General Debiasing for Multimodal Sentiment Analysis

Teng Sun
Shandong University
stbestforever@gmail.com

Juntong Ni
Shandong University
juntongni02@gmail.com

Wenjie Wang*
National University of Singapore
wenjiewang96@gmail.com

Liqiang Jing
Shandong University
jingliqiang6@gmail.com

Yinwei Wei
National University of Singapore
weiyinwei@hotmail.com

Liqiang Nie
Harbin Institute of Technology (Shenzhen)
nieliqiang@gmail.com

ABSTRACT

Existing work on Multimodal Sentiment Analysis (MSA) utilizes multimodal information for prediction yet unavoidably suffers from fitting the spurious correlations between multimodal features and sentiment labels. For example, if most videos with a blue background have positive labels in a dataset, the model will rely on such correlations for prediction, while "blue background" is not a sentiment-related feature. To address this problem, we define a general debiasing MSA task, which aims to enhance the Out-Of-Distribution (OOD) generalization ability of MSA models by reducing their reliance on spurious correlations. To this end, we propose a general debiasing framework based on Inverse Probability Weighting (IPW), which adaptively assigns small weights to the samples with larger bias (*i.e.,* the severer spurious correlations). The key to this debiasing framework is to estimate the bias of each sample, which is achieved by two steps: 1) disentangling the robust features and biased features in each modality, and 2) utilizing the biased features to estimate the bias. Finally, we employ IPW to reduce the effects of large-biased samples, facilitating robust feature learning for sentiment prediction. To examine the model's generalization ability, we keep the original testing sets on two benchmarks and additionally construct multiple unimodal and multimodal OOD testing sets. The empirical results demonstrate the superior generalization ability of our proposed framework. We have released the code to facilitate the reproduction https://github.com/Teng-Sun/GEAR.

CCS CONCEPTS

• **Information systems → Sentiment analysis**.

KEYWORDS

Multimodal Sentiment Analysis, Debiasing, Out-of-distribution Generalization

ACM Reference Format:
Teng Sun, Juntong Ni, Wenjie Wang, Liqiang Jing, Yinwei Wei, and Liqiang Nie. 2023. General Debiasing for Multimodal Sentiment Analysis. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612051

1 INTRODUCTION

Sentiment analysis, a classical language understanding task, has attracted wide attention from the academy and industry. The early work chiefly utilizes users' textual reviews to analyze their sentiment [15, 21]. However, single textual modality usually has problems of polysemy and ambiguity [3, 43]. Along with the advance of social media, more and more people begin to adopt multimodal information to express their sentiments, such as video and audio [11, 12, 25, 27, 28]. The cross-modal consistency and complementarity provide rich semantic information for sentiment analysis. Therefore, Multimodal Sentiment Analysis (MSA) has been a popular research field in recent years [16, 22, 38, 40, 45].

Previous studies of MSA predominantly pay attention to representation learning and multimodal fusion. For representation learning, some researchers utilize techniques like adversarial learning [18] and multi-task learning [8] to map features from different modalities into a shared representation space. Self-supervised learning [39] is also used to incorporate unimodal information into the fusion model to aid representation learning. For multimodal fusion, previous work manages to learn cross-modal representation using sophisticated fusion mechanisms, such as tensor-based fusion [41] and graph-based fusion [44]. In addition, some studies [24, 31] also attempt to integrate modalities via pre-trained transformers (*e.g.,* BERT [5] and XLNet [37]).

However, existing studies usually suffer from fitting the spurious correlation between multimodal features and sentiment labels. As shown in Figure 1, the word "movie" in Figure 1(a) and the attribute "blue background" in Figure 1(b) show strong correlations with negative and positive sentiment labels, respectively. However, "movie" and "blue background" are not reliable cues for identifying sentiment. Due to the short-cut bias [7], the MSA models will easily learn such spurious correlations for prediction, impairing the generalization ability in the Out-Of-Distribution (OOD) testing data, where the correlations between multimodal features and sentiment labels differ from those in the training data. For instance, Sun *et al.* [29] pointed out the spurious correlations between textual words and sentiment labels. Nevertheless, there are also spurious correlations in video and audio modalities in addition to textual modality.

*Wenjie Wang (wenjiewang96@gmail.com) is corresponding author.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612051

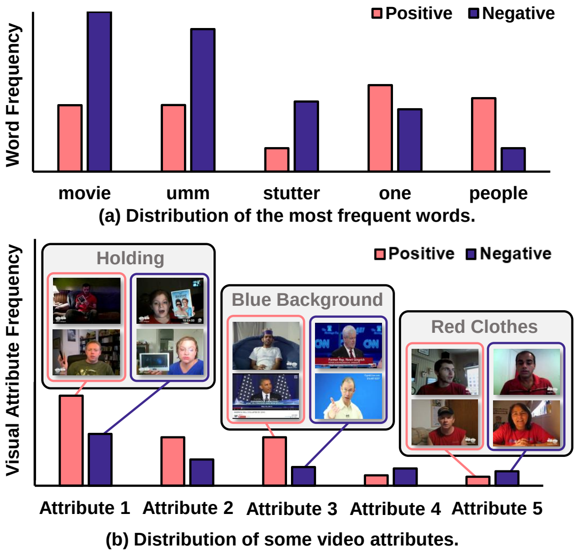


Figure 1: The distribution of the top-5 most frequent words and visual attributes.

To address the above problems, we first propose a general debiasing task for MSA, which aims to enhance the OOD generalization ability of MSA models by reducing the bad effect of multimodal spurious correlations. To mitigate the effect of spurious correlations, a widely used method is Inverse Probability Weighting (IPW), where a sample with strong bias will be assigned a small weight for training. To implement IPW, the key lies in estimating the bias of each sample, which depends on two steps: 1) disentangling the robust features (*i.e.,* sentiment-related features such as smiling face) and biased features (*i.e.,* sentiment-irrelevant features such as the blue background) in each modality, and 2) utilizing the biased features to estimate the sample bias.

To disentangle multimodal features for estimating bias weights, we propose a General dEbiAsing fRamework (GEAR) with three stages. First, we design three pairs of robust extractors and biased extractors, where each pair is used to extract the robust features and biased features in a modality. Second, to disentangle biased features, prior studies usually consider using Generalized Cross Entropy (GCE) loss [46] to train the biased extractor and amplify the prejudice for bias estimation. However, such GCE loss cannot be applied to debiasing MSA since MSA is usually formulated as a regression task instead of the classification task [41]. Toward this challenge, we propose a novel Generalized Mean Absolute Error (GMAE) loss, which is specially designed to disentangle biased features in the MSA task. We then estimate the bias weight from biased features by calculating the absolute error between the outputs of the three biased extractors and sentiment labels. The underlying philosophy is that the biased features with strong correlations will have a lower absolute error and vice versa. Third, to reinforce the generalization ability, we use the estimated bias weights to adjust IPW-based Mean Absolute Error (MAE) loss for debiasing training and fuse the robust features of three modalities for prediction.

To evaluate the generalization ability of MSA models, we construct four OOD testing sets while keeping the original testing set as an Independent and Identical Distribution (IID) testing set on two benchmarks. The empirical results demonstrate the superior generalization ability of GEAR on OOD testing sets while maintaining comparable IID performance with state-of-the-art methods. To sum up, our contributions are threefold.

- To the best of our knowledge, we are the first to formulate a general debiasing MSA task from multiple modalities. Meanwhile, to examine the generalization ability of MSA models, we contribute several multimodal OOD testing datasets.
- We propose a novel framework GEAR, which strengthens the generalization ability of MSA models by disentangling the robust and biased features via a novel GMAE loss and estimating the bias weight of each sample for IPW-enhanced debiasing training.
- We conduct extensive experiments on two datasets (*i.e.,* MOSEI and MOSI [43]), and the experimental results demonstrate the superior generalization ability of GEAR.

2 RELATED WORK

• **Multimodal Sentiment Analysis.** In recent years, a substantial number of researchers have explored the MSA. The prior researchers mainly focused on representation learning and multimodal fusion. For representation learning, previous studies mainly are in three variants: 1) Shift-based models shift textual representations based on aligned nonverbal behaviors (*i.e.,* audio and vision modality) [35]. 2) Shared subspace learning models map all the modalities simultaneously into modality-invariant and modality-specific representations [8]. And 3) Self-supervised models generate the unimodal labels by self-supervised learning strategy and use multi-task learning to train the model [39]. For multimodal fusion, according to the fusion stages, two multimodal fusion strategies are applied: 1) early fusion [24, 30, 31, 36, 42] means that the features of different modalities are combined together in an early stage. And 2) late fusion [4, 17, 41] indicates that the intra-modal representation is learned first and inter-modal fusion is performed last.

Although existing studies have achieved great success, they ignore the spurious correlations between modalities and sentiment labels. Hence, Sun *et al.* [29] is the first to settle this issue, which introduces a model-agnostic counterfactual reasoning framework (CLUE) for MSA that can leverage the positive aspects of text-based modality and mitigate potential drawbacks. However, CLUE can only handle the case of spurious correlations in a single textual modality, and cannot satisfactorily deal with multiple modalities with spurious correlations. To this end, we presented a general debiasing MSA network to improve the OOD generalization ability.

• **General Debiasing Methods.** The existing general debiasing methods can be divided into three categories. 1) Debiasing with known bias types and labels. Many debiasing methods [1, 10, 34] require explicit bias types and bias annotations for each training sample. 2) Debiasing with known bias types. To eliminate the costs of bias annotations, some bias-tailored studies [2, 32] only require bias types. 3) Debiasing with unknown bias types. The above assumptions face limitations since manually discovering bias types strongly relies on expert knowledge and laborious labeling [20, 33]. The following work estimates the bias of each sample without knowing its bias types and labels. Nam *et al.* [20] trained a debiased classifier from the biased classifier by utilizing GCE and relative difficulty score. Lee *et al.* [14] learned debiased representation via disentangled feature augmentation. Fan *et al.* [6] applied the methods from the previous two works to graph data debiasing.

However, multimodal data have the complex bias which is infeasible to be recognized. To this end, we resorted to the third category. Yet most existing methods are designed for image datasets and could not effectively conduct debiasing with multimodal data. Thus, we designed a debiasing framework specified for multimodal features.

3 METHODOLOGY

In this section, we first give the task formulation, and then we present the overall framework and the detailed modules.

3.1 Task formulation

3.1.1 Traditional MSA Task. Let $\mathcal{D} = \{T_i, A_i, V_i, Y_i\}_1^N$ denote the MSA training set with N training samples. Each quadruple is from a video segment, where T_i, A_i, V_i and Y_i denote text, audio, video, and the corresponding sentiment label of the i-th sample, respectively. The traditional MSA task aims to develop a model $\mathcal{F}_\theta$ which jointly utilizes three modalities (*i.e.,* T_i, A_i and V_i) to predict sentiment label Y_i as follows,

$$\tilde{Y}_i = \mathcal{F}_\theta(T_i, A_i, V_i), \tag{1}$$

where θ represents the learnable parameters and $\tilde{Y}_i$ denotes the predicted label of the i-th sample. For clarity, we temporally omit the subscript i of the training samples.

3.1.2 General Debiasing MSA Task. Traditional MSA models rely on complementary and consistent multimodal information for sentiment prediction. However, existing MSA models suffer from spurious correlations between multimodal features and sentimental labels. As mentioned before, the word "movie" is highly correlated with the negative sentiment, and the attribute "blue background" has a strong correlation with the positive sentiment in the MOSEI dataset. Trained with such biased data, models tend to predict samples with the word "movie" as negative samples and "blue background" as positive samples, which strongly deteriorates the generalization performance of MSA models.

To reduce the negative influence caused by multimodal spurious correlations, we formulate the general debiasing MSA task, which aims to evaluate the generalization ability of different MSA models on the OOD testing set. To achieve this, we propose an algorithm to automatically build OOD testing sets based on the original IID testing set. Specifically, the OOD testing sets have significantly different multimodal sentiment correlations from the training set. The different multimodal distributions in OOD testing sets and the training set are able to effectively evaluate whether MSA models have strong debiasing ability.

3.2 Framework

As can be seen in Figure 2, the overall framework is as follows: 1) Disentangled Representation Learning Module: for a given sample with three modalities, we first disentangle the robust and biased features of each modality by robust and biased extractors, respectively. Then, we swap the robust and biased features, which synthesizes more diverse samples and facilitates the disentanglement. 2) Bias Estimation Module: we devise GMAE loss to boost the training of bias extractors. In addition, we calculate the absolute values between the prediction based on multimodal biased features and sentiment labels. The absolute values of three modalities are used to estimate the bias weight of each sample. 3) Debiasing Optimization Module: we fuse the multimodal robust features by multi-head self-attention and employ IPW-enhanced MAE loss for training robust extractors. We utilize IPW to re-weight the samples by bias weights, which discourages the influence of samples with a large bias. Each module will be elaborated on in the following sub-sections.

3.3 Disentangled Representation Learning

To disentangle the robust features and biased features in each modality, we simultaneously train three pairs of biased extractors and robust extractors. In addition, to facilitate disentanglement, we swap the robust and biased latent vectors and synthesize more diverse samples.

3.3.1 Robust and Biased Extractors. To extract robust and biased features in each modality, we present three pairs of the robust extractors E_R^m and biased extractors $E_B^m, m \in \{t, a, v\}$.

Extractors for Textual Modality. In the textual modality, due to the great success of large pre-trained transformer-based language models, we utilize the pre-trained BERT as the backbone to extract textual representations of the raw text. Similar to the existing studies [24], we select the first [CLS] token in the last layer as the whole textual representation. After that, we employ the linear layers to map the features to the low-dimension semantic space. We feed raw text T into the textual robust extractors and biased extractors to gain robust and biased latent vectors of text (*i.e.,* $\mathbf{v}_\kappa^t$). The whole structures of the textual robust and biased extractors are formulated as follows,

$$\mathbf{v}_\kappa^t = E_\kappa^t(T) = \mathbf{W}_\kappa^t(BERT_\kappa^t(T)) + \mathbf{b}_\kappa^t, \tag{2}$$

where $\kappa \in \{R, B\}$, $\mathbf{v}_\kappa^t \in \mathbb{R}^{d_s}$, $\mathbf{W}_\kappa^t \in \mathbb{R}^{d_s \times d_t}$, $\mathbf{b}_\kappa^t \in \mathbb{R}^{d_s}$, d_s and d_t denote the dimensions of the latent vectors and BERT's output.

Extractors for Acoustic and Visual Modalities. In acoustic and visual modalities, we employ the hand-crafted features extracted by Yu *et al.* [39] from the raw data, $\mathbf{A} \in \mathbb{R}^{l_a \times d_a}$ and $\mathbf{V} \in \mathbb{R}^{l_v \times d_v}$. Here, l_a and l_v are the sequence lengths of audio and video, respectively. d_a and d_v are the extracted hand-crafted features dimension of audio and video, respectively. Then, we use the 1-layer Long Short-Term Memory (LSTM) [9] to capture the temporal information. Similar to previous work [8, 39], we select the final states vector of LSTM as the whole modality representation. The robust and biased extractors of audio and video are similar to the ones of text except for the backbone, where we replace BERT with LSTM. The structures of the robust and biased extractors of audio and video are as follows,

$$\begin{cases} \mathbf{v}_\kappa^a = E_\kappa^a(\mathbf{A}) = \mathbf{W}_\kappa^a(LSTM_\kappa^a(\mathbf{A})) + \mathbf{b}_\kappa^a, \\ \mathbf{v}_\kappa^v = E_\kappa^v(\mathbf{V}) = \mathbf{W}_\kappa^v(LSTM_\kappa^v(\mathbf{V})) + \mathbf{b}_\kappa^v, \end{cases} \tag{3}$$

where $\mathbf{v}_\kappa^{a/v} \in \mathbb{R}^{d_s}$ denote the robust and biased latent vectors of audio and video, $\mathbf{W}_\kappa^{a/v} \in \mathbb{R}^{d_s \times d'_{a/v}}$, $\mathbf{b}_\kappa^{a/v} \in \mathbb{R}^{d_s}$, and $d'_{a/v}$ are the dimension of the output of LSTM.

3.3.2 Diversify Samples via Swap. We argue that the diversity of samples is of vital importance for disentanglement. With the visual

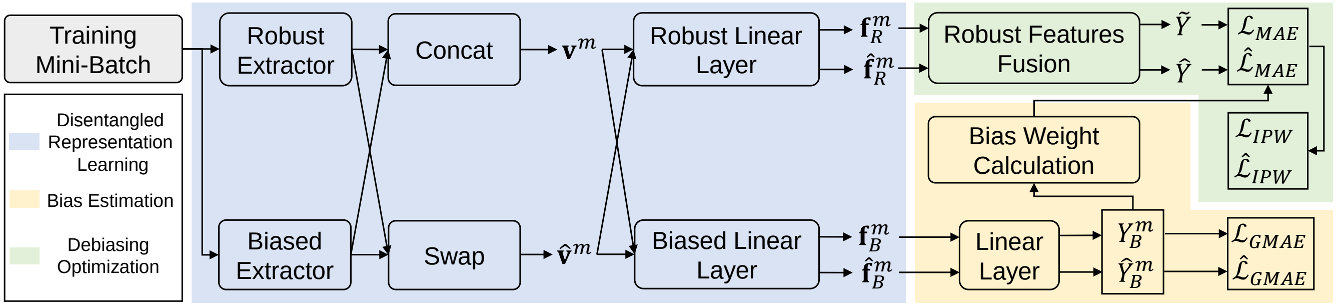


Figure 2: Illustration of the proposed general debiasing framework, which consists of disentangled representation learning, bias estimation, and debiasing optimization.

modality as an example, the facial expressions reflect sentiment precisely. It is interesting to note that yellow can make people feel happy, so abundant samples contain the pair of (*Smile*, *Yellow*) which means a smile and yellow background coexist in an image. On the contrary, purple tends to make people feel sad, so ample samples contain (*Frown*, *Purple*). In the two pairs, the facial expressions are robust attributes and the colors of the background are biased attributes. While the above architecture disentangles the robust features and bias features, E_R^m and E_B^m are still mainly trained with small amounts of samples that have poor diversity. Thereby, the above architecture is able to disentangle (*Smile*, *Yellow*) and (*Frown*, *Purple*), but not (*Smile*, *Purple*) and (*Frown*, *Yellow*) due to the limited number of such samples, which deteriorates the model's performance. Thus, we need more samples with rich diversity (*e.g.*, (*Smile*, *Purple*) and (*Frown*, *Yellow*)). To this end, we swap the latent vectors to synthesize more diverse samples.

By three pairs of robust and biased extractors for three modalities, we get robust and biased latent vectors of each modality (*i.e.,* $\mathbf{v}_R^m$ and $\mathbf{v}_B^m$). To diversify samples, we utilize these preliminary disentangled features for swap. We propose to generate diverse samples in latent embedding space by swapping biased latent vectors. More specifically, we replace each biased vector $\mathbf{v}_B^m$ with a randomly selected biased vector $\hat{\mathbf{v}}_B^m \in \mathbb{R}^{d_s}$ in the same mini-batch.

To synthesize diverse samples, we concatenate robust and corresponding biased latent vectors, and also concatenate robust and randomly selected biased latent vectors. In detail, concatenated vectors $\mathbf{v}^m$ are as follows,

$$\mathbf{v}^m = [\mathbf{v}_R^m; \mathbf{v}_B^m], \tag{4}$$

where $\mathbf{v}^m \in \mathbb{R}^{2d_s}$ denote the latent vectors that are combined with the robust latent vectors and biased latent vectors without swapping. Then, we concat $\mathbf{v}_R^m$ and $\hat{\mathbf{v}}_B^m$ to obtain $\hat{\mathbf{v}}^m$ as follows,

$$\hat{\mathbf{v}}^m = [\mathbf{v}_R^m; \hat{\mathbf{v}}_B^m], \tag{5}$$

where $\hat{\mathbf{v}}^m \in \mathbb{R}^{2d_s}$ represent the latent vectors that are combined with robust latent vectors and swapped biased latent vectors. Thus, by swapping, we acquire additional latent vectors $\hat{\mathbf{v}}^m$ that have the same robust latent vector but a different biased latent vector with $\mathbf{v}^m$. By this, we can get more samples with diverse (robust features, biased features) combinations.

To make the disentangled representation learning module meet more diverse samples and gain a stronger ability of disentanglement, $\mathbf{v}^m$ and $\hat{\mathbf{v}}^m$ are both fed into pairs of robust and biased linear layers (L_R^m, L_B^m), which extract robust features $\mathbf{f}_R^m, \hat{\mathbf{f}}_R^m \in \mathbb{R}^{d_s}$ and biased features $\mathbf{f}_B^m, \hat{\mathbf{f}}_B^m \in \mathbb{R}^{d_s}$ of each modality as follows,

$$\begin{cases} \mathbf{f}_\kappa^m = L_\kappa^m(\mathbf{v}^m) = ReLU(\mathbf{W}_\kappa^m \mathbf{v}^m + \mathbf{b}_\kappa^m), \\ \hat{\mathbf{f}}_\kappa^m = L_\kappa^m(\hat{\mathbf{v}}^m) = ReLU(\mathbf{W}_\kappa^m \hat{\mathbf{v}}^m + \mathbf{b}_\kappa^m), \end{cases} \tag{6}$$

where $\mathbf{W}_\kappa^m \in \mathbb{R}^{(d_s)\times 2d_s}$, $\mathbf{b}_\kappa^m \in \mathbb{R}^{d_s}$, and $ReLU(\cdot)$ is the relu activation function [19].

3.4 Bias Estimation

To estimate the bias precisely, we need high-quality bias features. Thus, bias estimation has two steps, 1) training biased extractors to acquire high-quality bias features and 2) utilizing bias features to estimate bias in each modality and calculate bias weight of samples.

3.4.1 GMAE loss. To facilitate biased extractors to gain high-quality bias features, we develop GMAE loss. It is known that the biased features are easier to learn than the robust features in the early stage of training [20]. Based on this observation, prior studies employ GCE loss to train a biased model by amplifying the learning of "easier" bias. To be specific, GCE loss can make the biased model emphasize the "easier" samples with strong agreements between the predictions of the biased model and the labels, which amplifies the "prejudice" of the biased model. This is because the "easier" samples in the early training stage are more likely to be biased and hence the model makes more accurate predictions for biased samples. However, GCE loss is elaborately designed for classification tasks and cannot be employed for a regression task such as MSA. Thus, we develop GMAE loss to amplify the prejudice specifically for the regression task. The designed GMAE loss is as follows,

$$\begin{cases} Y_B^m = \mathbf{w}_B^m \mathbf{f}_B^m + b_B^m, \\ \mathcal{L}_{GMAE}^m(Y, Y_B^m) = -2\ln(e^{|Y - Y_B^m|} + 1) + 2|Y - Y_B^m|, \end{cases} \tag{7}$$

where $\mathbf{w}_B^m \in \mathbb{R}^{1\times d_s}$ and $b_B^m \in \mathbb{R}$ are trainable parameters. $\ln(\cdot)$ denotes natural logarithm, and $|\cdot|$ denotes absolute value. To calculate GMAE loss, we forward the biased features for prediction, $Y_B^m \in \mathbb{R}$ are the sentiment predictions based on biased features $\mathbf{f}_B^m$ without swapping process.

The gradient of the GMAE loss up-weights the gradient of MAE loss when the sample has a low absolute value between the prediction and the label as follows,

$$\nabla\mathcal{L}_{GMAE}(Y, Y_B^m) = \frac{2}{1 + e^{|Y - Y_B^m|}} \nabla\mathcal{L}_{MAE}(Y, Y_B^m). \tag{8}$$

We assign greater weights to samples that are predicted well by the biased model, *i.e.*, the lower $|Y - Y_B^m|$, the higher $\frac{2}{1+e^{|Y-Y_B^m|}}$. In addition, GMAE loss is able to keep the gradient weight between 0 and 1, which avoids gradient explosion and makes the training process more stable. To sum up, by GMAE loss, a sample with an "easier" biased feature could gain low absolute value and high gradient weight while training, which helps the biased model amplify the "prejudice".

Meanwhile, to make biased models learn biased features from swapped samples, we utilize swapped labels $\hat{Y}^m$ for calculating GMAE loss as follows,

$$\begin{cases} \hat{Y}_B^m = \mathbf{w}_B^m \hat{\mathbf{f}}_B^m + b_B^m, \\ \hat{\mathcal{L}}_{GMAE}^m(\hat{Y}^m, \hat{Y}_B^m) = -2\ln(e^{|\hat{Y}^m - \hat{Y}_B^m|} + 1) + 2|\hat{Y}^m - \hat{Y}_B^m|, \end{cases} \tag{9}$$

where swapped labels $\hat{Y}^m$ are along with the same selected sample of $\hat{\mathbf{v}}^m$ to make biased models focus on the bias information, and $\hat{Y}_B^m \in \mathbb{R}$ are the sentiment predictions based on biased features $\hat{\mathbf{f}}_B^m$ with swapping process.

3.4.2 Bias weight Calculation. The biased extractors are trained with amplifying the "prejudice" by GMAE loss so that biased models are good at utilizing biased features for prediction. The more precisely the biased model predicts, the more biased the sample is. Thus, we employ the absolute value calculated between the prediction and the label to measure how much each modality is likely to be biased. The smaller the absolute value, the larger the bias in the modality. Then, we estimate the bias weight of a sample by calculating the minimum or average value of absolute value in each modality and taking the inverse as follows,

$$\begin{cases} \psi_{min}(Y, Y_B^m) = \frac{1}{\min(|Y-Y_B^t|,|Y-Y_B^a|,|Y-Y_B^v|)}, \\ \psi_{avg}(Y, Y_B^m) = \frac{1}{\mathrm{avg}(|Y-Y_B^t|,|Y-Y_B^a|,|Y-Y_B^v|)}, \end{cases} \tag{10}$$

where $\psi(\cdot)$ denotes the bias weight estimation function of a sample, the larger the bias weight is, the more bias a sample has. We regard the two equations in Eqn.(10) as *MinStrategy* and *AvgStrategy*, respectively. We consider that *MinStrategy* selects the most biased modality to indicate how much a sample is biased and *AvgStrategy* estimates the bias degree of a sample based on the biased degree of the three modalities simultaneously.

3.5 Debiasing Optimization

In this module, we first fuse multimodal robust features by multi-head self-attention. To learn robust representations from biased data with spurious correlations, we use IPW-enhanced MAE loss for training, where a sample with strong bias will be assigned with a small weight for training. Finally, we calculate the overall training objective for debiasing optimization.

3.5.1 Robust Features Fusion. Due to the superior performance of sentiment analysis brought by multimodal features, we develop a multimodal fusion mechanism for final prediction. And to reinforce the generalization ability of the model, we utilize the robust features for fusion.

First, we stack the three robust features (from Eqn.(6)) into a matrix $\mathbf{M} = [\mathbf{f}_R^t, \mathbf{f}_R^a, \mathbf{f}_R^v] \in \mathbb{R}^{3\times d_s}$. Then, in order to make each vector aware of its companion cross-modal features, we employ a multi-head self-attention on these features. By doing this, each feature is given the opportunity to gain consistent and complementary information from other features that could contribute to the overall sentiment analysis. Specifically, suppose we have U attention heads, and the attention function of the i-th attention head can be formulated as follows,

$$\begin{cases} \mathbf{Q}_i = \mathbf{M}\mathbf{W}_i^q, \mathbf{K}_i = \mathbf{M}\mathbf{W}_i^k, \mathbf{V}_i = \mathbf{M}\mathbf{W}_i^v, \\ \mathbf{O}_i = softmax(\mathbf{Q}_i\mathbf{K}_i^T/\sqrt{d_s})\mathbf{V}_i, \end{cases} \tag{11}$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\ \mathbf{O}_i \in \mathbb{R}^{3\times\frac{d_s}{U}}$ are the query, the key, and the value projected from the matrix $\mathbf{M}$, respectively. $\mathbf{W}_i^{q/k/v} \in \mathbb{R}^{d_s\times\frac{d_s}{U}}$ are learnable matrices in the i-th attention head. The multi-head self-attention outputs a matrix $\hat{\mathbf{M}} \in \mathbb{R}^{3\times d_s}$ as follows,

$$\hat{\mathbf{M}} = [\mathbf{O}_1; \ldots; \mathbf{O}_U]\mathbf{W}^o, \tag{12}$$

where $\mathbf{W}^o \in \mathbb{R}^{d_s\times d_s}$, and each $\mathbf{O}_i$ here is calculated based on Eqn.(11). Finally, we take the multi-head self-attention output $\hat{\mathbf{M}} \in \mathbb{R}^{3\times d_s}$ and construct a joint-vector $\mathbf{f}_o \in \mathbb{R}^{3d_s}$ using concatenation. The final sentiment predictions are then generated by a classifier as follows,

$$\tilde{Y} = \mathbf{w}\mathbf{f}_o + b, \tag{13}$$

where $\mathbf{w} \in \mathbb{R}^{3d_s}$ and $b \in \mathbb{R}$. Meanwhile, we also fuse the robust features $\hat{\mathbf{f}}_R^m$ disentangled from swapped concatenated vectors in the same way as above to calculate $\hat{Y}$.

3.5.2 IPW-enhanced MAE loss. For the regression task, existing studies mostly utilize MAE loss as follows,

$$\mathcal{L}_{MAE} = |Y - \tilde{Y}|. \tag{14}$$

However, MAE loss treats samples with/without bias equally. Training with MAE loss, robust extractors cannot focus on samples without bias to learn robust features, which acquire features that contain biased features from biased samples. We thus can train robust extractors by making them focus on learning unbiased samples. Our robust extractors are unable to extract robust features from unbiased samples since they cannot recognize which sample is unbiased. To learn robust features from biased data with spurious correlations, a widely used method is IPW [23], where a sample with strong bias will be assigned with a small weight for training. This helps robust extractors focus on learning robust features of unbiased samples. Thus, we utilize IPW-enhanced MAE loss for training, which re-weight the MAE loss by bias weight as follows,

$$\mathcal{L}_{IPW} = \mathcal{L}_{MAE} \cdot \frac{1}{P(x|Bias(x)) + 1}. \tag{15}$$

This IPW implies that if a sample $x = [T, A, V]$ is more likely associated with its biased features (*i.e.*, $Bias(x)$), we should underweight the loss to discourage such a biased sample. We calculate

$P(x|Bias(x))$ as follows,

$$P(x|Bias(x)) \propto \psi(Y, Y_B^m) \cdot (|Y - \tilde{Y}|), \tag{16}$$

where $\psi(\cdot)$ is illustrated in Eqn.(10), and $|Y - \tilde{Y}|$ is the absolute value without gradients.

Meanwhile, we also calculate IPW-enhanced MAE loss for the swapped sample. To make robust extractors focus on robust information, we employ Y as the label of the swapped sample and the same weight for MAE loss as follows,

$$\begin{cases} \hat{\mathcal{L}}_{MAE} = |Y - \hat{Y}|, \\ \hat{\mathcal{L}}_{IPW} = \hat{\mathcal{L}}_{MAE} \cdot \frac{1}{P(x|Bias(x))+1}. \end{cases} \tag{17}$$

3.5.3 Training Objective. The overall learning objective of the model is performed by minimizing,

$$\mathcal{L} = \mathcal{L}_{IPW} + \lambda \mathcal{L}_{GMAE}^m + \beta(\hat{\mathcal{L}}_{IPW} + \lambda \hat{\mathcal{L}}_{GMAE}^m), \tag{18}$$

where λ and β are adjusted for weighting the importance of GMAE loss and swap, respectively. To ensure that robust modules and biased modules focus on robust attributes and biased attributes, respectively, $\mathcal{L}_{IPW}$ and $\hat{\mathcal{L}}_{IPW}$ are backpropagated to robust features fusion, robust linear, and robust extractor, $\mathcal{L}_{GMAE}^m$ and $\hat{\mathcal{L}}_{GMAE}^m$ are backpropagated to linear, biased linear, and biased extractor.

4 EXPERIMENTS

In this section, we conducted extensive experiments on two widely-used benchmark datasets, (*i.e.,* MOSI and MOSEI), to answer the following research questions.

- **RQ 1:** Does GEAR outperform state-of-the-art MSA baselines on the OOD testing sets?
- **RQ 2:** How does GEAR perform on the IID testing set?
- **RQ 3:** How does each component affect GEAR?
- **RQ 4:** How is the qualitative performance of GEAR?

4.1 Experimental Settings

4.1.1 Datasets. To demonstrate the effectiveness of our GEAR, we conducted extensive experiments on MOSI and MOSEI datasets, which are widely used in the MSA task.

- **MOSI** [43] is a publicly released MSA dataset. It collects 2,199 utterance-video clips of 93 monologue videos from YouTube platform[1], each of which is labeled with a continuous sentiment score ranging from -3 (strongly negative) to 3 (strongly positive).
- **MOSEI** [44] is an expanded version of MOSI. In MOSEI, 3,837 monologue videos are also collected from YouTube, involving 250 topics and 22,856 utterance-level labeled instances, each of which is also labeled with a continuous sentiment score ranging from -3 to 3.

The video clips in the two datasets consist of textual descriptions, acoustic tracks, and visual keyframes, which provide multimodal information to reflect the sentiment.

4.1.2 IID and OOD Settings. We removed the spurious correlations by discarding samples from the IID testing set to build the OOD testing sets. Due to the different bias types across modalities, we employed different strategies to construct the OOD datasets for different modalities. For the OOD Text set, following Sun *et al.* [29], we adopt the same method as described in this paper and refer to it for further details. We first obtained the distribution of word frequency in different sentiment categories. Then, we used the simulated annealing algorithm for dataset construction, which iteratively optimizes the OOD Text set to make the distribution of all words on different sentiment categories as same as possible, (*e.g.,* the word "movie" has an equal number of positive and negative categories). For the OOD Audio set and OOD Video set, the attributes of the audio and video are not recognizable by humans, and thus the distributions of each attribute are inaccessible. To mitigate this issue, we employ K-means clustering on hand-crafted features provided by [39] of the audio or video to obtain k clusters for each modality ($k = 100$ for audio and video). We assume that each cluster derived from K-means represents an attribute, and for the two OOD datasets, we ensure that all attributes appear equally in different sentiment categories by random sampling. For example, if there are four positive samples and six negative samples in a cluster, we randomly sample four samples from the six negative samples to ensure that the number of samples in each category is the same, (*e.g.,* make "blue background" appear equally in positive and negative categories). For OOD TAV set, we first obtained the distribution of word frequency and attributes in different sentiment categories as mentioned above. Following the existing work [29], we employed the simulated annealing algorithm to make the distribution of all words and all attributes on different sentiment categories as same as possible simultaneously.

4.1.3 Evaluation Tasks and Metric. Due to space limitations and following the latest work [29], we mainly focus on the metrics of accuracy and F1 score. Two distinct formulations have been considered in the past. The first is $negative/non\text{-}negative$ where $negative$ denotes a class with sentiment scores < 0 and $non\text{-}negative$ class with sentiment scores >= 0. Second, recent work [8] employs a more accurate formulation of $negative/positive$ classes where negative and positive classes are assigned with < 0 and > 0 sentiment scores, respectively. For the fair competition, we reported both $negative/non\text{-}negative$ and $negative/positive$ results. We converted the predicted score on the regression task into these two formulations, and then we used Accuracy and Weighted-F1 to measure the performance of the models.

4.1.4 Baselines. To evaluate the performance of GEAR, we employed the following methods for comparison.

- **MISA [8]:** The model projects modalities into model-specific and model-invariant vectors, capturing cross-modal interactions.
- **MAG-BERT [24]:** This baseline employs the nonverbal representations with sentimental polarity to shift lexical representations within the pretrained language model.
- **Self-MM [39]:** This model develops a unimodal sentiment label-generating module based on a self-supervised method to aid in learning modality-specific representations.
- **CENet [31]:** This baseline employs an attention-based gate to capture asynchronous emotion cues from unaligned data.
- **Cube-MLP [26]:** This baseline develops MLPs to mix features on three dimensions: sequence, modality, and channel.
- **CLUE [29]:** This framework captures the direct effect of textual modality via an extra text model and estimates the total effect by an MSA model.

[1] https://www.youtube.com.

Table 1: IID and OOD testing performance (%) comparison among different methods on MOSEI datasets. For *Acc-2* and *F1*, we reported results on both these metrics using the segmentation marker -/- where the left-side score is for *neg./nonneg.* while the right-side score is for *neg./pos*. The AVG (OOD) means the average result over four OOD sets. *Imp.* denotes the improvement of our model compared to the best-performing baseline. The best result is highlighted in bold and the second-best result is underlined. $^{\dagger}p < 0.05$ under McNemar's Test for accuracy improvement compared with all baselines.

Model	MOSEI											
	IID		OOD Text		OOD Audio		OOD Video		OOD TAV		AVG (OOD)	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*
MISA	82.91/85.47	83.11/85.28	79.06/81.23	79.00/81.06	80.09/81.73	80.03/81.89	80.64/81.70	80.57/81.86	77.32/79.60	77.26/79.80	79.28/81.07	79.22/81.15
MAG-BERT	83.14/84.61	83.27/84.40	79.15/80.06	79.00/79.86	80.40/81.07	80.32/81.23	80.65/80.84	80.55/80.99	77.83/78.58	77.73/78.79	79.51/80.14	79.40/80.22
Self-MM	**84.69**/85.35	**84.69**/85.10	80.53/80.60	80.31/80.39	81.53/81.47	81.42/81.63	81.40/81.01	81.28/81.15	78.75/78.65	78.62/78.85	80.55/80.43	80.41/80.51
CENet	82.77/85.13	83.00/84.97	78.61/80.73	78.56/80.57	80.32/81.81	80.26/81.96	80.89/81.70	80.81/81.85	77.29/79.28	77.22/79.48	79.28/80.88	79.21/80.97
CubeMLP	83.38/85.05	83.48/84.81	79.33/80.44	79.16/80.22	80.65/81.48	80.56/81.63	81.05/81.37	80.95/81.52	77.98/79.00	77.88/79.19	79.75/80.57	79.64/80.64
CLUE	83.99/85.06	83.90/85.26	80.91/81.09	**81.14**/81.33	81.03/80.72	81.16/80.57	81.54/80.95	81.70/80.82	78.48/78.48	78.65/78.06	80.49/80.31	80.66/80.20
GEAR†	84.06/**85.88**	84.30/**85.79**	**80.99/82.33**	80.97/**82.24**	**82.45/83.48**	**82.42/83.64**	**82.51/83.05**	**82.48/83.21**	**79.85/81.22**	**79.82/81.41**	**81.45/82.52**	**81.42/82.63**

Table 2: IID and OOD testing performance (%) comparison among different methods on MOSI datasets. The explanations of notations are the same as those in Table 1.

Model	MOSI											
	IID		OOD Text		OOD Audio		OOD Video		OOD TAV		AVG (OOD)	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*
MISA	82.66/84.2	82.65/84.24	78.42/79.53	78.41/79.54	81.24/83.14	81.22/83.13	81.33/82.91	81.31/82.91	76.97/79.19	76.95/79.20	79.49/81.19	79.47/81.20
MAG-BERT	83.04/84.81	83.00/84.82	79.20/80.76	79.18/80.76	**82.29/84.37**	**82.24/84.35**	81.59/83.62	81.57/83.61	77.55/79.92	77.51/79.92	80.16/82.17	80.13/82.16
Self-MM	82.95/84.81	82.87/84.79	79.20/81.17	79.17/81.17	81.36/83.66	81.29/83.61	81.66/83.76	81.63/83.74	78.23/80.54	78.18/80.52	80.11/82.28	80.07/82.26
CENet	82.27/84.09	82.14/84.03	79.79/81.28	79.71/81.22	80.68/82.88	80.58/82.80	81.19/83.26	81.12/83.22	78.04/80.33	77.97/80.30	79.93/81.94	79.85/81.89
CubeMLP	81.73/83.64	81.60/83.58	79.40/81.28	79.30/81.21	79.94/82.10	79.85/82.04	80.52/82.70	80.45/82.65	78.23/80.33	78.14/80.27	79.52/81.60	79.44/81.54
CLUE	82.79/84.68	82.87/84.71	78.97/81.07	79.03/81.09	80.56/82.80	80.60/82.83	81.32/83.55	81.35/83.56	77.38/79.79	77.43/79.81	79.56/81.80	79.60/81.82
GEAR†	**83.29/84.96**	**83.22/84.95**	**80.47/82.10**	**80.44/82.09**	82.22/84.31	82.16/84.27	**82.53/84.39**	**82.50/84.37**	**79.98/82.09**	**79.94/82.08**	**81.30/83.22**	**81.26/83.20**

4.1.5 Implementation Details. We implemented all baselines and our GEAR using Pytorch[2]. To optimize the parameters of the models, we adopted Adam [13] optimizer with a learning rate of 5*e*-5 for BERT and 1*e*-3 for other modules. Note that, as we needed well-disentangled robust and biased features for swap, we began to swap after certain epochs e_s. We employed a grid search strategy to identify the optimal hyperparameters for our model. Specifically, we set the batch size N to 32, the latent vector dimension d_s to 32, the head number U to 4, and the swap weight β to 0.3. Additionally, we set the swap epochs e_s to 8 and 11, and the GMAE weight λ to 10 and 18, for MOSEI and MOSI, respectively. Besides, we employed the early stopping strategy, which stops the training if the accuracy score/loss does not increase/decrease for 8 successive epochs. For all baselines, we used the grid search strategy to find the optimal parameter settings to achieve the best performance. For a fair comparison, we reported the average experimental results on accuracy and F1 score over three random seeds.

4.2 Model Comparison (RQ1 & RQ2)

We conducted experiments on the IID and OOD testing sets of MOSEI and MOSI datasets, respectively. As shown in Tables 1 and 2, we had the following observations. 1) The average accuracy on *neg./pos* was observed to increase by 1.46% on the MOSEI dataset and by 1.14% on the MOSI dataset. GEAR achieves clear margins over the prior methods on OOD sets, which demonstrates that GEAR has superior debiasing ability over existing methods. After conducting significant tests, $p < 0.05$ proves that our results are significant. 2) In particular, the improvements are most obvious in the OOD TAV testing set. The possible reason is that the OOD TAV testing set removes spurious correlations in all three modalities and the models' general debiasing ability removed bias in three modalities. 3) The improvement of GEAR on the IID testing set is smaller than on OOD testing sets, which indicates that the bias between the training set and the IID testing set is very small and GEAR's debiasing ability cannot be fully utilized. 4) On the unimodal and multimodal OOD testing sets, all methods perform worse than on IID testing sets. This demonstrates that methods indeed suffer from the spurious correlation for prediction. In spite of this, the performance decrease of GEAR is the least compared with that of baseline methods. And 5) CLUE performs well on the OOD Text set of the MOSEI dataset, for which we reasoned that CLUE is specifically designed for reducing spurious correlations between textual words and sentiment labels. However, on the other two unimodal testing sets and the multimodal testing set, CLUE performs worse than our GEAR. One reasonable explanation is that CLUE has limited ability in debiasing the acoustic and visual modalities.

4.3 Ablation Study (RQ3)

To verify the effectiveness of the main components in the proposed model, we conducted extensive ablation studies on OOD TAV, OOD Text, OOD Audio, and OOD Video datasets. We introduced several variants for analysis. (1) **w/o-IPW**. In this variant, we replaced $\mathcal{L}_{IPW}$ and $\hat{\mathcal{L}}_{IPW}$ with $\mathcal{L}_{MAE}$ and $\hat{\mathcal{L}}_{MAE}$ by removing the weight of MAE loss in Eq.(15) and Eq.(17). (2) **w/o-GMAE**. To verify the effect of the proposed GMAE loss, we trained the biased model

[2] https://pytorch.org.

Table 3: Ablation study results (%) for *neg./pos.* results of GEAR on OOD TAV testing sets of MOSEI and MOSI dataset. The best results are highlighted in boldface.

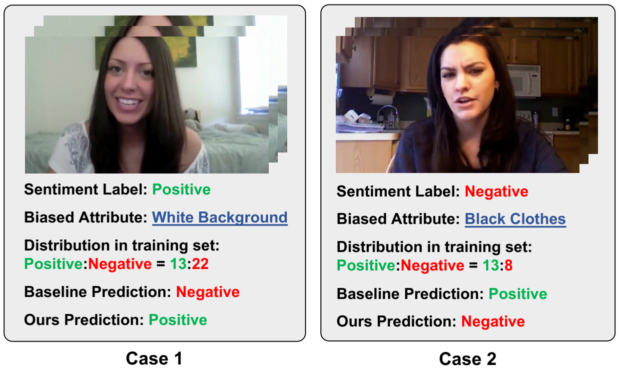
Model	MOSEI		MOSI	
	Acc	*F1*	*Acc*	*F1*
GEAR (OOD TAV)	**81.22**	**81.41**	**82.09**	**82.08**
w/o-IPW	78.84	79.03	81.37	81.29
w/o-GMAE	79.88	80.08	80.12	80.12
w/o-Swap	80.32	80.51	81.47	81.49
GEAR (OOD Text)	**82.33**	**82.24**	**82.10**	**82.09**
w/o-Text	81.22	81.06	81.07	81.05
GEAR (OOD Audio)	**83.48**	**83.64**	**84.31**	**84.27**
w/o-Audio	81.75	81.91	81.77	81.78
GEAR (OOD Video)	**83.05**	**83.21**	**84.39**	**84.37**
w/o-Video	82.06	82.21	84.04	84.02

with standard MAE loss instead of our proposed GMAE loss by replacing GMAE loss in Eq.(18) with MAE loss. (3) **w/o-Swap**. We set swap epochs e_s to an extremely large number to remove the swap operation. (4) **w/o-Text, w/o-Audio, and w/o-Video**. We removed the bias estimation of text, audio, or video modality, respectively. In other words, we calculated the bias weight only using two modalities in Eq.(10). These variants are tested on their corresponding OOD testing sets.

Table 3 shows the results of the ablation studies. First, after employing the model w/o-IPW, we can see the performance drops significantly. This phenomenon shows assigning a small weight to a sample with a strong bias for debiasing is indeed indispensable. Second, the setting of w/o-GMAE obtains worse results than the original model. This is because our proposed GMAE loss can train a model to be biased by amplifying the prejudice. However, the model trained with standard MAE loss not only exploits the biased attribute but also partially learns the robust attribute, which can hurt the debiasing ability of our overall algorithm by estimating the bias of each modality inaccurately. Third, GEAR w/o-Swap performs marginally better than almost all models in MOSEI and MOSI datasets, but not as well as GEAR. Compared to GEAR w/o-Swap, GEAR gains the relative improvements of 0.90% and 0.62% evaluated by *Acc* for the two datasets, respectively. This shows that the diversity of samples for disentanglement is crucial for optimal performance. Furthermore, we observed that after removing the bias estimation of any modality, the performance on the corresponding dataset drops significantly, which confirms that GEAR has superior general debiasing ability with each modality.

4.4 Case Study (RQ4)

To gain more insights into our model, we randomly selected four cases to explain how the spurious correlations in video modality affect the traditional MSA model and why GEAR is able to handle such spurious correlations in the testing set. We illustrated the binary (*neg./pos.*) results of Self-MM and GEAR on four testing samples from MOSI datasets in Figure 3 because the Self-MM shows the best overall performance (*i.e.,* AVG (OOD)). To learn the debiasing ability of our model, for each case, we recognized its biased attribute and counted the sentiment frequency of this biased attribute in the training dataset. There are 93 videos in the MOSI dataset, each of which has several clips. The clips in the same video have the same biased attributes, and thus, for convenience, our biased attributes statistics was at the video level. However, the dataset is labeled at the clip level. To obtain the video-level label, we selected the dominant label of all video clips as the video label. In detail, for a video, if there are more positive video clips than negative video clips, then the video is considered positive.



Sentiment Label: Positive

Biased Attribute: White Background

Distribution in training set: Positive:Negative = 13:22

Baseline Prediction: Negative

Ours Prediction: Positive

Case 1

Sentiment Label: Negative

Biased Attribute: Black Clothes

Distribution in training set: Positive:Negative = 13:8

Baseline Prediction: Positive

Ours Prediction: Negative

Case 2

Figure 3: Two testing cases of the baseline (*i.e.,* Self-MM) and GEAR on the *neg./pos.* results.

Taking Case 1 as an example, we can see that there is a woman with a smile and white background. By manual recognition, we found a total of 35 videos with white backgrounds from the training set, 13 with positive sentiment labels, and 22 with negative sentiment labels. And the white background is a kind of superficial feature that can be captured easily by models. Thus, the white background is a biased attribute that induces spurious correlations with sentiment labels. The traditional MSA model cannot reduce the spurious correlations, it thus predicts Case 1 as negative sentiment. Different from Self-MM, GEAR is able to handle this biased case by employing robust features such as the smile of the women for prediction. This demonstrates that GEAR has strong debiasing ability. A similar observation can be found in Case 2.

5 CONCLUSION AND FUTURE WORK

In this work, we first point out the spurious correlations between multimodal input data and sentiment labels and formulate a general debiasing multimodal sentiment analysis task. We design a novel general debiasing framework for multimodal sentiment analysis, GEAR for short, which strengthens the generalization ability via disentangling the robust features and bias features of textual, acoustic, and visual modalities, estimating the bias weight, and training with IPW-enhanced loss. Extensive experiments on two datasets (*i.e.,* MOSEI and MOSI) confirm the existence of spurious correlations and also indicate the superior generalization ability of GEAR on OOD testing sets. In future work, we will explore new strategies such as invariant feature learning to learn better disentangle biased features and facilitate bias estimation.

ACKNOWLEDGMENTS

This material is in part based on research sponsored by Defense Advanced Research Projects Agency (DARPA) under agreement number HR0011-22-2-0047.

REFERENCES

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. 2020. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*. PMLR, 528–539.

[3] Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 117–125.

[4] Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal End-to-End Sparse Model for Emotion Recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 5305–5316.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, 4171–4186.

[6] Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. 2022. Debiasing Graph Neural Networks via Learning Disentangled Causal Substructure. In *Advances in Neural Information Processing Systems*.

[7] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11, 665–673.

[8] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1122–1131.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8, 1735–1780.

[10] Youngkyu Hong and Eunho Yang. 2021. Unbiased classification through bias-contrastive and bias-balanced learning. In *Advances in Neural Information Processing Systems*. 26449–26461.

[11] Jun Hu, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2020. Multi-modal Attentive Graph Pooling Model for Community Question Answer Matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 3505–3513.

[12] Jun Hu, Shengsheng Qian, Quan Fang, Youze Wang, Quan Zhao, Huaiwen Zhang, and Changsheng Xu. 2021. Efficient Graph Deep Learning in TensorFlow with tf_geometric. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 3775–3778.

[13] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

[14] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. 2021. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*. 25123–25133.

[15] Xiaojiang Lei, Xueming Qian, and Guoshuai Zhao. 2016. Rating Prediction Based on Social Sentiment From Textual Reviews. In *IEEE Transactions on Multimedia*, Vol. 18. 1910–1921.

[16] Han Liu, Yinwei Wei, Jianhua Yin, and Liqiang Nie. 2023. HS-GCN: Hamming Spatial Graph Convolutional Networks for Recommendation. In *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35. IEEE, 5977–5990.

[17] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2247–2256.

[18] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the Conference on Artificial Intelligence*. AAAI, 164–172.

[19] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*. Omnipress, 807–814.

[20] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*. 20673–20684.

[21] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 271–278.

[22] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 973–982.

[23] Jiaxin Qi, Kaihua Tang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. 2022. Class is invariant to context and vice versa: on learning invariance for out-of-distribution generalization. In *European Conference on Computer Vision*. Springer, 92–109.

[24] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2359.

[25] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. In *Image and Vision Computing*, Vol. 65. Elsevier, 3–14.

[26] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 3722–3729.

[27] Teng Sun, Liqiang Jing, Yinwei Wei, Xuemeng Song, Zhiyong Cheng, and Liqiang Nie. 2023. Dual Consistency-enhanced Semi-supervised Sentiment Analysis towards COVID-19 Tweets. In *IEEE Transactions on Knowledge and Data Engineering*. IEEE, 1–13.

[28] Teng Sun, Chun Wang, Xuemeng Song, Fuli Feng, and Liqiang Nie. 2022. Response Generation by Jointly Modeling Personalized Linguistic Styles and Emotions. In *ACM Trans. Multim. Comput. Commun. Appl.* ACM, 52:1–52:20.

[29] Teng Sun, Wenjie Wang, Liqaing Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. 2022. Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 15–23.

[30] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*. ACL, 6558–6569.

[31] Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2022. Cross-modal Enhancement Network for Multimodal Sentiment Analysis. *IEEE Transactions on Multimedia*, 1–13.

[32] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. 2019. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*.

[33] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2022. User-controllable recommendation against filter bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1251–1261.

[34] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. 2022. Causal Representation Learning for Out-of-Distribution Recommendation. In *WWW*. 3562–3571.

[35] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the Conference on Artificial Intelligence*. AAAI, 7216–7223.

[36] Kaicheng Yang, Hua Xu, and Kai Gao. 2020. CM-BERT: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 521–528.

[37] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*.

[38] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Joint Visual-Textual Sentiment Analysis with Deep Neural Networks. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1071–1074.

[39] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the Conference on Artificial Intelligence*. AAAI, 10790–10797.

[40] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based Feature Reconstruction Network for Robust Multimodal Sentiment Analysis. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 4400–4407.

[41] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1103–1114.

[42] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the Conference on Artificial Intelligence*. AAAI.

[43] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *IEEE Intelligent Systems* 36, 6, 82–88.

[44] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2236–2246.

[45] Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Effective Sentiment-relevant Word Selection for Multi-modal Sentiment Analysis in Spoken Language. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 148–156.

[46] Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*. 8792–8802.