.*, 2021] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *CVPR*, 2021.

[Pan *et al.*, 2020] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *EMNLP Findings*, 2020.

[Pan *et al.*, 2023] Hang Pan, Jiawei Chen, Fuli Feng, Wentao Shi, Junkang Wu, and Xiangnan He. Discriminative-invariant representation learning for unbiased recommendation. In *IJCAI*, 2023.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS*, 2017.

[Pearl and Mackenzie, 2018] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[Pearl *et al.*, 2016] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[Poria *et al.*, 2016] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. A deeper look into sarcastic tweets using deep convolutional neural networks. In *COLING*, 2016.

[Qian *et al.*, 2020] Chen Qian, Fuli Feng, Lijie Wen, Li Lin, and Tat-Seng Chua. Enhancing text classification via discovering additional semantic clues from logograms. In *SIGIR*, 2020.

[Qian *et al.*, 2021] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. Counterfactual inference for text classification debiasing. In *ACL*, 2021.

[Qin *et al.*, 2023] Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In *ACL Findings*, 2023.

[Radford *et al.*, 2021] 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. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Riloff *et al.*, 2013] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, 2013.

[Schifanella *et al.*, 2016] Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. Detecting sarcasm in multimodal social platforms. In *ACM MM*, 2016.

[Sun *et al.*, 2022] Teng Sun, Wenjie Wang, Liqiang Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *ACM MM*, 2022.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for unaligned multimodal language sequences. In *ACL*, 2019.

[Tu *et al.*, 2023] Geng Tu, Ran Jing, Bin Liang, Min Yang, Kam-Fai Wong, and Ruifeng Xu. A training-free debiasing framework with counterfactual reasoning for conversational emotion detection. In *EMNLP*, 2023.

[Wang and Culotta, 2021] Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification

via automatically generated counterfactuals. In *AAAI*, 2021.

[Xin *et al.*, 2023] Yifei Xin, Dongchao Yang, Fan Cui, Yujun Wang, and Yuexian Zou. Improving weakly supervised sound event detection with causal intervention. In *ICASSP*, 2023.

[Xu *et al.*, 2020] Nan Xu, Zhixiong Zeng, and Wenji Mao. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *ACL*, 2020.

[Yang *et al.*, 2024a] Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. Towards multimodal sentiment analysis debiasing via bias purification. *arXiv*, 2024.

[Yang *et al.*, 2024b] Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. Robust emotion recognition in context debiasing. *CVPR*, 2024.

[Zhang *et al.*, 2016] Meishan Zhang, Yue Zhang, and Guohong Fu. Tweet sarcasm detection using deep neural network. In *COLING*, 2016.

[Zhang *et al.*, 2020] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *ACL*, 2020.

[Zhou *et al.*, 2023] Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. Causal-Debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *ACL*, 2023.

[Zhu *et al.*, 2022] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. Generalizing to the future: Mitigating entity bias in fake news detection. In *SIGIR*, 2022.

[Zhu *et al.*, 2024a] Zhihong Zhu, Xuxin Cheng, Guimin Hu, Yaowei Li, Zhiqi Huang, and Yuexian Zou. Towards multimodal sarcasm detection via disentangled multi-grained multi-modal distilling. In *COLING*, 2024.

[Zhu *et al.*, 2024b] Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. Dance with labels: Dual-heterogeneous label graph interaction for multi-intent spoken language understanding. In *WSDM*, 2024.