

# A Multibias-mitigated and Sentiment Knowledge Enriched Transformer for Debiasing in Multimodal Conversational Emotion Recognition

Jinglin Wang<sup>1\*</sup>, Fang Ma<sup>1\*</sup>, Yazhou Zhang<sup>2,3</sup>, Dawei Song<sup>1†</sup>

<sup>1</sup>Beijing Institute of Technology

<sup>2</sup>Zhengzhou University of Light Industry

<sup>3</sup>State Key Lab. for Novel Software Technology, Nanjing University

{jinglinwang, mfang, dwsong}@bit.edu.cn

yzhou\_zhang@bit.edu.cn

## Abstract

Multimodal emotion recognition in conversations (mERC) is an active research topic in natural language processing (NLP), which aims to predict human’s emotional states in communications of multiple modalities, e.g., natural language and facial gestures. Innumerable implicit prejudices and preconceptions fill human language and conversations, leading to the question of whether the current data-driven mERC approaches produce a biased error. For example, such approaches may offer higher emotional scores on the utterances by females than males. In addition, the existing debias models mainly focus on gender or race, where multibias mitigation is still an unexplored task in mERC. In this work, we take the first step to solve these issues by proposing a series of approaches to mitigate five typical kinds of bias in textual utterances (i.e., gender, age, race, religion and LGBTQ+) and visual representations (i.e., gender and age), followed by a Multibias-Mitigated and sentiment Knowledge Enriched bi-modal Transformer (MMKET). Comprehensive experimental results show the effectiveness of the proposed model and prove that the debias operation has a great impact on the classification performance for mERC. We hope our study will benefit the development of bias mitigation in mERC and related emotion studies.

## 1 Introduction

Whether people realize them or not, innumerable implicit prejudices and preconceptions fill human language, and are conveyed in almost all data sources, such as news, reviews and conversations (Misra et al., 2016). Such prejudices are known to hurt specific groups, and infringe their rights. For example, these two utterances, “older people are not interested in digital technology”, or

\*Jinlin Wang and Fang Ma contribute equally to this work and share the co-first authorship.

†Dawei Song is the corresponding author.



Figure 1: An illustrative example of gender bias in the multimodal dialogue dataset (MELD). There are three speakers in this conversation: Rachel, Joey, and Monica. In the bracket at the end of each utterance is the emotion type (e.g., joy, surprise, etc.) of it. The underlined sentences and the words highlighted in red express an apparent gender bias, and such bias is further amplified by Joey’s facial gestures.

“women are pleasant to look slim”, reveal the age and gender biases.

Recent research has shown that pre-trained word representations, e.g., word embeddings (in which each word is represented as a vector in the semantic space), tend to amplify the bias in the data (Kurita et al., 2019; Webster et al., 2020). The male names have been proved more likely to be associated with career-related terms than female names, by calculating the similarity between their embeddings (Caliskan et al., 2017). African-American names are also shown to be more likely to be associated with unpleasant terms than European-American names (Nadeem et al., 2020). Unconsciously learning such implicit biases from a dataset that is sampled from all kinds of data sources, leads to the fact that the learned models may further amplify the harmful bias (such as gender or race) when they make decisions (Goyal et al.,

2019; Srinivasan and Bisk, 2021). In addition, the biased error will propagate to downstream tasks. For example, coreference resolution systems exhibit a gender bias due to the use of biased word embeddings (Rudinger et al., 2018). Facial recognition applications have also been proved to perform worse for the inter-sectional group “darker females” than for either darker individuals or females (Buolamwini and Gebru, 2018).

In view that human language is multi-modal in nature, human bias also exists in multimodal conversations, e.g., textual and visual utterances. Figure 1 shows an example of the gender bias in a multimodal dialogue dataset<sup>1</sup>. Joey makes an association with a beautiful female nurse and expresses a significant smile when Rachel says “cute nurse”, but when Rachel says “they are male nurses”, he shows a disappointed-looking facial expression, although his textual response seems neutral.

Therefore, human bias naturally resides in the multimodal expression of emotions in conversations.

There has been a great body of literature in debiasing for computer vision (Buolamwini and Gebru, 2018) and pre-trained language models (Wang et al., 2020). However, the existing debias models mainly focus on only one kind of prejudice, e.g., gender or race, where multibias mitigation is still an unexplored task in mERC. This leaves us with a research question: *Whether the current data-driven multi-modal emotion recognition in conversations approaches produce a biased error or not?*

To answer this question, we first propose a series of approaches for debiasing multiple types of bias in multimodal (i.e., textual and visual) conversations. For textual utterances, we propose mitigating five types of bias, including gender, age, race, religion, and LGBTQ+ in word embedding.

For visual utterances, we first propose a subspace-projection-based debiasing approach to mitigate two typical visual biases, i.e., gender and age. It constructs a subspace for each type of visual bias and identifies the type of bias in the visual representation by projecting the representation into the corresponding subspace.

To incorporate the proposed multimodal debiasing methods into the mERC task that involves

conversational context modeling, cross-modality interactions capturing and the use of sentiment knowledge, we propose a **Multibiases Mitigated and sentiment Knowledge Enriched Transformer (MMKET)** as a unified framework. Specifically, it is a bimodal Transformer involving a contextual attention layer to capture the contextual interactions, a bimodal cross-attention layer to capture the cross-modal interactions, and a sentiment attention layer to enrich the debiased representation with sentiment knowledge.

Empirical evaluation has been carried out on two benchmark datasets, and the experimental results shows that the proposed multimodal debiasing methods can effectively mitigate the corresponding biases. We also prove that debiasing the representation of multimodal utterances has a remarkable impact on the performance of mERC models.

## 2 Generation of Bias

Models and algorithms have never independently created bias. Social bias is exhibited in multiple components of a NLP system, including the training corpus, pre-trained models (e.g., word embeddings), and algorithms themselves (Caliskan et al., 2017; Garg et al., 2018).

**Datasets: The Soil of Bias.** The dataset is the basis of model training. The bias in the dataset comes from the unbalanced dataset samples and biased labels. For example, gender bias manifests itself in training data that features more examples of men than women, an unbalanced dataset. In the process of label annotation, the annotators will transfer personal bias to the data, where the algorithm absorbs, thus produces a biased model. Sometimes, such bias is due to a lack of domain expertise (Plank et al., 2014) or preconceived notions and stereotypes held by the annotators (Sap et al., 2019).

**Bias in Word Embeddings.** Word embeddings are often trained from large and human-created corpora that contain multifarious biased raw data. Recent literature has demonstrated that gender bias is encoded in word embeddings (May et al., 2019). For example, Bolukbasi highlights that “programmer” is more closely associated with “man” while “homemaker” is more closely associated with “woman” in word2vec embeddings trained on the Google News dataset (Bolukbasi et al., 2016). And word embeddings connect medical doctors more frequently to male pronouns than female pro-

<sup>1</sup>Trigger Warning: This paper contains examples of biases and stereotypes seen in society and language representations. These examples may be potentially triggering and offensive. These examples are meant to bring light to and mitigate these biases, and it is not an endorsement.

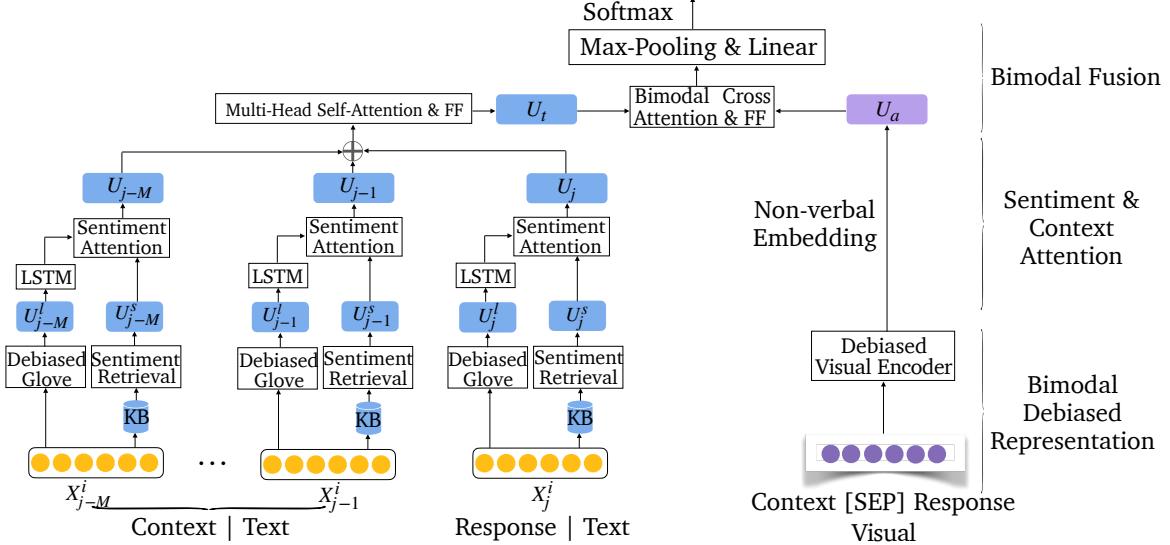


Figure 2: Overall architecture of our proposed Multibias-Mitigated and sentiment Knowledge enriched Transformer model.

Bias Type	Word Pairs
Gender	woman-man, girl-boy, she-he, mother-father, daughter-son, gal-guy, female-male
Race	slave-secondary, group-tribe, easy task-cake walk, master-primary
Age	young-old, health-disease, work-retirement, education-pension
Religion	Christian-Muslim, Christianity-Islam, Christ-Allah, Jesus-Muhammad
LGBTQ+	homosexuals-they, husband/wife-spouse, dad/father/mom/mother-parent

Table 1: The five pre-defined sets of word pairs, where the main difference between each pair of words captures the corresponding bias.

nouns (Caliskan et al., 2017). Furthermore, pre-trained word embeddings are often used without access to the original data. Social bias in word embeddings will propagate to downstream tasks, which can further amplify social bias. Studies show that machine translation systems tend to link occupations to their stereotypical gender, e.g., linking “doctor” to “he” and “nurse” to “she” (Prates et al., 2019).

### 3 Debiasing Methods

#### 3.1 Mitigating Multiple Biases in GloVe

The recent debiasing models (Bolukbasi et al., 2016; Wang et al., 2020) have only focused on removing gender bias in pre-trained word embeddings, particularly GloVe (Pennington et al., 2014), which has surfaced several social biases (Spliethöver and Wachsmuth, 2021). In this paper, we propose to mitigate five types of biases in GloVe embeddings, i.e., gender, race, religion, age, and LGBTQ+. Methodologically, we extend the existing Double-Hard Debias method, to multi-

ple types of bias.

**Hard Debias (Bolukbasi et al., 2016).** Hard Debias is a commonly adopted debiasing strategy in NLP. It projects a pre-trained word embedding vector into a subspace orthogonal to an inferred bias subspace (i.e., direction of a particular type of bias), which is constructed based on a set of pre-defined word pairs (e.g., *young* vs. *old*) characterizing the bias.

To extend it to multiple types of bias mitigation, we manually define a set of  $n$  characterizing word pairs for each type of bias based on typical data biases. Table 1 shows a range of representative examples.

**Double Hard Debias (Wang et al., 2020).** Wang et al. discovered that word frequency twists the bias direction, and proposed the Double-Hard Debias method. To find an intermediate subspace that can mitigate the effect of word frequency on the bias direction, Wang uses the clustering accuracy of highly biased words as an indicator to iteratively test the principal components of the word

embedding space.

Specifically, the Double Hard Debias method includes the following steps (taking age bias for example):

(a) Let  $W$  be the vocabulary of the word embeddings we aim to debias. Pick the top biased young and elderly words  $W_y, W_e \in W$ , according to the Cosine similarity of their embeddings to the age direction computed earlier.

(b) Calculate the principal components of  $W$  (measured by their projections onto the age direction) as the candidate frequency direction. Repeat steps (c)-(e) for each candidate dimension  $u_i$  respectively.

(c) The top biased word embeddings are mapped to an intermediate space orthogonal to  $u_i$  to mitigate the frequency-related bias:  $w'_{bias} = w_{bias} - (u_i^T w_{bias}) u_i$ , where  $w_{bias} \in W_y, W_e$ .

(d) Apply the Hard Debias method. The characterizing word pairs  $D_{t_1}, D_{t_2}, \dots, D_{t_n} \subset W$  are used here to subtract the bias projection from the top biased word embeddings:  $\hat{w}_{bias} = \text{Hard Debias}(w'_{bias})$ . The detailed steps of Hard Debias can be found in the original paper.

(e) Cluster the  $\hat{w}_{bias}$ , and then compute the corresponding accuracy and append to  $S_{debias}$ .

The purpose of debiasing is to make the top biased words (e.g., words about young and elderly) less separable. So the lower clustering accuracy in  $S_{debias}$ , the better debiasing effect that removing  $u_i$  has (i.e. the top biased words are mixed up). In other words, we filter out the  $u_i$  that causes the most significant decrease in the clustering accuracy and then remove it. Let  $j = \arg \min_i S_{debias}$ , we get the frequency-debiased word embeddings:  $w' = w - (u_j^T w) u_j$ , where  $w \in W$ . Then, apply the Hard Debias method to  $w'$  to obtain the output age-debiased word embedding:  $\hat{w} = \text{Hard Debias}(w')$ .

The algorithm operates on the five types of bias sequentially, i.e., the debiased word embedding of the first type serves as the input for the second, and so forth. Finally, we get the multibias-mitigated pre-trained word embeddings, which can be used in our proposed MMKET model.

### 3.2 Mitigating Multiple Biases in Visual Representation

Recent research shows that gender and age bias accounts for a noticeable portion of visual bias (Dzordowski et al., 2020). To mitigate them, we propose

two methods: Visual Hard Debias and Projection Debias methods. The Visual Hard Debias method can mitigate the superficial gender and age bias in visual representation. Then we devise the Projection Debias method to further mitigate finer-grained visual bias.

**Visual Hard Debias.** We assume that, for each type of visual bias, there is a pre-defined set of  $n$  image pairs  $V_1, V_2, \dots, V_n \in S$  (e.g., male-female or young-old), which represent the bias. The images are selected randomly from IMDB-WIKI (Rothe et al., 2018), a publicly available face image dataset with gender and age labels. Let  $\vec{v}$  denote an image's visual representation. Let  $u_i = \sum_{p \in V_i} \vec{p} / |V_i|$  be the mean of the image representations of  $V_i$  in the pre-defined image set. The visual bias subspace  $VB$  is spanned by the first  $k (\geq 1)$  eigen-vectors of  $VC$ , by applying Singular Value Decomposition (SVD) on it.

$$VC := \sum_{i=1}^m \sum_{v \in V_i} (\vec{v} - u_i)^T (\vec{v} - u_i) / |V_i| \quad (1)$$

Here the  $k$  is set to 1. As a result, the bias subspace  $VB$  becomes a bias direction  $\overrightarrow{VB}$ . After getting the visual bias subspace, each image representation  $\vec{v}$  is debiased through:  $\tilde{v} = \vec{v} - (\overrightarrow{VB}^T \cdot \vec{v}) \overrightarrow{VB}$ .

**Projection Debias.** In order to further mitigate the finer-grained gender and age bias in the image representation, we propose a new visual debias method, namely Projection Debias. Specifically, it projects the image representation twice into the bias subspaces (e.g., *male* vs. *female*, *young* vs. *old*) respectively. By subtracting the two projections from the original visual representation, we get the final debiased representation. Figure 3 shows the Projection Debias method on gender bias.

First we use the IMDB-WIKI to define four sets of images  $U_i$ , where  $i \in [1, 2, 3, 4]$ , corresponding to the female, male, young and old respectively. Then we compute the bias subspace as:

$$B_i = \vec{u}_i^T \otimes \vec{u}_i \quad (2)$$

where  $i \in [1, 2, 3, 4]$ ,  $\vec{u}_i$  is the first principal component of  $U_i$  computed through Principal Component Analysis (PCA). T means the transpose operation, and  $\otimes$  means outer product operation. Then we can get the corresponding projection-debiased visual

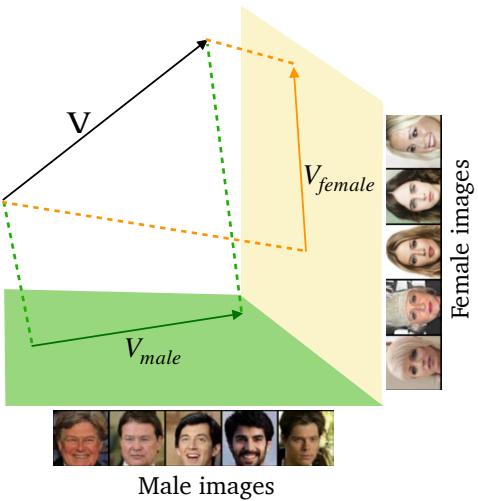


Figure 3: Projection Debias.

representation through:

$$\hat{v} = \tilde{v} - \sum_{i=1}^4 (B_i \times \tilde{v}) \quad (3)$$

Then we get the two-step debiased visual representation  $\hat{v}$  for an image.

## 4 The Proposed MMKET Model

We outline the Multibias-mitigated and Sentiment Knowledge Enriched Transformer (MMKET) model (c.f. Figure 2). In the proposed framework, we apply the Transformer (Vaswani et al., 2017) to leverage the debiased contextual and multimodal (text and visual) clues to predict the emotions of the target utterance, due to its ability to capture the context and fast computation. The main ideas are: (1) a multi-modal encoder to create textual and visual representations of contexts and responses, including debiased word embedding (GloVe) and debiased visual representation from the pre-trained EfficientNet Network (Tan and Le, 2019). (2) The text representation is enriched by sentiment knowledge. (3) The context-aware attention mechanism is proposed to effectively incorporate conversational context. (4) The text representation and non-verbal embedding are forwarded through a self-attention layer and a feed-forward sublayer to perform multimodal fusion.

### 4.1 Task Definition

Suppose our dataset has  $N$  data-points, we can represent the  $i$ -th data as  $\{U_j^i, Y_j^i\}$ ,  $U_j^i = (X_j^i, V_j^i)$ , where  $i \in \{1, 2, \dots, N\}$ ,  $j \in \{1, 2, \dots, N_i\}$ , which

is a collection of  $\{\text{utterance}, \text{label}\}$  pairs,  $N$  denotes the number of conversations, and  $N_i$  denotes the number of utterances in the  $i$ -th conversation. Each utterance consists of two modalities: text (X), video (V). We align the visual features with their corresponding tokens in the text modality. Therefore, both two modalities have the same length. Given an utterance, our task is to predict its emotion label. The objective of the task is to maximize the following function:

$$\Theta = \prod_{i=1}^N \prod_{j=1}^{N_i} p(Y_j^i | U_j^i, U_{j-1}^i, \dots, U_1^i; \theta) \quad (4)$$

where  $U_{j-1}^i, \dots, U_1^i$  denote contextual utterances and  $\theta$  denotes the model parameters set. We denote the number of contextual utterances as  $M$ .

### 4.2 Bimodal Encoder Layer

We extract textual and visual features via the bimodal encoder respectively. For text representation, we use a debiased word embedding layer to convert each token  $t$  in  $X^i$  into a vector representation  $\vec{t} \in \mathbb{R}^d$ , where  $d$  denotes the size of word embedding. Moreover, the debiased GloVe embeddings (through our debiasing methods presented in Sec. 3.1) are used for initialization in the word embedding layer. Let

$$\vec{t} = \text{Embed}(t) \quad (5)$$

as described in the previous part, we use a sentiment embedding layer to convert each token  $t$  in the utterance into a corresponding sentiment features score  $\vec{S}_i$  as an additional information source vector. The resulting textual embeddings are fed into the Transformer encoders to further refine textual representation.

For the visual representation, each input video clip is scaled to  $480 \times 360$ , and the pre-trained EfficientNet (Tan and Le, 2019) is used to extract the features. The Transformer encoders are used to learn the visual representations.

### 4.3 Sentiment Knowledge Attention

In philosophy and psychology, sentiment and emotion are closely related, corresponding to internal and external human affection (Evans, 2002). Sentiment refers to human's subjective experience and mental attitude, which involves long-term and deep human cognition (Dolan, 2002). Therefore, we hypothesise that the sentiment knowledge will help

the task of emotion recognition. Correspondingly, we propose a sentiment knowledge attention mechanism to capture and counterpoise the sentiment representation for each token. Specifically, a gated unit is used to combine the sentiment representation and the original utterance representation.

In our model, we use a commonsense emotion lexicon NRC\_VAD (Mohammad, 2018) as the sentiment knowledge source. The NRC Valence, Arousal, and Dominance (VAD) lexicon include a list of more than 20,000 English words and their valence, arousal, and dominance scores.

For a given word and a dimension (V/A/D), the scores range from 0 to 1.

In general, for each word token  $t$  in  $X_j^i$ , we only retrieve its valence values from the NRC\_VAD dictionary, which is the ‘positive-negative’ dimension. The final sentiment knowledge representation for each text utterance  $X_j^i$  is a list of valence scores:  $[V(t_1), V(t_2), \dots, V(t_n)]$ . The valence scores of tokens that are not included in NRC\_VAD are set to 0.5. The sentiment knowledge representation of each text utterance will be used to enrich the text representation and serve the multi-bias mitigation. The gate value  $g_i$  for each token  $x_i$  is calculated as follows:

$$g_i = \sigma(W_g h_i + b_g) \quad (6)$$

where  $h_i$  is the hidden vector of token  $x_i$  from the previous *lstm* layer,  $W_g$  is a learnable linear transformation and  $b_g$  is the bias. Then the attention output  $\vec{T}$  is calculated as a weighted combination of sentiment enriched and original attention scores:

$$\vec{T}_i = g_i \vec{t}_i + (1 - g_i) \vec{S}_i \vec{t}_i \quad (7)$$

#### 4.4 Bimodal Cross Attention

We use a bimodal cross attention layer (Hasan et al., 2021), which is a multi-head self-attention mechanism, to learn the joint representation of  $U_l$  and  $U_v$ ,  $U_l = \vec{T}_i$ , where  $U_l$  represents sentiment-enriched textual representation and  $U_v$  denotes sentiment-enriched visual representation.

Specifically, we create corresponding sets of queries ( $Q_l, Q_v$ ), keys ( $K_l, K_v$ ), and values ( $V_l, V_v$ ) to learn the interaction between textual and visual modalities ( $U_l, U_v$ ). The modal representation and query set is attached to a multi-head cross attention layer. We also add the normalization layer and residual connections layer after each cross attention layer. Let

$$M_{l,v} = \text{BimodalCrossAttention}(U_l, U_v) \quad (8)$$

#### 4.5 Classification

The bimodal fusion representation is gained from the bimodal cross attention layer, which is shown in Eq. 8. We then add a maxpooling layer to extract the most salient features across the time dimension and yield a one-dimensional vector. Let

$$M_{l,v} = \text{MaxPooling}(M_{l,v}) \quad (9)$$

$$P = \text{softmax}(M_{l,v} W + b) \quad (10)$$

where  $P$  represents the output probability,  $W \in \mathbb{R}^{d \times l}$  and  $b \in \mathbb{R}^l$  denote parameters,  $l$  denotes the number of classes.

### 5 Experiments

#### 5.1 Datasets

**IEMOCAP** (Busso et al., 2008): A multimodal dataset containing emotional dialogues. Each video contains a single dynamic dialogue, segmented into utterances. The emotion labels of utterances include neutral, happiness, sadness, anger, frustrated, and excited.

**MELD** (Poria et al., 2019): A dataset of TV show scripts collected from Friends, which is a multimodal emotion classification dataset. The emotion labels of the dataset include happiness, surprise, sadness, anger, disgust, and fear.

Both datasets contain textual, visual, and acoustic information for every utterance. We only focus on the textual and visual modalities in this work. Table 2 shows the statistics of the datasets. In all our experiments, 300-dimensional GloVe is leveraged to initialize word embeddings, pre-trained EfficientNet network is used to extract the corresponding feature vectors of images. The dimensionality of hidden states is set to 300. We use adam as an optimizer with a learning rate of 0.0001 and train. The coefficient of L2 regularization is  $10^{-5}$ . The batch size is 64. The network is subjected to regularization in the form of Dropout.

#### 5.2 Evaluation Metrics

**Debiasing.** We use k-Means clustering to verify the effectiveness of the debiasing methods. For each type of bias, we take the top 100/500/1000 of the original GloVe embeddings and 100/300/500 of the visual features by calculating their cosine similarity with the specific bias directions. Then, we cluster them into two groups and compute the

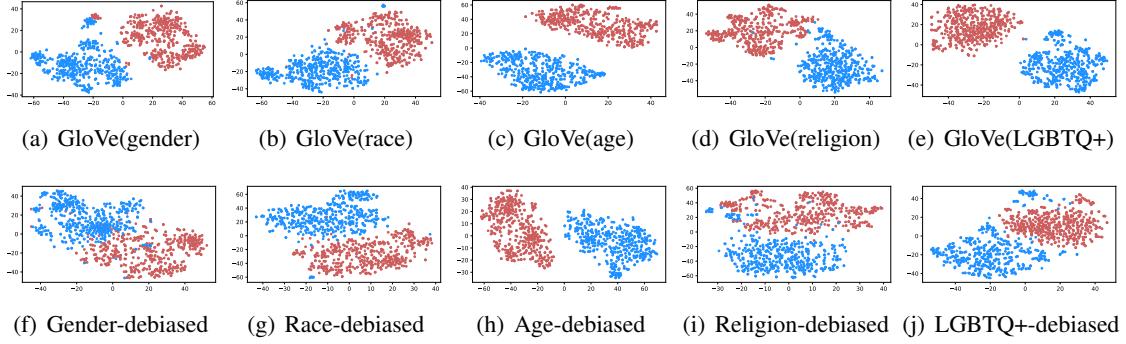


Figure 4: TSNE visualization of clustering top 500 most biased embeddings (a-e) and their debiased embeddings (f-j).

Dataset	# dialogues.	# utterances.
IEMOCAP	train	100
	dev	20
	test	31
MELD	train	1039
	dev	114
	test	280

Table 2: The data statistics of IEMOCAP and MELD.

alignment accuracy for the bias. To visualize the difference, we applied tSNE projection on word embeddings and the image features.

**Our Proposed MMKET Model.** We evaluate our proposed MMKET model on IEMOCAP and MELD, and adopt the F1 score on the test set as our evaluation metric.

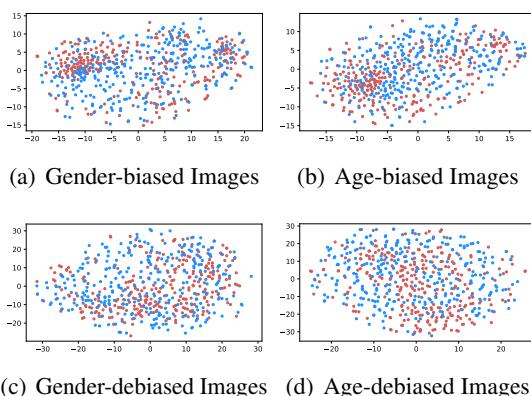


Figure 5: tSNE visualization of top 300 most biased images (a-b) and their debiased counterparts (c-d).

Embeddings	Top100	Top500	Top1000
GloVe	100.0	99.9	99.7
Gender-debiased GloVe	86.0	68.7	<b>55.3</b>
GloVe	100.0	100.0	99.3
Age-debiased GloVe	100.0	99.2	<b>98.9</b>
GloVe	86.5	75.3	54.5
Race-debiased GloVe	86.5	75.1	<b>54.4</b>
GloVe	99.5	95.7	96.6
Religion-debiased GloVe	97.0	86.8	<b>81.6</b>
GloVe	100.0	99.7	99.3
LGBTQ+-debiased GloVe	94.5	<b>90.7</b>	91.1

Table 3: K-Means clustering accuracy (%) of top 100/500/1000 biased words.

Mitigated Bias	IEMOCAP	MELD
None	57.11	53.93
Gender	56.22	53.22
Race	56.63	53.41
Age	56.76	53.69
Religion	56.09	53.14
LGBTQ+	56.89	53.20
<b>5 Biases</b>	<b>55.85</b>	<b>52.86</b>

Table 4:  $T + V$  results (%) on IEMOCAP and MELD dataset, which is mitigated the specific bias.

## 6 Results and Analysis

### 6.1 Effects of Debiasing

**Mitigating Multiple Biases in GloVe.** Table 3 shows the result of K-Means clustering on the original GloVe and the debiased ones. Lower accuracy means fewer bias cues can be learned. The accuracy appears to decrease after the debiasing operation, suggesting the debias method works effectively in embeddings. More intuitively, in the upper row of Figure 4, word embeddings are divided into two clear parts. In the lower row, the two parts have mixed up, though different biases have varied effects. Among the five proposed biases, gender and religious bias were mitigated most. LGBTQ+ bias also reduced, while racial and age bias did not

decrease significantly. We speculate that the racial bias is more implicit in textual data given that the accuracy of the original GloVe is already close to 50. As for the age bias, we consider the bias words like “old” are widely used as unbiased meanings, i.e. “an old tree”, “a seven-year-old boy”, which decreased the effect of debiasing. Mitigating the racial and age bias will be left to our future work.

**Mitigating Multiple Biases in Visual Representation.** Table 5 shows the clustering result of biased images. As shown in Figure 5, the visual representation of images from IMDB-WIKI are projected into a 2D space. Our proposed debiasing methods mix up the images to a noticeable extent, indicating that gender and age bias are mitigated in image representation.

Visual Representation	Top100	Top300	Top500
Gender-biased Images	74.3	70.8	64.5
Gender-debiased Images	61.0	59.3	<b>53.3</b>
Age-biased Images	67.2	62.3	59.6
Age-debiased Images	60.7	53.8	<b>52.5</b>

Table 5: K-Means clustering accuracy (%) of top 100/300/500 biased images. Lower accuracy means less bias cues.

## 6.2 Debiased mERC Results

We make the first step to explore the role of bias plays in mERC tasks. Human emotions contain prejudice, so removing the bias will decrease the emotion classification accuracy, which can explain the results in Table 4 and Table 6. Compared to the single modal results (Table 4), our MMKET model makes full use of the rich information in the Bimodal data and the connection between them, which greatly improves the performance of the algorithm.

Mitigated Bias		IEMOCAP	MELD
Text	Visual		
None	None	58.29	56.35
	Gender&Age	57.61	55.09
5 Biases	None	57.56	55.64
	Gender&Age	56.23	54.25

Table 6:  $T + V$  results (%) of mERC on IEMOCAP MELD dataset, by mitigating different types of bias.

## 6.3 Ablation Studies

To further investigate how the sentiment knowledge affects the debias method and mERC, we conduct extensive ablation experiments with the weight of

Dataset	0	0.3	0.5	0.7
MELD	55.93	<b>56.35</b> <sub>+0.42</sub>	56.24 <sub>+0.31</sub>	56.09 <sub>+0.16</sub>
IEMOCAP	57.60	<b>58.29</b> <sub>+0.69</sub>	57.96 <sub>+0.36</sub>	57.22 <sub>-0.37</sub>
Debiased-MELD	54.14	<b>54.42</b> <sub>+0.28</sub>	54.31 <sub>+0.17</sub>	54.25 <sub>+0.11</sub>
Debiased-IEMOCAP	56.37	<b>56.57</b> <sub>+0.20</sub>	56.40 <sub>+0.03</sub>	56.38 <sub>+0.01</sub>

Table 7: Analysis of the weight of sentiment knowledge from 0 to 0.7.

sentiment knowledge of different values, whose results are included in Table 7. The sentiment knowledge improves model performance significantly, but less on the debiased model. One possible reason is that biases themselves imply the emotions of humans, so mitigating biases will reduce the effect of sentiment knowledge.

## 7 Conclusion

In this work, we extend the types of bias in the embedding level (e.g., gender, age, race, religion, and LGBTQ+) and innovatively propose the Projection Debias to mitigate gender and age bias in visual representation. We also present a Multibias-mitigated and Sentiment Knowledge Enriched Transformer (MMKET), taking the first step to explore how the debiasing operation affects the algorithm in multimodal emotion recognition in conversation (mERC). We conduct extensive experiments to show the effectiveness of the proposed model and prove that debias operation and sentiment knowledge has a great impact on the classification performance for the task of mERC. Due to the difference of the biases, the effect of debiasing also varies, which requires further research. Our model also has a few limitations. For example, we only select to mitigate two typical visual biases, while other types of bias are ignored. Such efforts will be left to our future work. We hope our study will benefit the development of bias mitigation in mERC and other emotion studies.

## Acknowledgements

This research was supported in part by Natural Science Foundation of Beijing (grant number: 4222036) and Huawei Technologies (grant number: TC20201228005). This work was supported by National Science Foundation of China under grant No. 62006212, the fund of State Key Lab. for Novel Software Technology in Nanjing University (grant No. KFKT2021B41), and the Industrial Science and Technology Research Project of Henan Province (grant No. 222102210031).

## References

- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*.
- Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*.
- Raymond J Dolan. 2002. Emotion, cognition, and behavior. *science*, 298(5596):1191–1194.
- Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*.
- Dylan Evans. 2002. *Emotion: The science of sentiment*. Oxford University Press, USA.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*.
- Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*.
- Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multi-modal humor.
- Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.
- Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proc. of ACL*.
- Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proc. of CVPR*, pages 2930–2939.
- Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proc. of ACL*.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proc. of ACL*.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. of ACL*.
- Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*.
- Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proc. of ACL*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. of ACL*.
- Maximilian Spliethöver and Henning Wachsmuth. 2021. Bias silhouette analysis: Towards assessing the quality of bias metrics for word embedding models. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*.
- Tejas Srinivasan and Yonatan Bisk. 2021. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *arXiv preprint arXiv:2104.08666*.
- Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. of ICML*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proc. of ACL*.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.