Mitigating Gender Bias for Neural Dialogue Generation with Adversarial Learning

Haochen Liu[1], Wentao Wang[1], Yiqi Wang[1], Hui Liu[1], Zitao Liu[2]*, Jiliang Tang[1]

[1] *Michigan State University, East Lansing, MI, USA*

[2] *TAL Education Group, Beijing, China*

{liuhaoc1,wangw116,wangy206,liuhui7}@msu.edu; liuzitao@100tal.com; tangjili@msu.edu

Abstract

Dialogue systems play an increasingly important role in various aspects of our daily life. It is evident from recent research that dialogue systems trained on human conversation data are biased. In particular, they can produce responses that reflect people's gender prejudice. Many debiasing methods have been developed for various natural language processing tasks, such as word embedding. However, they are not directly applicable to dialogue systems because they are likely to force dialogue models to generate similar responses for different genders. This greatly degrades the diversity of the generated responses and immensely hurts the performance of the dialogue models. In this paper, we propose a novel adversarial learning framework **Debiased-Chat** to train dialogue models free from gender bias while keeping their performance. Extensive experiments on two real-world conversation datasets show that our framework significantly reduces gender bias in dialogue models while maintaining the response quality.

1 Introduction

The elimination of discrimination is an important issue that our modern-day society is facing. Learning from humans' behaviors, machine learning algorithms have been proven to inherit the prejudices from humans (Mehrabi et al., 2019). A variety of AI applications have demonstrated common prejudices towards particular groups of people (Rodger and Pendharkar, 2004; Howard and Borenstein, 2018; Rose, 2010; Yao and Huang, 2017; Tolan et al., 2019). It is evident from recent research that learning-based dialogue systems also suffer from discrimination problems (Liu et al., 2019a; Dinan et al., 2019). Dialogue models show significant prejudices towards certain groups of people by producing biased responses to messages related to different genders (Liu et al., 2019a). A biased dialogue system will produce improper speeches, which can bring in bad experiences to users or even cause negative social impacts (Wolf et al., 2017; Liu et al., 2019b, 2020). Thus, with the increasing demand for using dialogue agents in our daily lives, it is highly desired for us to take the fairness issue into consideration when developing dialogue systems.

The gender bias[1] in dialogues comes from different dimensions – the gender of the person that speakers are talking about (speaking-about), and the gender of the speaker (speaking-as) and the addressee (speaking-to) (Dinan et al., 2020). In this work, we focus on mitigating the gender bias of dialogue systems in the speaking-about dimension. It is the most common format of gender bias in dialogues which exists under both speaker-given dialogue scenario, where the personas of the speaker or the addressee are known (Li et al., 2016; Zhang et al., 2018), and speaker-agnostic dialogue scenario, where the information of the speakers is unknown. Given messages with the same content for different genders, dialogue models could produce biased responses, which have been measured in terms of their politeness and sentiment, as well as the existence of biased words (Liu et al., 2019a). Table 1 shows one example from a generative dialogue model trained on the Twitter dialogue corpus. When we change the words in the messages from "he" to "she", the responses produced by the dialogue model are quite different. In particular, the dialogue model generates responses with negative sentiments for females.

There are debiasing methods in natural language processing such as data augmentation (Dinan et al.,

* The corresponding author: Zitao Liu

[1] We focus on two genders (i.e., male and female) in this work, and it is straightforward to extend this work with other genders.

Table 1: Examples of gender bias in dialogue systems.

Message	Response
Really wishes **he** could take at least one step on this husker floor...	I'm sure he's going to be a great guest.
Really wishes **she** could take at least one step on this husker floor...	I'm sure she's a little jealous.

2019) and word embeddings regularization (Liu et al., 2019a). Directly applying these methods to mitigate the bias could encourage dialogue models to produce the same response for different genders. This strategy can lead to producing unreasonable responses such as "he gave birth to a baby" and also reduce the diversity of the generated responses. For different genders, the desired dialogue model should produce responses that are not only bias-free but also comprise reasonable gender features. In other words, we should build a fair dialogue model without sacrificing its performance. To achieve this goal, we face three key challenges. First, dialogues contain various gender-related contents. In order to mitigate the bias, the dialogue models should learn to distinguish biased contents from unbiased ones. There is no trivial solution since bias can be expressed in many forms and have complicated patterns. Second, even if the first challenge is addressed, eliminating biased contents in responses by the dialogue models remains hard. Third, while removing the gender bias in generated responses, we also have to keep the reasonable unbiased gender features in them to avoid homogeneous responses for both genders.

In this paper, we propose a novel framework **Debiased-Chat** to train bias-free generative dialogue models. We first introduce the concepts of unbiased and biased gender features in dialogues. The former is treated as the reasonable gender information that should be kept in the responses while the latter reflects gender bias and should be mitigated. Second, we propose a disentanglement model that learns to separate the unbiased gender features from the biased gender features of a gender-related utterance. Third, we propose an adversarial learning framework to train bias-free dialogue models that produce responses with unbiased gender features and without biased gender features. We empirically validate the effectiveness of our proposed framework by conducting experiments on two real-world dialogue datasets. Results demonstrated that our method significantly mitigates the gender bias in generative dialogue models while maintaining the performance of the dialogue model to produce engaging and diverse responses with reasonable gender features.

2 The Proposed Framework

In this section, we detail the proposed framework. Note that in this work, we focus on the classical generative Seq2Seq dialogue model for single-turn dialogue generation while we leave other settings such as the multi-turn case as future work. We first define two key concepts. We refer to the reasonable and fair gender features in a response as the **unbiased gender features** of the response. They include gendered terms and words or phrases specially used to describe one gender. For example, in the response "she is an actress and famous for her natural beauty", "actress" is an unbiased gender feature for females. We call the unreasonable and discriminatory gender features in a response as the **biased gender features**. According to the definition of the bias in dialogue models in (Liu et al., 2019a), any offensive, sentimental expressions and biased words correlated with one gender are considered as its biased gender features. For instance, given the same context with different genders as shown in Table 1, for the response to females, "I'm sure she's a little jealous", the word "jealous" is a biased gender feature under the context.

2.1 An Overview

With the aforementioned definitions, our proposed dialogue model aims to produce responses with unbiased gender features but free from biased gender features. Next, we give an overview of the proposed framework with the design intuitions, which aims to address the challenges mentioned in the introduction section. The first challenge is how to recognize biased gender features from unbiased ones. Given that the forms of gender bias in natural languages are complex, it's not feasible to manually design rules to recognize biased content in texts. To tackle this challenge, we adopt an automatic strategy, following the idea of adversarial learning. We propose a disentanglement model (right of Figure 1) to learn to separate the unbiased gender features $\mathbf{f}^{(\mathbf{u})}$ and the semantic features $\mathbf{f}^{(\mathbf{s})}$ of a gender-related utterance. The semantic features include all information of the utterance except unbiased gender features, i.e., the content information and possibly biased gender features. We collect a set of unbiased gendered utterances and train the

disentanglement model with objectives that the extracted unbiased gender features can be used for a discriminator to infer the gender of the utterance while the rest semantic features cannot. Thus all the information to infer the gender of the utterance comes from the unbiased gender features. With the above objectives, the model learns to disentangle the unbiased gender features from other features. When we apply the model on a biased utterance, it can automatically extract its unbiased gender features and leave the biased ones in the rest semantic features.

To address the second challenge (remove biased gender features in dialogues) and the third challenge (reserve unbiased gender features in dialogues), we propose our framework to train bias-free dialogue models (left of Figure 1). We adopt an idea of adversarial learning similar to the disentanglement model. Given a response from the dialogue model, its two disentangled feature vectors are fed into two discriminators D_1 and D_2 respectively, to predict the gender of the dialogue[2]. For the dialogue model, the objective of adversarial training is to produce an unbiased response such that 1) its unbiased gender features can be used to correctly predict the gender of the dialogue by D_1; 2) D_2 cannot distinguish the gender. The intuition of the design is below. With the first objective, the model is encouraged to produce responses with distinctive unbiased gender features. Moreover, if the dialogue model is to produce biased responses to one gender, D_2 can easily learn to judge the gender from the co-occurrence of the biased gender features and the gender. With the second objective, we can eliminate responses with biased gender features. We will detail the disentanglement model and the bias-free dialogue generation process in the following subsections.

2.2 The Disentanglement Model

2.2.1 Unbiased Gendered Utterance Corpus

Given the dialogue corpus $\mathbf{D}$, we collect all the gender-related utterances from it. Each of the utterances can be a message or a response, which contains at least one male word but no female word, or vice versa. Then, we filter out all utterances that could be biased. Following the bias measurements in (Liu et al., 2019a), we remove all the utterances which 1) are offensive, or 2) show strong positive or negative sentiment polarity, or 3) contain career or family words. The rest utterances form an **Unbiased Gendered Utterance Corpus** $\mathbf{U} = \{(U_i, g_i)\}_{i=1}^{M}$, where U_i is the i-th utterance and g_i is its gender label. The corpus is used to train the disentanglement model.

[2]We assume that the message and the response of a single-turn dialogue are always related to the same gender. We call it the gender of the dialogue.

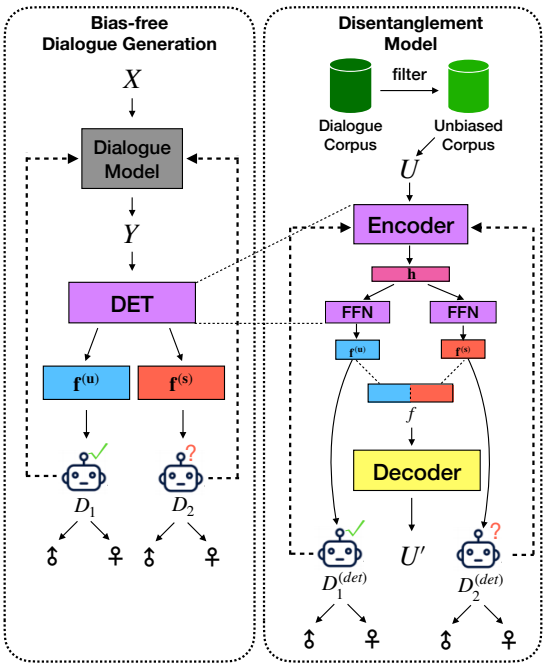

Figure 1: An overview of our proposed framework. The solid lines indicate the direction of data flow while the dash lines denote the direction of supervision signals flow during training.

2.2.2 Model Design

The illustration of the disentanglement model is shown on the right of Figure 1.

Autoencoder. We adopt an autoencoder as the disentanglement model, in which both the encoder and the decoder are implemented using recurrent neural networks (RNN) with gated recurrent unit (GRU) cells (Cho et al., 2014). The encoder learns to encode an utterance U into a latent vector $\mathbf{h} \in \mathbb{R}^d$. The latent vector $\mathbf{h}$ is then mapped into the space of unbiased gender features $\mathbb{R}^u$ and the space of the semantic features $\mathbb{R}^s$ by two 1-layer feedforward networks respectively, to get the unbiased gender features $\mathbf{f^{(u)}}$ and the semantic features $\mathbf{f^{(s)}}$. The concatenation of the unbiased gender and the semantic features $\mathbf{f} = [\mathbf{f^{(u)}} : \mathbf{f^{(s)}}]$ is then fed into the decoder to reconstruct the original utterance U.

Discriminators. In the autoencoder, to disen-

tangle the latent representation $\mathbf{h}$ into the unbiased gender features $\mathbf{f^{(u)}}$ and the semantic features $\mathbf{f^{(s)}}$, we take advantage of the idea of adversarial learning. We first train two discriminators $D_1^{(det)}$ and $D_2^{(det)}$ to distinguish whether the utterance U is related to male or female based on the unbiased gender features $\mathbf{f^{(u)}}$ and the semantic features $\mathbf{f^{(s)}}$, respectively. The discriminators are implemented via one-layer feedforward neural networks, which predict the probability distribution of the genders $\mathbf{p^{(u)}} \in \mathbb{R}^2$ and $\mathbf{p^{(s)}} \in \mathbb{R}^2$ based on $\mathbf{f^{(u)}}$ and $\mathbf{f^{(s)}}$, respectively.

Adversarial Training. In the adversarial training process, we hope that the discriminator $D_1^{(det)}$ can make predictions correctly, while $D_2^{(det)}$ cannot. The outputs of the discriminators are used as signals to train the disentanglement model so that it will assign the gender-related information into the unbiased gender features $\mathbf{f^{(u)}}$ while ensuring that the semantic features $\mathbf{f^{(s)}}$ do not include any gender information. Thus, we define two losses in terms of the discriminators $D_1^{(det)}$ and $D_2^{(det)}$ as:

$$L_{D_1^{(det)}} = -(\mathbb{I}\{g=0\} \log \mathbf{p}_0^{(u)} + \mathbb{I}\{g=1\} \log \mathbf{p}_1^{(u)}) \quad (1)$$

$$L_{D_2^{(det)}} = -(\mathbf{p}_0^{(s)} \log \mathbf{p}_0^{(s)} + \mathbf{p}_1^{(s)} \log \mathbf{p}_1^{(s)}) \quad (2)$$

where g is the gender label of the utterance and $\mathbf{p}_i$ is the i-th element of $\mathbf{p}$. $L_{D_1^{(det)}}$ is the cross-entropy loss function on $\mathbf{p^{(u)}}$. Minimizing $L_{D_1^{(det)}}$ will force $D_1^{(det)}$ to make correct predictions. $L_{D_2^{(det)}}$ is the entropy of the predicted distribution $\mathbf{p^{(s)}}$. Minimizing it makes $\mathbf{p^{(s)}}$ close to an even distribution, so that $D_2^{(det)}$ tends to make random predictions.

To further ensure that only $\mathbf{f^{(s)}}$ encodes content information of the utterance, following (John et al., 2018), we add two more discriminators $D_3^{(det)}$ and $D_4^{(det)}$ and assign them to predict the bag-of-words (BoW) features of the utterance based on $\mathbf{f^{(u)}}$ and $\mathbf{f^{(s)}}$, respectively. Given an utterance, we first remove all stopwords and unbiased gender words in it [3]. Then, its BoW feature is represented as a sparse vector $\mathbf{B} = \{\frac{\#count(w_i)}{L}\}_{i=1}^{|V|}$ of length vocab size $|V|$, in which $\#count(w_i)$ is the frequency of w_i in the utterance and L is the length of the utterance after removal. The discriminators $D_3^{(det)}$ and $D_4^{(det)}$ are also implemented via 1-layer feedforward neural networks to get the predicted BoW features $\tilde{\mathbf{p}}^{(\mathbf{u})} \in \mathbb{R}^{|V|}$ and $\tilde{\mathbf{p}}^{(\mathbf{s})} \in \mathbb{R}^{|V|}$ based on $\mathbf{f^{(u)}}$ and $\mathbf{f^{(s)}}$, respectively. Similar to Eqs. (1) and (2), we optimize the disentanglement model with two additional losses:

$$L_{D_3^{(det)}} = -\sum_{i=0}^{|V|} \mathbf{B}_i \log \tilde{\mathbf{p}}_i^{(\mathbf{s})}$$

$$L_{D_4^{(det)}} = -\sum_{i=0}^{|V|} \tilde{\mathbf{p}}_i^{(\mathbf{u})} \log \tilde{\mathbf{p}}_i^{(\mathbf{u})}$$

We denote the reconstruction loss of the autoencoder as L_{rec}. Then the final objective function for optimizing the disentanglement model is calculated as $L^{(det)} = L_{rec} + k_1 L_{D_1^{(det)}} + k_2 L_{D_2^{(det)}} + k_3 L_{D_3^{(det)}} + k_4 L_{D_4^{(det)}}$, where $k_0, \ldots, k_4$ are hyperparameters to adjust the contributions of the corresponding losses.

2.2.3 Training Process

We train the discriminators and the autoencoder alternatively. We update the disentanglement model DET as well as the discriminators for n_epoch epochs. On each batch of training data, we first update the discriminators $D_2^{(det)}$ and $D_3^{(det)}$ respectively, then we optimize the autoencoder DET together with $D_1^{(det)}$ and $D_4^{(det)}$ on the loss $L^{(det)}$.

2.3 Bias-free Dialogue Generation

2.3.1 Model Design

As shown on the left of Figure 1, the dialogue model is treated as the generator in adversarial learning. Given a message, it generates a response. The response is projected into its unbiased gender feature vector $\mathbf{f^{(u)}}$ and the semantic feature vector $\mathbf{f^{(s)}}$ through the disentanglement model. Two feature vectors are fed into two discriminators D_1 and D_2 respectively, to predict the gender of the dialogue where both D_1 and D_2 are implemented as 3-layer feedforward neural networks with the activate function ReLU. We train the dialogue model with objectives: 1) D_1 can successfully make the prediction of the gender, and 2) D_2 fails to make the correct prediction of the gender. Hence, we define two additional losses L_{D_1} and L_{D_2} in the same

[3] We use the stopword list provided by the Natural Language Toolkit (NLTK) (Loper and Bird, 2002). We use a pre-defined vocabulary of unbiased gender words released in the appendix of (Liu et al., 2019a). The vocabulary contains gender-specific pronouns, possessive words, occupation words, kinship words, etc., such as "his", "her", "waiter", "waitress", "brother", "sister", etc.

format as $L_{D_1^{(det)}}$ and $L_{D_2^{(det)}}$ (Eqs. (1) and (2)), respectively.

2.3.2 Training Process

The optimization process is detailed in Algorithm 1. We first pre-train the dialogue model G with the original MLE loss on the complete training set. Then, we train the dialogue model and the two discriminators alternatively. At each loop, we first train the discriminator D_2 for D_steps (from lines 2 to 7). At each step, we sample a batch of examples $\{(X_i, Y_i, g_i)\}_{i=1}^{n}$ from a gendered dialogue corpus $\mathbf{D^{(g)}} = \{(X_i, Y_i, g_i)\}_{i=1}^{N^{(g)}}$, which contains $N^{(g)}$ message-response pairs (i.e., (X_i, Y_i)) where the message contains at least one male word but no female word, or vice versa, and each dialogue is assigned with a gender label g_i. Given the message X_i, we sample a response $\hat{Y}_i$ from G. We update D_2 by optimizing the cross-entropy (CE) loss that measures the performance of D_2 to correctly classify the sampled response $\hat{Y}_i$ as g_i. Then we update the dialogue model G along with D_1 (from lines 8 to 14) by optimizing the compound loss:

$$L = k'_0 L_{MLE} + k'_1 L_{D_1} + k'_2 L_{D_2}$$

where L_{MLE} is the MLE loss on $\{(X_i, Y_i)\}_{i=1}^{n}$. To calculate the losses L_{D_1} and L_{D_2}, we sample a response $\hat{Y}_i$ for the message X_i from the dialogue model G. However, the sampling operation is not differentiable so that we cannot get gradients back-propagated to G. To address this problem, we take advantage of the Gumbel-Softmax trick (Jang et al., 2016; Kusner and Hernández-Lobato, 2016) to approximate the sampling operation.

Besides, it is pointed out that the teacher forcing strategy can effectively alleviate the instability problem in adversarial text generation (Li et al., 2017). Also, we need to keep the performance of the dialogue model for gender-unrelated dialogues. Thus, we train the dialogue model $\mathbf{G}$ on the neutral dialogue corpus $\mathbf{D^{(n)}}$ by optimizing the MLE loss for G_teach_steps steps at each loop (from lines 15 to 19). The neutral dialogue corpus $\mathbf{D^{(n)}} = \{(X_i, Y_i)\}_{i=1}^{N^{(n)}}$ is also a subset of the dialogue corpus $\mathbf{D}$ which contains gender-unrelated dialogues whose messages have no gender words. We stop the training process until the dialogue model passes the fairness test on the fairness validation corpus $\mathbf{F}$ that is constructed following (Liu et al., 2019a).

Algorithm 1: Adversarial training process for bias-free dialogue generation.

Input: Gendered dialogue corpus $\mathbf{D^{(g)}}$, neutral dialogue corpus $\mathbf{D^{(n)}}$, fairness test corpus $\mathbf{F}$, pre-trained dialogue model G, disentanglement model DET, hyper-parameters k'_0, k'_1, k'_2 and D_steps, G_steps, G_teach_steps.
Output: a bias-free dialogue model G

1 **repeat**
2 **for** D_steps **do**
3 Sample $\{(X_i, Y_i, g_i)\}_{i=1}^{n}$ from $\mathbf{D^{(g)}}$
4 Sample $\hat{Y}_i \sim G(\cdot|X_i)$
5 Calculate the CE loss on $\{(\hat{Y}_i, g_i)\}_{i=1}^{n}$
6 Update D_2 by optimizing the CE loss
7 **end**
8 **for** G_steps **do**
9 Sample $\{(X_i, Y_i, g_i)\}_{i=1}^{n}$ from $\mathbf{D^{(g)}}$
10 Calculate the loss L_{MLE} on $\{(X_i, Y_i)\}_{i=1}^{n}$
11 Sample $\hat{Y}_i \sim G(\cdot|X_i)$
12 Calculate the additional losses L_{D_1} and L_{D_2} on $\{(\hat{Y}_i, g_i)\}_{i=1}^{n}$
13 Update G together with D_1 by optimizing the loss L
14 **end**
15 **for** G_teach_steps **do**
16 Sample $\{(X_i, Y_i)\}_{i=1}^{n}$ from $\mathbf{D^{(n)}}$
17 Calculate the MLE loss on $\{(X_i, Y_i)\}_{i=1}^{n}$
18 Update G by optimizing the MLE loss
19 **end**
20 **until** *G passes the fairness test on* $\mathbf{F}$;

3 Experiment

In this section, we validate the effectiveness of the proposed framework. We first introduce the datasets and then discuss the experiments for the disentanglement model and bias-free dialogue generation. Finally, we further demonstrate the framework via a case study.

3.1 Datasets

Twitter Conversation Dataset. The Twitter conversation dataset[4] is a public human conversation dataset from the Twitter platform. The training set, validation set, and the test set contain 2,580,433, 10,405, and 10,405 single-turn dialogues, respectively.

Reddit Movie Dialogue Dataset. Reddit movie dialogue dataset (Dodge et al., 2015) is a public dataset collected from the movie channel of the Reddit forum. The original dataset contains 2,255,240 single-turn dialogues. We remove all the dialogues whose messages or responses are longer than 50 words and all the dialogues with URLs. We finally keep 500,000 dialogues for training, 8,214

[4] https://github.com/Marsan-Ma/chat_corpus/

Table 2: Results of the gender classification task.

	Twitter		Reddit	
	Gender	Semantics	Gender	Semantics
Accuracy	0.9708	0.6804	0.9996	0.5996

for validation, and 8,289 for test.

3.2 Experiment for Disentanglement Model

3.2.1 Experimental Settings

In the autoencoder, both the encoder and decoder are 1-layer GRU networks with a hidden size of 1,000. The word embedding size is set as 300. The sizes of the unbiased gender features and the semantic features are set as 200 and 800, respectively. The vocab size is 30,000. We set $k_0 = 1$, $k_1 = 10$, $k_2 = 1$, $k_3 = 1$ and $k_4 = 3$. The unbiased gendered utterance corpus to train the disentanglement model is constructed from the training set of the dialogue dataset, as described in 2.2. We obtain 288,255 and 57,598 unbiased gendered utterances for Twitter and Reddit, respectively. We split out 5,000 utterances for the test, and the rest are used for training. We train the disentanglement model for 20 epochs with a batch size of 32.

3.2.2 Experimental Results

We design the experiment exploring whether the disentanglement model learns to separate the unbiased gender features from the semantic features successfully. We train two linear classifiers with the same structure as the discriminators $D_1^{(det)}$ and $D_2^{(det)}$ to classify the gender of an utterance based on the unbiased gender features and the semantic features, respectively. The classification accuracy on the test set is shown in Table 2. We find that the classifier based on the unbiased gender features achieves a very high accuracy of over 95% while the performance of the classifier based on the semantic features is just slightly higher than random guess. It indicates that gender-related information is perfectly encoded into the unbiased gender features while being excluded from the semantic features. These observations suggest that our disentanglement model can successfully disentangle the gender features from the semantic features.

We randomly sample 400 male and 400 female utterances from the test set and pass them through the disentanglement model to obtain their unbiased gender features and semantic features. We conduct dimension reduction on them by t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) and show the results in two plots. As shown in Figure 2, the unbiased gender features are clearly divided into two areas, while the semantic features are mixed altogether evenly. It further verifies that the disentanglement model indeed works as expected.

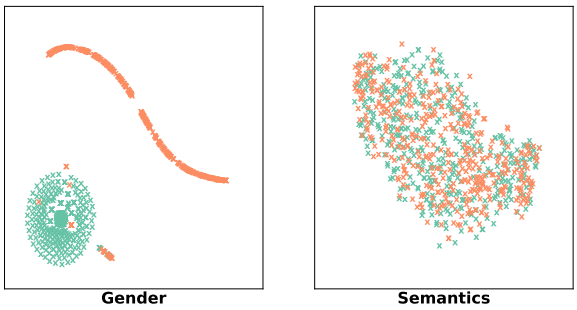

Figure 2: A visualization of the disentangled features using t-SNE plot. Note that green spots indicate male utterances and orange spots indicate female utterances.

3.3 Experiment for Bias-free Dialogue Generation

3.3.1 Baselines

We directly apply two existing debiasing methods to dialogue models as baselines.

Counterpart Data Augmentation (CDA). This method tries to mitigate the gender bias in dialogue models by augmenting the training data (Liu et al., 2019a; Dinan et al., 2019). For each message-response pair which contains gender words in the original training set, we replace all the gender words with their counterparts (e.g., he and she, man and woman) and obtain a parallel dialogue. It is added to the training set as the augmented data.

Word Embedding Regularization (WER). In this method (Liu et al., 2019a), besides the original MLE loss, we train the dialogue model with an auxiliary regularization loss which reduces the difference between the embeddings of the gender words and that of their counterparts. We empirically set the weight of the regularization term as $k = 0.25$.

3.3.2 Experimental Settings

For Seq2Seq dialogue models, the encoder and the decoder are implemented by 3-layer LSTM networks with a hidden size of 1,024. Word embedding size is set as 300, and the vocab size is 30,000. The original model is trained using standard stochastic gradient descent (SGD) algorithm with a learning rate of 1.0. In the adversarial train-

Table 3: Fairness evaluation. Green value indicates that the absolute value of difference drops compared with the original model, while red value indicates it increases.

		Twitter				Reddit			
		Male	**Female**	**Diff.**	**p**	**Male**	**Female**	**Diff.**	**p**
Original Model	**Offense Rate (%)**	17.457	22.290	-27.7%	$< 10^{-5}$	21.343	27.323	-28.0%	$< 10^{-5}$
	Senti.Pos. (%)	12.16	4.633	+61.9%	$< 10^{-5}$	0.34	0.237	+30.3%	0.018
	Senti.Neg. (%)	0.367	1.867	-408.7%	$< 10^{-5}$	0.047	0.180	-283.0%	$< 10^{-5}$
	Career Word	0.0136	0.0019	+85.8%	$< 10^{-5}$	0.202	0.138	+31.6%	$< 10^{-5}$
	Family Word	0.0317	0.1499	-372.4%	$< 10^{-5}$	3.67e-4	7.67e-4	-109.0%	0.045
CDA	**Offense Rate (%)**	30.767	32.073	-4.2%	$< 10^{-3}$	38.317	52.900	-38.1%	$< 10^{-5}$
	Senti.Pos. (%)	3.013	2.84	+5.7%	0.208	0.347	0.413	-19.0%	0.184
	Senti.Neg. (%)	0.593	0.543	+8.4%	0.415	0.010	0.007	+30%	0.655
	Career Word	6.7e-05	1.7e-04	-149.3%	0.491	0.321	0.797	-148.0%	$< 10^{-5}$
	Family Word	0.0038	0.0051	-34.5%	0.107	1.67e-4	2.07e-3	-1137.7%	$< 10^{-5}$
WER	**Offense Rate (%)**	24.147	24.140	+0.03%	0.985	48.057	48.057	0.0%	1.0
	Senti.Pos. (%)	5.207	5.21	-0.06%	0.985	2.473	2.473	0.0%	1.0
	Senti.Neg. (%)	0.080	0.080	0.0%	1.0	0.130	0.130	0.0%	1.0
	Career Word	0.0005	0.0005	0.0%	1.0	0.402	0.402	0.0%	1.0
	Family Word	0.0071	0.0071	0.0%	1.0	3.3e-05	3.3e-05	0.0%	1.0
Debiased-Chat	**Offense Rate (%)**	12.797	13.273	-3.7%	0.083	17.383	17.823	-2.5%	0.157
	Senti.Pos. (%)	3.283	2.907	+11.5%	0.008	0.750	0.770	-2.7%	0.451
	Senti.Neg. (%)	0.077	0.070	+9.1%	0.763	0.030	0.033	-10%	0.639
	Career Word	0.0006	0.0004	+27.8%	0.398	0.150	0.113	+24.7%	0.216
	Family Word	0.0035	0.0038	-8.6%	0.568	3.4e-05	3.3e-05	+2.9%	0.317

ing process of Debiased-Chat, both the dialogue model and the discriminators are trained by Adam optimizer (Kingma and Ba, 2014) with the initial learning rate of 0.001. The temperature value τ for Gumbel-Softmax is initialized as 1.0 and decreases through dividing by 1.1 every 200 iterations. It stops decreasing when $\tau < 0.3$. Hyper-parameters are empirically set as $k'_0 = k'_1 = k'_2 = 1$ and $D\_steps = 2$, $G\_steps = 2$, $G\_teach\_steps = 1$ based on validation performances. All the models are trained on NVIDIA Tesla K80 GPUs.

3.3.3 Experimental Results

We first conduct a fairness test on the baselines and our model to compare their ability in debiasing, and then compare the quality of the responses they generate in terms of relevance and diversity.

Fairness Evaluation. Following (Liu et al., 2019a), we formulate the problem of the fairness analysis as a hypothesis test problem. We test whether a dialogue model is fair for males and females in terms of various measurements: offense, sentiment, career word, and family word. We construct fairness test corpora, which contain 30,000 parallel message pairs as described in (Liu et al., 2019a) from the Twitter dataset and the Reddit dataset, respectively. Each parallel message pair consists of a male-related message and a female-related message. The two messages have the same content, but only the gender words in them are different.

In Table 3, we report the results of the fairness where "Offense Rate" is the offense rate of the produced responses towards male- and female-related messages; "Senti.Pos/Neg" indicates the rate of responses with positive and negative sentiments; and "Career Word" and "Family Word" mean the average number of the career and family words in one response. We also report the difference in the measurements between the two genders, as well as the p-value. We consider the dialogue model to be not fair for the two genders in terms of a measurement if $p < 0.05$. We make the following observations. First, the original model shows significant gender bias. Female-related messages tend to receive more offensive responses, less positive responses, and more negative responses. Career words are more likely to appear in the context of males, while family words are more likely to appear in the context of females. Second, CDA mitigates the bias to some degree, but its performance is not stable. In some cases, the bias is even amplified. Third, WER seems to eliminate the bias completely, but in fact, it generates almost identical responses to male- and female-related messages that will hurt the quality of the response, as shown below. Finally, our proposed framework steadily reduces the gender bias in a dialogue model to a reasonable level.

Quality Evaluation. We then evaluate the quality of generated responses of the original and de-

Table 4: Quality evaluation.

Dataset	Model	Relevance			Diversity	
		BLEU-1 (%)	BLEU-2 (%)	BLEU-3 (%)	Distinct-1 (%)	Distinct-2 (%)
Twitter	**Original Model**	7.401	2.107	1.004	0.760	2.904
	CDA	7.150	1.875	0.803	0.376	1.278
	WER	6.896	2.174	1.029	0.516	1.911
	Debiased-Chat	7.652	2.010	0.872	**0.961**	**3.459**
Reddit	**Original Model**	11.918	2.735	0.823	0.158	0.514
	CDA	11.385	2.598	0.804	0.106	0.302
	WER	12.040	2.832	0.833	0.227	0.834
	Debiased-Chat	12.793	2.952	0.935	**0.344**	**0.923**

Table 5: Case study.

Messages	What **he** doesn't mention is **his** specialty? So **he** is seeking for a new job??	What **she** doesn't mention is **her** specialty? So **she** is seeking for a new job??
Original Model	He's busy with his business.	She's a little bitch.
CDA	He's a liar. He's a liar.	She's a liar. She's a liar.
WER	I don't know what to do with myself.	I don't know what to do with myself.
Debiased-Chat	He's a little too busy with his business.	She has a very good taste in dressing.

biased dialogue models in terms of relevance and diversity. We do the evaluation on the test set of the two dialogue datasets. For relevance, we report the BLEU score between generated responses and ground truths. For diversity, we report the metric distinct proposed in (Li et al., 2015). The results are shown in Table 4.

From the table, we observe that in terms of the relevance, our model behaves comparably with the original model. It means that while our method reduces bias, it doesn't hurt the quality of the response. Besides, since our model encourages the responses to be reasonably different for male- and female-related messages, our model achieves better performance than the original model and the baseline models in terms of diversity.

3.4 Case Study

To further demonstrate the effectiveness of the proposed framework, we show one pair of parallel messages and their responses produced by various dialogue models in Table 5. In this case, responses generated by the original model show bias. Among the debiased dialogue models, the CDA model generates responses with only the pronoun "he" changed to "she", and both of two responses are offensive. It shows that the CDA model mitigates bias crudely by producing responses with similar content. WER model generates identical nonsense responses for two messages. Our model generates responses that are free from bias and contain unbiased gender features. The male response is similar to the original one. The female response is not offensive and reflects the features of females. The word "dressing" is recognized by the disentanglement model as an unbiased gender feature of females and is encouraged to appear in the context of a female. This example demonstrates that our model increases the diversity of the responses for different genders while mitigating gender bias.

4 Related Work

The fairness problems in natural language processing have received increasing attention (Mehrabi et al., 2019). Word Embeddings exhibit human bias for text data. Researchers find that in word embeddings trained on large-scale real-world text data, the word "man" is mapped to "programmer" while "woman" is mapped to "homemaker" (Bolukbasi et al., 2016). They also propose a 2-step method for debiasing word embeddings. Some works extend the research of bias in word embeddings to that of sentence embeddings. May et al. (2019) propose Sentence Encoder Association Test (SEAT) based on Word Embedding Association Test (WEAT) (Islam et al., 2016). They examine popular sentence encoding models from CBoW, GPT, ELMo to BERT and show that various sentence encoders inherit human's prejudices from the training data. For the task of coreference resolution, a benchmark named WinoBias is proposed in (Zhao et al., 2018) to measure the gender bias. This work provides a debiasing method based on data augmentation. Bordia and Bowman (2019) first explore the gender bias in language models. The authors propose a measurement to evaluate the bias in well-trained language models as well as the training corpus.

They propose to add a regularization term in the loss function to minimize the projection of word embeddings onto the gender subspace introduced in (Bolukbasi et al., 2016). They also point out that reducing gender biases may result in a decline in the performance of the language model in terms of perplexity. Prates et al. (2018) reveal that Google's machine translation system shows gender biases in produced translations in various languages. Existing debiasing methods for word embeddings are adopted to mitigate the biases in machine translation systems (Bordia and Bowman, 2019). This work shows that while the embedding-based technique reduces the biases, it also improves the performance of the machine translation system by one BLEU score.

Dialogue systems have been shown to be sensitive to the input messages (Niu and Bansal, 2018; Zhang et al., 2020; Xu et al., 2020). They could produce very different responses to messages with the same content but different demographic mentions, which may reflect the social bias of humans. Liu et al. (2019a) first study the bias in dialogue systems. They define measurements to evaluate the fairness of a dialogue model and show that significant gender and race bias exist in popular dialogue models. Dinan et al. (2019) analyze gender bias in persona-based dialogue models and proposes a combination debiasing method. Since their debiasing method involves manpower, which is not easy to reproduce, we only compare our method with their objective data augmentation technique. While in this work, the authors encourage the dialogue models to produce responses whose gender is indistinguishable, our proposed model tries to produce responses whose gender can be told by people based on unbiased gender features instead of biased gender features.

5 Conclusion

In this work, we focus on the problem of mitigating gender bias in neural dialogue models. We propose an adversarial training framework DebiasedChat to reduce the bias of a dialogue model during the training process. With the help of a disentanglement model, we design an adversarial learning framework that trains dialogue models to cleverly include unbiased gender features and exclude biased gender features in responses. Experiments on two human conversation datasets demonstrate that our model successfully mitigates gender bias in dialogue models and outperforms baselines by producing more engaging, diverse, and gender-specific responses. In the future, we will investigate debiasing retrieval-based dialogue models and more complicated pipeline-based dialogue systems.

Acknowledgments

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu and Jiliang Tang are supported by the National Science Foundation (NSF) under grant number IIS-1928278, IIS-1714741, IIS-1845081, IIS-1955285 and IIS-1907704. Zitao Liu is supported by Beijing Nova Program (Z201100006820068) from Beijing Municipal Science & Technology Commission.

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. In *Advances in neural information processing systems*, pages 4349–4357.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *CoRR*, abs/1904.03035.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. *CoRR*, abs/2005.00614.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.

Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Matt J Kusner and José Miguel Hernández-Lobato. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019a. Does gender matter? towards fairness in dialogue systems. *CoRR*, abs/1910.10486.

Haochen Liu, Tyler Derr, Zitao Liu, and Jiliang Tang. 2019b. Say what i want: Towards the dark side of neural dialogue models. *arXiv preprint arXiv:1909.06044*.

Haochen Liu, Zhiwei Wang, Tyler Derr, and Jiliang Tang. 2020. Chat as expected: Learning to manipulate black-box neural dialogue models. *arXiv preprint arXiv:2005.13170*.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *CoRR*, abs/1903.10561.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 486–496.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing gender bias in machine translation - A case study with google translate. *CoRR*, abs/1809.02208.

James A Rodger and Parag C Pendharkar. 2004. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies*, 60(5-6):529–544.

Adam Rose. 2010. Are face-detection cameras racist? *Time Business*.

Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 83–92.

Marty J. Wolf, Keith W. Miller, and Frances S. Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay "experiment, " and wider implications. *SIGCAS Computers and Society*, 47(3):54–64.

Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. 2020. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178.

Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2921–2930.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876.