Neurocomputing 493 (2022) 422–434

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Examining and mitigating gender bias in text emotion detection task

Odbal [a], Guanhong Zhang [b,*], Sophia Ananiadou [c]

[a] Anhui Vocational and Technical College, China
[b] Hefei University, China
[c] The University of Manchester, England, United Kingdom

ARTICLE INFO

Article history:
Received 26 July 2021
Revised 15 March 2022
Accepted 9 April 2022
Available online 14 April 2022
Communicated by Zidong Wang

Keywords:
Gender bias
Text emotion detection
Bias examine
Debiasing
Adversarial training

ABSTRACT

Gender bias is an important problem that affects models of natural language, and the propagation of such biases could be harmful. Much research focuses on gender biases in word embeddings, and there are also some works on gender biases in subsequent tasks. However, very limited prior work has been done on gender issues in emotion detection tasks. In this paper, we investigate the effect of gender in text emotion detection. Existing methods for gender biases require gender balanced and gender-swapping data, and might influence the performance of the target task due to removing more information related to sensitive attributes. We present different solutions to measuring and mitigating gender bias in emotion detection. To measure gender bias, we first prepare datasets annotated with emotional classes and gender information. Then, we compare the performance of emotion recognition models from gender balanced samples, and also analyze gender prediction results from emotion related data. Our experiment results show that there exists gender bias in emotion detection: the models trained on the female data often achieve better results than the male models, and the female models and the male models report the opposite trends on the recognition of some emotions. We also attempt to mitigate gender bias by developing various approaches including products of experts, introducing weights and variants of focal loss, as well as adversarial training. Compared to other debiasing methods, adversarial trainings represent tpr reduction approximately 0.02–0.03 while simultaneously less harming performance by below 1.0 points on our prepared datasets. Further, we show that efficient parameters can lead to further improvements.

1. Introduction

Gender biases have been paid attention to by many researchers within the natural language processing (NLP) community, because they highly affect models of natural language, and the propagation of such biases could result in some dangerous stereotypes when they come to real-world downstream applications. This is dissatisfying in practice, for example, a recruitment support system based on a database that is trained only on men systematically ranks women lower when they are compared with similarly qualified men [5]. Microsoft's AI chat bot Tay learnt abusive language from Twitter within the first 24 h of its release, which forced Microsoft to shutdown the bot [47]. We believe it matters if a NLP system exists gender bias. However, solving gender bias is a hard problem because such bias exists in multiple parts of a NLP system, including training data, pretrained models (e.g. word embeddings), and algorithms. In addition, we do not have a universal definition about gender bias and evaluation criteria used to measure it even though there has been a common understanding that gender bias is the preference or prejudice toward one gender over the other [44].

Much of NLP research focuses on gender biases in word embeddings, attempting to identify and debiase them (e.g, [6,59,4]). There are also some works on gender bias in subsequent tasks. For example, [38] evaluated three publicly-available, off-the-shelf coreference resolution systems and found systematic gender bias in each: for many occupations, systems strongly prefer to resolve pronouns of one gender over another. [43] presented evidence for gender bias in machine translation (MT), showing that both commercial systems and academic MT models were significantly prone to translate based on gender stereotypes rather than more meaningful context. Other examples of research on gender bias are coreference resolution in [58], language modeling in [29], abusive language detection in [33], sentiment analysis in [24] and machine translation presented in [41] etc. In Table 1, we list common examples of gender bias in the above tasks.

In the field of sentiment analysis and emotion detection, however, very limited prior work has been done on gender issues (e.g., [23,24,48]). This may be due to the fact that emotions are

* Corresponding author.
E-mail address: ustcghzhang@ustc.edu.cn (G. Zhang).

https://doi.org/10.1016/j.neucom.2022.04.057
0925-2312/

Table 1
We list some examples on gender bias in NLP tasks.

Task	Example	Reference
Machine Translation	English: The doctor asked the nurse to help her in the procedure. Spanish: El doctor le pidio a la enfermera que le ayudara con el procedimiento	[43]
Abusive Language	"You are a good woman" was considered "sexist"	[33]
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor"	[44]
Word Embeding	Analogies such as "man: woman, computer programmer: homemaker"are automatically generated by models trained on biased word embeddings.	[6]
Sentiment Analysis	More than 75% of the systems tend to mark sentences involving one gender/race with higher intensity scores than the sentences involving the other gender/race.	[24]
Coreference Resolution	A man and his son get into a terrible car crash. The father dies, and the boy is badly injured. In the hospital, the surgeon looks at the patient and exclaims, "I can't operate on this boy, he's my son!" Many first-time listeners have difficulty assigning both the role of "mother" and "surgeon" to the same entity.	[38]

treated as physical types that have little to do with the words we use, and gender seems irrelevant whether a text is produced by a woman or a man. However, cross-gender differences in the use of language to express emotions have been examined in sociolinguistics. Early literature [3] pointed out that typical female language features are different from typical male language features. [31] reported females used more references to positive emotion, and made more adjectives related to anger. [32] mentioned that references to other-directed negative emotions (e.g., anger) were predominant for boys, and inner-directed negative emotions (such as sadness, fear, guilt and shame) were characteristic of girls. [14] claimed that one of the most obvious gender differences is the men's lower salience of the words sadness and happiness than women's. Obviously, gender differences in the linguistic expression of emotion are discovered, and they will influence word choice or syntax. We thus ask whether gender is an important factor that could affect text emotion detection models, whether these models yield gender biased predictions, and whether debiasing is needed for such tasks.

Various approaches have been proposed in the literature to analyze and solve the problem of gender bias. Building and using gender balanced and gender-swapping data (e.g., "He went to the park" vs "She went to the park") are described to show gains (e.g., [29]) in measuring and reducing gender biases. Debiasing word embeddings are also useful for NLP models (e.g., [6,59,4]). There have also been a series of papers that attempt to reduce gender bias by adjusting training algorithms (e.g., [57,56,28,9]). However, commonly used methods such as gender-swapping data is not compatible with the text emotion detection task because gender information will rarely be seen in emotional sentences (e.g., I never knew a detention was so hard to get.). The debiased embeddings might not carry adequate information for downstream NLP tasks if too much information is removed from word embeddings. Adjusting training algorithms necessitate designing reasonable models and training schemes to balance blinding information related to the sensitive attribute (e.g., gender) and retaining information necessary for the target task.

In this work, we present different solutions to measuring and mitigating gender bias in text emotion detection. Like the above work, we first examine gender bias on the performance of neural network models for classification tasks. Unlike them, we observe gender bias from two perspectives: performance differences of emotion detection models across gender specific datasets and gender prediction results from emotion specific datasets. And, our experiments are performed using random splits rather than gender-swapping datasets. Additionally, in order to avoid gender biased predictions, we propose various training schemes instead of a single strategy. These methods work by adjusting the training loss to reduce biases.

Specifically, we first randomly split emotion datasets across genders: female, male and mixed. We define female datasets as a set of sentences written by women, and male datasets as a sequence of sentences written by men. Mixed datasets are sentences including both group datasets. We further develop emotion detection models trained on gender specific datasets, and compare performance differences when classifying emotions across different models. We also explore gender prediction models trained on emotion specific datasets to check gender bias from a view of predicting gender. Both models are developed based on convolutional neural nets (CNNs) and Transformer. In addition, we propose various methods for reducing the effects of gender bias in emotion detection models. We first present three general approaches, namely Products of Experts, Introducing weights, and the variant of Focal Loss to vary the training loss. We also propose a different solution to adjust the training algorithm based on adversarial nets, in which emotion detection and gender prediction models are jointly trained in an adversarial way. This method assumes that the generalization capability of the model could be enhanced to defend against the influence of sensitive attributes, such as gender.

Efforts on analyzing gender bias, however, require training datasets annotated with both emotion labels and gender information. Both are not simultaneously found in most existing datasets: sufficient emotion datasets are not annotated with gender information or datasets with gender information lack affective labeling. To study gender bias in text emotion detection, we make use of an emotion dataset, ISEAR, which contains both emotion and gender information. And, we augment another emotion dataset, CrowdFlower (a twitter emotion data), with gender tags (female and male) based on author names, profile photos, and descriptions.

To summarize, we make the following contributions:

1. We annotate a public available emotion dataset with gender information which could be used for both emotion detection and gender identification tasks.
2. Through extensive experiments, we investigate the effect of gender on classification performance, and show gender bias exists in our tasks.
3. We propose various training schemes to reduce gender bias in emotion detection tasks, and our experiments show they can be optimal if one takes into account the tpr-gap/accuracy trade-off.

The plan of this paper is as follows: In Section 2, we review previous work. In Section 3, we prepare our datasets. In Section 4, we give evaluating strategies for gender bias and discuss the results. In Section 5, we describe several methods for mitigating gender bias. Section 6 gives conclusion and future work.

2. Related Work

We are inspired by related work on examining and mitigating gender bias, fairness in machine learning models as well as generative adversarial networks.

Examining and Mitigating Gender Bias. As mentioned in the introduction, there are several approaches to evaluating and alleviating gender biases for various NLP downstream tasks. Designing gender-swapping test sets to evaluate the performance difference of the model on male and female data could be easily implemented for many NLP tasks [57,38,58,12,29]; [24]; [37]; [44]. [58] created an augmented dataset using gender-swapping technique to the original dataset and trained the model on the union of the original and data-swapped sets, while [29] measured and mitigated bias with counterfactual data augmentation (CDA). Their approach used causal interventions that broke associations between gendered and gender-neural words. Another approach is de-biasing word embeddings that attempt to mitigate bias in embeddings (e.g., [6,59,4]). A recent survey of bias in NLP [5] found that one third of all research works focused on bias in word embeddings [17]. More recently, [33] used transfer learning from a gender unbiased abusive tweets dataset and fine-tuning on a gender biased sexist tweets dataset. Another different approach to the above adjusting training scheme is to modify training objectives using adversarial training (e.g. [15]). We borrow ideas from this line of research in that our learning process jointly trains the emotion detection model and gender identification model.

Similar to the above work on measuring gender bias, we use the idea of comparing the performance difference of the model trained on the male and female data. The main differences are (i) the random splits of datasets, (ii) the observation of the performance difference of emotion detection model on gender specific data and gender identification rate from emotion specific data.

Fairness in Machine Learning Models. Fairness is becoming one of the most popular topics in machine learning in recent years. Two lines of research are particularly common. The first is evaluating fairness of the machine learning systems using a variety of measures. For example, [20] proposed Equalized odds, also called Separation, Positive Rate Parity for discrimination against a specified sensitive attribute in supervised learning. [25] proposed calibration techniques to achieve fairness. Predictive Rate Parity, also called Sufficiency, appeared in [54], and Counterfactual fairness was proposed in [39], which provides a possible way to interpret the causes of bias. There are also other definitions of fairness that have been proposed in the literature (see [16,51]).

The second line of research is improving fairness. [55] formulated fairness as an optimization problem of finding an intermediate representation of the data that best encodes the data while simultaneously obfuscating aspects of it, removing any information about membership with respect to the protected group. [21] found that the status quo of empirical risk minimization (ERM) amplifies representation disparity over time, and distributionally robust optimization (DRO) was proposed to mitigate bias, providing an upper bound on the risk incurred by minority groups and performing well in practice. [1] achieved fairness by means of optimizing the tradeoff between accuracy and any (single) definition of fairness given training-time access to protected attributes. The article[10] utilized product-of-expert (PoE), which was originally introduced in [22], for avoiding datasets biases. [35] proposed end-to-end debiasing techniques to adjust the cross entropy loss for reducing the biases learned from the training dataset by down-weighting the biased examples. [49] used the "soft" target labels to compute the loss function. We extend the above ideas from known and hard examples to female datasets as well as male datasets.

Generative Adversarial Networks. [18] proposed the adversarial nets framework. They introduced a framework for estimating generative models via an adversarial process, in which two models were simultaneously trained: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G was to maximize the probability of D making a mistake.

In recent years, the general family of adversarial nets methods have gained significantly more attention in NLP. Recent attempts at using adversarial nets to deal with the problem of NLP include adversarial examples and attacks [2,15,36,26,52,60], where researchers focus on learning or creating adversarial examples, attacks or rules to improve the robustness of NLP systems. Another line of research is adversarial training [53,30,40,13], that focuses on adding noise, randomness, penalty or adversarial loss during optimization. Similar to their work, we use the idea of adversarial training to reduce the influence of sensitive attributes (e.g., gender) to emotion detection models.

Although adversarial training methods can combine any representation learning based neural networks, and perform optimization for the NLP models, a key challenge for applying adversarial training techniques to NLP tasks is the model design issue since it seems that the discriminator might overpower the generator. In this paper, we design an adversarial classifier and an emotional classifier using adversarial training as a means to hide the protected attributes from the representation process.

3. Data Preparation

Even though several datasets have been annotated and published for classification of text emotion [7], most of them lack gender information, which increases the difficulty of analyzing the problem of gender bias in emotion detection tasks.

In order to analyze gender bias in emotion detection, we first prepare adequate sized datasets labeled by emotion categories along with gender information. We use two publicly available emotion datasets: ISEAR [42] and CrowdFlower [7]. ISEAR contains 7,659 entries in terms of gender (4,201 sentences from females and 3,458 from males) containing seven major emotions: joy, fear, anger, sadness, disgust, shame, and guilt. CrowdFlower, annotated via crowdsourcing with one label per tweet, consists of 40,000 tweets with emotion categories: empty, sadness, enthusiasm, neural, worry, sadness, love, fun, hate, happiness, relief, boredom, surprise and anger. However, this corpus lacks gender information. We enrich CrowdFlower semi-automatically by using the "author" tag, which associates with author profiles. In order to automatically associate authors' names with the corresponding gender, we compiled a dictionary of English names based on Facebook profiles by [46] and the Social Security Administration's (SSA) baby names data set. The names dictionary contains about 8,444 names, and it is currently composed of 3,304 male names as well as 5,140 female names.

The annotation task is conducted as follows: firstly, we split the user name which corresponds to the "author" tag into separate words using regular expressions and look for the separate names from the first names dictionary which we have obtained. User names are found by recursively looking for names inside the string that corresponds to the entries in our dictionary of English names, while gender information can be directly annotated by consulting the names dictionary. For instance, the name Tommy can be found in the username TrustTommy, then the gender of user TrustTommy is labeled by male. If no name is found then: (1) if the user name contains repeated vowels (e.g. AprilLouiseeee), then we

remove repeated vowels in user name (e.g.Louise), and find dictionary names in modified username (e.g. Louise is female first name); (2) if the username contains leet speak characters (e.g. Ver0nica), then we replace user name "leet speak" characters with their equivalents (e.g. Veronica); (3) if the user name contains "Miss, Mrs, Mr" characters, we can directly annotate the user's gender (e.g. the gender of user Mr. Vose as male).

Secondly, we manually label gender information taking into consideration the profile picture of the user. If the picture does not correspond with the gender of the user, we check the description of the user profile (e.g. Painter, designer, craftsperson, female). Users without photos, celebrity-based pictures or descriptions are labeled as unknown. Finally, if users have blogging sites associated with their profiles, we follow those URLs and validate the data found with their profiles. By generating the gender information of Crowdflower in this way, we obtain 15,948 sentences for females and 14,790 sentences for males, excluding unknown gender.

Since CrowdFlower is a noisy dataset, we remove some sentences consisting of less than two words without any emotion words or emoticons, and provide the data containing the same numbers of female and male sentences by discarding female examples for experiments. Table 2 shows the number of sentences and labels in the ISEAR and CrowdFlower datasets in terms of gender and emotion categories.

4. Examining Gender Bias

4.1. Definition

The notation convention we employ is as follows. We use the symbol D to denote an annotated dataset, and D contains instances (X, Y, G), where X denotes a set of sentences x_i annotated with emotion labels $Y = \{y_i\}$ (e.g., joy, sadness) and sensitive attribute $G = \{g_i\}$ (e.g., male, female). In this paper, we only consider two genders: male and female.

We define gender bias in this context as the disproportionate influence of the emotion expressions of one gender over another in the output of the model. We assume that if there exists a function f such that $y_i = f(x_i, g_0)$ and $y_j = f(x_i, g_1)$ (e.g., g_0 indicates male and g_1 indicates female), we say that f exists bias.

Table 2
Data Statistic.

	Female	Male
ISEAR data		
anger	494	495
disgust	494	495
fear	494	494
joy	494	494
guilt	494	492
sadness	494	495
shame	494	495
	3,458	3,460
CrowdFlower data		
anger	48	36
boredom	79	55
empty	330	314
fun	662	666
love	1,627	1,321
sadness	2,185	1,818
surprise	822	845
happiness	2,077	1,924
hate	544	479
enthusiasm	280	290
relief	619	571
worry	2,185	3,139
	11,458	11,458

We also assume if we train a classifier C to predict g_0 with an accuracy beyond chance level, while to predict g_1 with an accuracy above this level, we consider C is biased.

4.2. Experimental Setup

To evaluate gender bias in emotion detection, we design two sets of experiments: (1) Experiment 1 examines the effect of gender on the performance of emotion detection models. (2) Experiment 2 tests gender predicted scores on the specific emotion related data.

We treat the task of experiment 1 as a multi-class classification task and design two baseline models with Convolutional Neural Network (CNN) and Transformer to classify emotion categories. We select an equal number of instances from both genders (male and female), considering 6,916 instances from the ISEAR dataset, where each gender consists of 3,458 sentences. Since the CrowdFlower data is emotion imbalanced, we choose instances from fun, happiness, hate, love, relief, sadness, surprise and worry by discarding anger, boredom, empty and enthusiasm examples so that the number of different emotion sentences are relatively balanced. In this case, CrowdFlower used in experiment 1 contains 11,656 instances where each gender consists of 5,828 sentences. In each training scenario, we train the models on: (1) a mixed-gender data containing an equal number of male and female sentences, (2) female data, as well as (3) male data. We call the female model when the model is trained on the female datasets, and the male model is a model trained on the male data. The used hyper-parameters of CNN are: convolution layers with 3 filters with the size of [3,4,5] followed by max pooling layer, feature map size = 100, dropout = 0.25. The pre-trained word embedding, the Global Vectors (GloVe) [34] are used in the embedding layer. We employ a standard bert based Transformer architecture [50], with 12 layers of encoder and 12 layers of decoder with model dimension of 1024 on 16 heads.

In experiment 2, we treat the task as a binary classification problem and use CNN and Transformer to distinguish female from male. Literature in sociolinguistics shows that men and women differ in the use of emotion terms. From this point forward, we formulate our task as a gender identification problem using emotion features, and we attempt to automatically identify the gender of a person in their context. Therefore, our datasets are grouped into various subsets according to different emotion categories and each subset contains equal numbers of sentences from two sexes (female and male). In our experiment, we divide the ISEAR dataset into 7 subsets, each consisting of 988 sentences with equal numbers of female and male. The CrowdFlower dataset is divided into 8 subsets, where each includes an equal number of sentences from the two groups. The used hyperparameters of CNN and Transformer in experiment 2 are similar to experiment 1.

In contrast with previous approaches [12,33,24], we do not develop a separate unbiased test set for each gender, but rather randomly generate training-validation-test splits at a 80:10:10 ratio on each data set, inspired by [45]. The test data contains an equal number of instances from both genders. We repeatedly conduct experiments on random splits (5–10 times), and average the final results. Both experiments use Accuracy and F1-score as evaluation metrics.

4.3. Results and Discussion

4.3.1. Emotion Detection Task

Tables 3 and 4 show the results of experiment 1 across all the emotion labels and individual tags, respectively. We note here that differences refer to performance differences of models.

We observe that models trained on the female data from both the ISEAR and CrowdFlower corpus achieve the highest accuracy and F1-score (in bold, Table 3). For example, Transformer achieves

Table 3
Emotion classification performance. Diff is the absolute performance difference between the model trained on the female data and the male data.

Gender	ISEAR		CrowdFlower		
	Accuracy	F1	Accuracy	F1	
CNN	Mixed	58.64	57.97	32.08	31.60
	Female	**61.09**	**60.62**	**35.84**	**35.31**
	Male	53.10	52.69	32.43	32.79
	Diff	7.99	7.93	3.41	2.52
Transformer	Mixed	66.03	65.91	43.89	42.15
	Female	**73.92**	**73.50**	**47.76**	**41.67**
	Male	65.95	62.96	43.18	39.44
	Diff	7.97	10.54	4.58	2.23

Table 4
Classification performance for each emotion category based on F1 score. Diff is the absolute difference of F1 between the model trained on the female data and the male data.

Emotion	ISEAR				
	F1(Mixed)	F1(Female)	F1(Male)	Diff	
CNN	anger	52.22	50.78	30.92	19.86
	disgust	63.19	72.35	57.19	15.16
	fear	61.76	72.51	74.39	1.88
	guilt	36.80	48.86	37.20	11.66
	joy	73.51	76.99	64.41	12.58
	sadness	67.66	59.01	59.68	0.67
	shame	50.64	42.11	45.75	3.64
Transformer	anger	50.25	70.71	55.56	15.15
	disgust	50.16	81.85	46.83	35.02
	fear	74.98	79.35	66.27	13.08
	guilt	63.73	49.05	81.67	32.62
	joy	83.87	86.66	85.68	0.98
	sadness	66.78	74.05	78.89	4.84
	shame	59.51	75.56	45.96	29.6
Emotion	CrowdFlower				
	F1(Mixed)	F1(Female)	F1(Male)	Diff	
CNN	fun	25.49	35.48	24.97	10.51
	happiness	22.03	26.50	22.42	4.08
	hate	36.66	42.96	37.01	5.95
	love	43.11	46.76	46.43	0.33
	relief	27.09	26.69	30.37	3.68
	sadness	31.98	33.96	35.52	1.56
	surprise	33.19	31.61	31.15	0.46
	worry	34.27	25.65	33.79	8.14
Transformer	fun	27.67	33.22	22.22	11
	happiness	28.05	52.56	55.56	3
	hate	50.02	39.00	50.00	11
	love	39.12	48.89	28.18	20.71
	relief	22.53	37.15	36.36	0.79
	sadness	56.83	47.71	34.78	12.93
	surprise	28.31	20.52	30.77	10.25
	worry	48.76	45.19	51.85	6.66

accuracy and F1-score: (73.92%, 73.5%) and (47.76%, 41.67%), respectively. Interestingly, the male models achieve the lower accuracy and F1-score on both datasets (e.g., (53.1%, 52.69%) with CNN and (65.95%, 62.96%) with Transformer on ISEAR data).

When we view the model generated on the mixed data as the baseline model, compared to the baseline, the female CNN model could increase the performance by around 3% on both the ISEAR and CrowdFlower datasets, while the male model decreases the performance or remains the same with the baseline. Transformer generates similar trend with CNN on both datasets. In general, female models are more accurate than male ones with above 7% on the ISEAR data and around 3% on the CrowdFlower data, showing biased predictions.

As shown in Table 4, we also observe clear differences across gender-based models as to specific emotions. Compared to the results of the models trained on the mixed data, the female models trained on the ISEAR data predict better disgust, fear and joy. For the male models, while we see lower scores are presented in disgust and shame, they predict better guilt. Our results show that the opposite trend is seen in gender-based models with regards to disgust. It is also noteworthy that while the differences of two gender models predicting disgust are larger, there is a smaller difference between sadness.

The CrowdFlower dataset also makes differences. From the recognition of the female models in fun, happiness, hate and love, we see much higher scores as compared to the baseline model but worse for surprise and worry. Compared to the results of the models on the mixed data, the male models show lower scores in emotions depicting fun, but achieve better scores in relief. It seems that fun shows the opposite trend across two genders. While we see a larger difference in fun and hate, the gap among relief and sadness show smaller.

In general, we observe that gender impacts emotion detection tasks, and emotion detection models exist gender bias. The performances of models trained on female datasets are better than those trained on male datasets, and female models are more accurate than male models in categorizing sentences that express positive emotions such as joy, happiness and love. In contrast, male models

achieve higher performance when identifying negative emotions such as guilt or relief. Our results partially support previous findings that women use more references to positive emotions such as love, happiness, joy (e.g., [31]). At the same time, the results show that the advantage of male model is significant for guilt and relief.

4.3.2. Gender Identification Task

We analyze the results of gender identification from a certain emotion data (Experiment 2).

The results in Table 5 show that across the models trained on the anger, disgust, shame and guilt subsets from the ISEAR dataset, the F1-score of men is considerably higher than women. The average F1-score of all models for women are 54.42% with CNN and 59.57% with Transformer vs. 60.83% with CNN and 59.70% with Transformer for men. We see similar patterns in the happiness, relief and sadness subsets in the CrowdFlower dataset, and the average F1 scores of men are higher than that of women (e.g., 54.01% vs. 55.74% with Transformer).

In terms of difference, we observe that the models trained on the sadness subset from the ISEAR corpus produce the higher differences when predicting female and male, while the joy and fear subsets from the ISEAR dataset account for the lower scores. For the CrowdFlower data, the higher differences are obtained from the relief, fun and hate subsets; The surprise, worry as well as love subsets from CrowdFlower corpus have the lower scores.

We assume that joy and love basically belong to the same emotion category, sadness and relief express similar emotions, and disgust as well as hate express a strong feeling of dislike. Based on the above considerations, we can conclude that the identification gap between female and male is larger in sadness related data, whereas joy related data shows the opposite trend. That is to say, the models trained on sadness-related data and joy-related data provide slightly higher prediction scores for one group over another. One possible explanation for the results could be that one gender group uses sadness related or joy related terms more often than another group, which (in) directly helps to generate biased features. However, the models trained on happiness and joy subsets, which belong to the same emotion category, get inconsistent results. For the other emotion categories, the pattern of results is not as clear-cut.

4.3.3. Discussion

Gender affects the performance of emotion detection models, and the female model often outperforms the male model when recognizing certain emotions. Considering specific emotions, happy emotions such as joy, happiness and love have higher recognition rates when attributed to females, whereas distressing words (i.e., guilt and relief) are mostly attributed to males. In these cases, the models trained on gender-unbalanced data tend to give biased results: if the training data contains more sentences written by women than men, the model might attain better performance and predict better happy-related emotions. Furthermore, the gender prediction gap between female and male is larger in the model trained on sadness related corpus, whereas the model trained on joy related data has the opposite trend. The results clearly show gender bias: the preference or prejudice toward one gender prediction over the other.

In this section, we do not compare our classification models with other state-of-the-art neural models (e.g, transformer) for emotion detection or gender prediction tasks because our focus is just examining gender bias. We thus explore CNN based and bert based transformer models within a simpler and easier-to-analyze setting. And also, due to the choice of experiment setup and various emotional categories, we do not compare our results with previous works (e.g., [24]). However, we believe that in general, the

Table 5
Gender identification results. Diff is the difference between f1 score of female and male, Diff =|F1 (**M**ale)- F1(**F**emale)|.

	ISEAR Data					
	sub_data	Acc	F1	F1(F)	F1(M)	Diff
CNN	joy	58.79	58.43	58.26	58.60	0.34
	fear	55.56	55.36	55.78	54.93	0.85
	anger	54.21	52.82	44.98	60.64	15.66
	disgust	58.25	57.55	52.31	62.80	10.49
	sadness	59.94	59.39	54.77	64.01	9.25
	shame	61.96	61.80	59.47	64.13	4.66
	guilt	58.25	58.05	55.39	60.72	5.33
Transformer	joy	64.76	64.87	66.32	62.78	3.54
	fear	54.29	52.91	56.07	51.19	4.88
	anger	59.05	59.55	58.01	59.31	1.3
	disgust	67.62	67.52	67.01	72.00	4.99
	sadness	51.43	51.50	57.53	44.13	13.4
	shame	55.24	53.19	47.12	58.09	10.97
	guilt	67.62	67.51	64.95	70.42	5.47
	CrowdFlower Data					
	sub_data	Acc	F1	F1(F)	F1(M)	Diff
CNN	fun	47.58	47.24	42.48	51.85	9.37
	happiness	50.97	50.85	46.99	54.40	7.41
	love	55.28	54.94	56.68	53.26	3.43
	worry	45.45	45.44	44.38	46.49	2.11
	hate	51.65	50.48	56.86	45.00	11.80
	sadness	47.81	47.84	46.88	48.71	1.83
	surprise	48.70	48.72	49.68	47.68	2.00
	relief	55.14	49.86	39.41	60.12	20.71
Transformer	fun	53.34	53.03	58.01	46.54	11.47
	happiness	53.85	53.91	52.45	54.84	2.39
	love	52.47	52.69	52.68	51.18	1.5
	worry	55.04	54.61	54.72	53.85	0.87
	hate	66.67	66.67	63.51	68.72	5.21
	sadness	59.42	59.22	56.00	59.72	3.72
	surprise	58.97	59.13	58.70	55.51	3.19
	relief	48.06	46.00	35.98	55.57	19.59

results generated in this section can include some key information to analyze the fairness of emotion detection models.

5. De-biasing Methods

5.1. General Approaches

Inspired by the work of [10,35], we present and compare three general approaches to mitigating gender bias: product of experts, introducing weights, and variant of focal loss.

5.1.1. Product of Experts

In the product of experts (PoE) method [22], a probability distribution is obtained by combining multiple probabilistic models of the same data by multiplying the probabilities together and then renormalizing, and n individual expert models can be combined by:

$$p(d|\theta_1,\ldots,\theta_n)=\frac{\prod_m p_m(d|\theta_m)}{\sum_i\prod_m p_m(c_i|\theta_m)}$$

where d is a data vector in a discrete space, θ_m is all the parameters of individual model m, $p_m(d|\theta_m)$ is the probability of d under model m, and i is an index over all possible vectors in the data space.

[10,35] used this method to combine predictions of the base model and the bias-only model. The idea is to compute the element-wise product between their predictions. Here, we follow and expand their methods to combine the predictions of the mixed, female and male models. The final distribution is then computed as:

$$p_i = log(p_i^t) + log(p_i^m) + log\left(p_i^f\right) \tag{1}$$

where we use the symbol p(·) to denote model-based probability distributions, and $p^t(\cdot)$, $p^m(\cdot)$ and $p^f(\cdot)$ denote the probability distributions generated by the mixed, male and female models, respectively. The training process is to combine the probability distributions of these models.

5.1.2. Introducing Weights

A common method for modifying the loss function is to introduce a weighting factor α. For example, $\alpha \in [0,1]$ for class 1 and $1-\alpha$ for class -1. In practice α may be set by inverse class frequency or treated as a hyperparameter to set by cross validation, and the α-introduced cross entropy (CE) loss can be written as:

$$CE(p_i) = -\alpha_i \log(p_i)$$

Here, p_i denotes the estimated probability of the model for the given class.

We propose a variant of this loss that is a weighted summation of the mixed, loss functions of the female and male models, expressed as:

$$VCE(p_i) = -\alpha_t \log(p_i^t) - \alpha_f \log\left(p_i^f\right) - \alpha_m \log\left(p_i^m\right) \tag{2}$$

Here, α_t, α_m and α_f are coefficients of the mixed, female and male models that control the importance of their corresponding losses.

5.1.3. Variant of Focal Loss

In the variant of the focal loss in [27], a modulating factor $(1-p_i)^\gamma$ is added to the cross entropy loss, with a tunable focusing parameter $\gamma \geqslant 0$. The operation is:

$$FL(p_i) = -\alpha_i(1-p_i)^\gamma \log(p_i) \tag{3}$$

The above scores (Eq. 3) are used in the search for the correct class for a sentence. We extend the equation to leverage predictions of the male and female models, and to reduce the relative importance of the most biased examples from each group. In this case, we propose a variant of Eq. 4, and the loss function is computed via:

$$VFL(p_i) = -\alpha\left(1-p_i^f-p_i^m\right)^\gamma \log(p_i^t) \tag{4}$$

The parameters α and γ are determined in such a way that the emotion prediction on held-out data is optimized.

5.2. Adversarial Training

In this section, we describe our adversarial training approach to mitigating the effects of gender bias.

In a nutshell, an adversarial training approach [19] involves simultaneous training of two network models, generator G and discriminator D. The generator G tries to generate noisy data aimed to fool D, while the discriminator D aims at classifying the real data and the fake data generated from G. Through the combination of the two learning processes, the G and D models facilitate each other interactively to individually reach their goals.

We use the idea of [19] and jointly train emotion detection and gender identification models using adversarial loss. The main difference is that we do not explore a generator but rather to adopt transfering representations of emotion detection into gender prediction. On the other hand, our learning algorithm is similar to multi-task learning, which introduces additional training objectives to a learning system to bias the learner with a broader understanding through solving related tasks. The end goal is to improve performance on a set of primary tasks through the inductive bias introduced by the additional tasks [8]. In multi-task learning, the goal is to minimize the loss across all tasks. However, our aim is to reduce the performance of gender prediction models and simultaneously improve the robustness of emotion detection models. For that purpose, we consider the first k layers of the network for emotion detection as an encoder, which maps the emotion features X_i to representation R_i. Then, the representations R_i are fed into a decoder network to output the emotion posteriors and are parts of input of another decoder network to classify the gender. But, the above representations would be noise for gender prediction. The relationship between the emotion detection and the gender identification model is given by

$$P_r(y,g|x) = \frac{p(y|x,g)\cdot p(g|x+noise)}{\sum_{y\prime,g\prime} p(y\prime|x,g)\cdot p(g\prime|x+noise)} \tag{5}$$

where X consists of the set of sentences, $X = x_1, x_2, \ldots, x_N$, $Y = y_1, y_2, \ldots, y_N$ indicates the set of all possible labeled emotion classes of X and $G = g_1, g_2, \ldots, g_N, g_i \in 0, 1$ refers to gender information of X.

The goal of joint training is to learn $X \rightarrow Y, G$, obtaining worse performance on gender prediction while the emotion detection is optimal. That is, among all possible target emotion labels, we will choose the label with the higher probability but lower probability of gender:

$$\widehat{Y}(X,G) = \arg\max_Y P_r(Y|X,G),$$

$$\widehat{G}(X+noise) = \arg\min_G P_r(G|X+noise) \tag{6}$$

Model. Our emotion detection models are based on CNN and Transformer, described in Section 4. In our framework, the first k layers of the network for CNN and the encoder of the transformer are viewed as encoders to generate a set of N representations

$R_n(Y,X), n = 1,\ldots,N$. Then, the representations R_i are fed into a decoder network to output the emotion posteriors $P(Y_i|X_i)$ simultaneously go through an adversary A which attempts to predict gender information g_i. A are learnable and can be implemented by neural networks. We will describe it below. The model is trained under the hybrid loss of $\mathscr{L}_E$ and $\mathscr{L}_A$, where $\mathscr{L}_E$ is the loss of emotion detection and $\mathscr{L}_A$ is the loss of adversary. During training, when the adversary A is optimal, it may be frozen, and the inputs may continue to be changed so as to drop the accuracy of the adversary. By tuning the entire pipeline from end to end, E will find the optimal task-specific features, but to the disadvantage of A. That is to say, our ultimate goal is for E to predict label y_i accurately and for A using an intermediate representation R_i to predict g_i poorly.

Adversary. Now that we have the framework of all pipelines, we can continue with designing the adversary. As different models have different abilities of processing data, we consider to develop three different adversaries.

- Linear model (Linear). We use 2 linear layers to generate a less complicated adversary.
- LSTM model (Adversarial_1). We create an adversary model with 1 embedding layer, 1 lstm layer and 2 linear layers. The LSTM model aims to map the intermediate convolutional representations C consisting of $c_1, c_2, \ldots, c_N$ into a sequence of hidden states $h_1, h_2, \ldots, h_N$ through a single-layered bidirectional LSTM. $H = BiLSTM(X + C) = [h_1, h_2, \ldots, h_N], H \in R^{d\times N}$, where d is the size of hidden layers and N is the length of the given sentence.
- Attention-based LSTM model (Adversarial_2). This adversary model integrates the attention mechanism into the LSTM model. After the LSTM produces a matrix H, we apply mean pooling and concatenate the results $\bar{h}$ with the hidden state of the last time step h_N. The attention layer aims to learn a normalized weight vector $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_N$ and a weighted hidden representation δ from H. $M = tanh(\bar{h} \oplus h_N), \alpha = softmax(w^T M), \delta = H\alpha^T$, where $\oplus$ denotes the concatenation operator.

Adversarial Training. We now describe our training algorithms to determine the model parameters. Every model has a specific set of free parameters. For example, the parameters for CNN consist of each layer's weights and biases. Thus, we first try to seek the training settings and hyperparameters that achieve good accuracy on the validation set.

The emotion classifier E is trained to predict the emotional label given an input sentence by using the baseline CNN or Transformer models described in Section 4, and we aim to minimize the cross-entropy loss $\mathscr{L}_E$ associated with predicting the emotional classes over the training data (X, Y). We consider the loss of the form:

$$\mathscr{L}_E(X;\theta_e) = -\sum_{i=1}^{K} logP(y_i|X;\theta_e)$$

For the adversary, it is trained to predict the gender information $g_i \in 0, 1$ given an input sentence x_i, and we initially aim to minimize the cross-entropy loss $\mathscr{L}_A$ associated with predicting the gender:

$$\mathscr{L}_A(X;\theta_a) = -\sum_{i=1}^{G} logP(g_i|X;\theta_a)$$

The final parameter values of the above models serve as the starting point for the adversarial nets. In the adversarial nets, the outputs of the first k CNN layers or the encoder of the transformer, C, are chosen as the feature vectors and are concatenated with word embedding as input features of the adversary model A, and the adversarial loss form is:

Algorithm 1: Adversarial Gender De-biasing

```
Initialize G, D with random weights.
for i: 1 to epochs do
  for minibatch B ⊂ X do
    pretrain emotion detection model
    minimize g_t ← ∇_{θ_e} L_E
    pretrain adversary model
    minimize g_adv ← ∇_{θ_a} L_A
  end for
end for
for i: 1 to iterations do
  for minibatch B ⊂ X do
    train adversary model
    maximize g_adv ← ∇_{θ_a} L_A
    train emotion detection model
    minimize g_t ← ∇_θ L
  end for
end for
```

$$\mathscr{L}_A(X + C;\theta_a) = -\sum_{G} logP(g_i|X + C;\theta_a)$$

We jointly train E and A by optimizing the prediction of E for y while simultaneously being penalized as A gets better at predicting g. That is, the emotion classifier tries to minimize its loss over the task specific predictions while A attempts to increase the adversary loss. Hence, the loss function is a weighted combination of $\mathscr{L}_E$ and $\mathscr{L}_A$. The detailed adversarial training algorithm is described in Algorithm 1.

In the absence of any knowledge on how to combine two losses models, our prior for combining two losses is:

$$\mathscr{L} = \lambda\mathscr{L}_E + (1 - \lambda)\mathscr{L}_A \tag{7}$$

Also, we combine $\mathscr{L}_E$ and $\mathscr{L}_A$ by using another different equation:

$$\mathscr{L} = \mathscr{L}_E - \lambda\mathscr{L}_A \tag{8}$$

5.3. Evaluation Metrics

We evaluate the results using two metrics: Accuracy and TPR (True Positive Rate) related to "Equality of Odds". For emotion detection tasks, we need to examine whether the learned target model maintains satisfactory performance by computing the accuracy: the higher the better. For the influence of gender bias, similar to [11], we define TPR-GAP which is the absolute value of the difference between the performance of female and male: the lower the better.

$$TPR - GAP = |TPR_f - TPR_m| \tag{9}$$

Specifically, we quantify this criterion by computing the difference in true positive rate (TPR) for each class, and further by averaging those quantities.

5.4. Experiments

We present in this section results of experiments involving general approaches and adversarial training.

5.4.1. Hyperparameters

We perform experiments on the ISEAR and CrowdFlower datasets. We split these two datasets into the training, validation and test data sets respectively, and the ratio of them are at 80:10:10. In the process, splitting guarantees that the test data contains an equal amount of sentences from both genders (male and female).

The base emotion detection model uses CNN and Transformer. The minibatch size was set to 64. For CNN, we fix the same regularization of $\lambda = 10^{-4}$ for all parameters. We also cross-validated on AdaGrad's learning rate which was eventually set to $\alpha = 0.1$ and word vector size (300). For Transformer, we use the AdamW optimizer (learning rate = 5e-5, epsilon value = 1e-8). We repeatedly conduct experiments on random splits, and the final results are averaged.

Our standard training schemes on general methods are: we set $\alpha = 0.2$ in the training of Introducing Weights (Intro_Weights); For Variant of Focal Loss, we set $\gamma = 6.0$ and $\alpha = 0.2$; and we choose Eq. 8, and $\lambda = 0.5$ in the Adversarial_1 and Adversarial_2 methods. Later in this section, we evaluate the effect of these parameters on Accuracy and TPR-GAP.

5.4.2. Comparison with other work

We compared our proposed methods with the other debiasing methods.

- Original training scheme: The basic emotion detection model trained on the original data, without any debiasing attempt.
- Debiases embeddings (GN-GloVe). Word embeddings can encode gender biases and correct word embeddings. In this work, we substitute the pretrained GloVe with Gender-neural Global Vectors [59] to verify the effectiveness of the debiased word embeddings in our tasks.
- Data Augmentation method (Data-Aug): We first swap all the gendered words using a bidirectional dictionary of gender pairs described by [29]. Then, we train the models using the augmented datasets.
- Adversarial network architecture (LSTM-MLP). We implement an adversarial network architecture similar to [15] which consists of a one layer LSTM network for representations and multi-layer perceptrons for the classifier and the adversarial. In our experiment, we construct 3 layers of perceptrons.

5.4.3. Results and Analysis

As shown in Table 6 and Table 7, the best accuracy of both CNN and Transformer are obtained by Data Augmentation method on both datasets, and this method also significantly reduce TPR-GAP from 0.063 to 0.033 with CNN and from 0.072 to 0.027 with Transformer on the ISEAR data, from 0.038 to 0.013 with CNN and from 0.029 to 0.012 using Transformer on the CrowdFlower data. Our adversarial training methods (Adversarial_1 and Adversarial_2) are also successful, producing TPR-GAP decreases of around 0.02 on both datasets, and these decreases do not significantly harm accuracy with below 1.0 compared to Original. That is, these two methods are able to reduce gender gap, while suffering a little bit of accuracy degradation. We see that Product_of_Experts and Variant_Focal_Loss achieve better results in terms of TPR-GAP, generating about 0.02 and 0.03 TPR-GAP decreases on the ISEAR data, respectively. However, they harm the accuracy, especially under Transformer architecture (e.g., 51.58% for Product_of_Experts compared to 71.82% for Original on the ISEAR data). These two methods show similar patterns on the CrowdFlower data. Introduce_Weight method also reduces accuracy. GN-Glove and Linear achieve better results in terms of TPR-GAP, achieving about 0.01 drop on both datasets but lose more accuracy. LSTM-MLP performs TPR-GAP reduction, but decreases the accuracy. Finally, we note that Adversarial_1 and Adversarial_2 achieve better bias mitigation between 0.02–0.03 but perform less accuracy decrease (with below 1.0 point) compared to other methods.

Table 6
Comparison of various methods for reducing gender bias on the ISEAR data.

	Method	Accuracy	TPR-GAP
CNN	Original	60.40	0.063
	P_o_E	54.62	0.042
	Intro_Weights	58.67	0.101
	Variant_Focal_Loss	59.54	0.047
	Linear	57.37	0.067
	Adversarial_1	60.98	0.043
	Adversarial_2	59.66	0.034
	GN-Glove	57.73	0.057
	Data-Aug	67.06	0.033
Transformer	Original	71.82	0.072
	P_o_E	51.58	0.044
	Intro_Weights	49.48	0.065
	Variant_Focal_Loss	54.41	0.036
	Linear	67.69	0.068
	Adversarial_1	70.64	0.036
	Adversarial_2	71.24	0.031
	Data-Aug	75.95	0.027
	LSTM-MLP	56.07	0.054

Table 7
Comparison of various methods for reducing gender bias on the CrowdFlower data.

	Method	Accuracy	TPR-GAP
CNN	Original	36.88	0.038
	P_o_E	35.65	0.010
	Intro_Weights	33.02	0.020
	Variant_Focal_Loss	37.00	0.010
	Linear	35.26	0.032
	Adversarial_1	36.66	0.014
	Adversarial_2	37.39	0.020
	GN-Glove	35.87	0.029
	Data-Aug	37.61	0.013
Transformer	Original	43.89	0.029
	P_o_E	37.22	0.020
	Intro_Weights	31.85	0.033
	Variant_Focal_Loss	39.13	0.015
	Linear	41.82	0.023
	Adversarial_1	43.25	0.014
	Adversarial_2	43.67	0.014
	Data-Aug	44.01	0.012
	LSTM-MLP	32.94	0.027

5.4.4. Effect of Parameters

In this section, we try to adjust several critical hyperparameters, i.e., the loss coefficients in Eq. 2 to obtain good results, and explore the influence of these hyperparameters to the performance of emotion detection models. In this experiment, we employ CNN as our emotion detection model to conduct experiments.

Introducing Weights. Since there are three hyperparameters, we evaluate their influence independently. Firstly, we vary the value of α_t without the female and male models losses, in Figs. 1 blue line. On the task of ISEAR data, we see the performance declines when α grows from 0.1 to 0.4, but TPR-GAP increases in this range. However, the results of the CrowdFlower data show different trends, resulting in the increase of both evaluation metrics. From 0.5 to 0.8, it shows fluctuation on both data sets. When α is larger than 0.8, the performance and TPR-GAP reduce on the ISEAR data, while TPR-GAP increases on the CrowdFlower data. Thus, a lower value for the mixed model (e.g., 0.2) may be preferable to achieve good fairness without too heavy performance loss.

Then, we alter the value of α_f under $\alpha_t = 0.2$ and without the male-model loss. The results are shown in Figs. 1 red line. From these results, we cannot find that the gap decreases with the increasing of α, and the change of the performance is relatively small on both datasets. Thus, the same value for α_f as α_t (0.2) can also achieve a good tradeoff between fairness and performance.

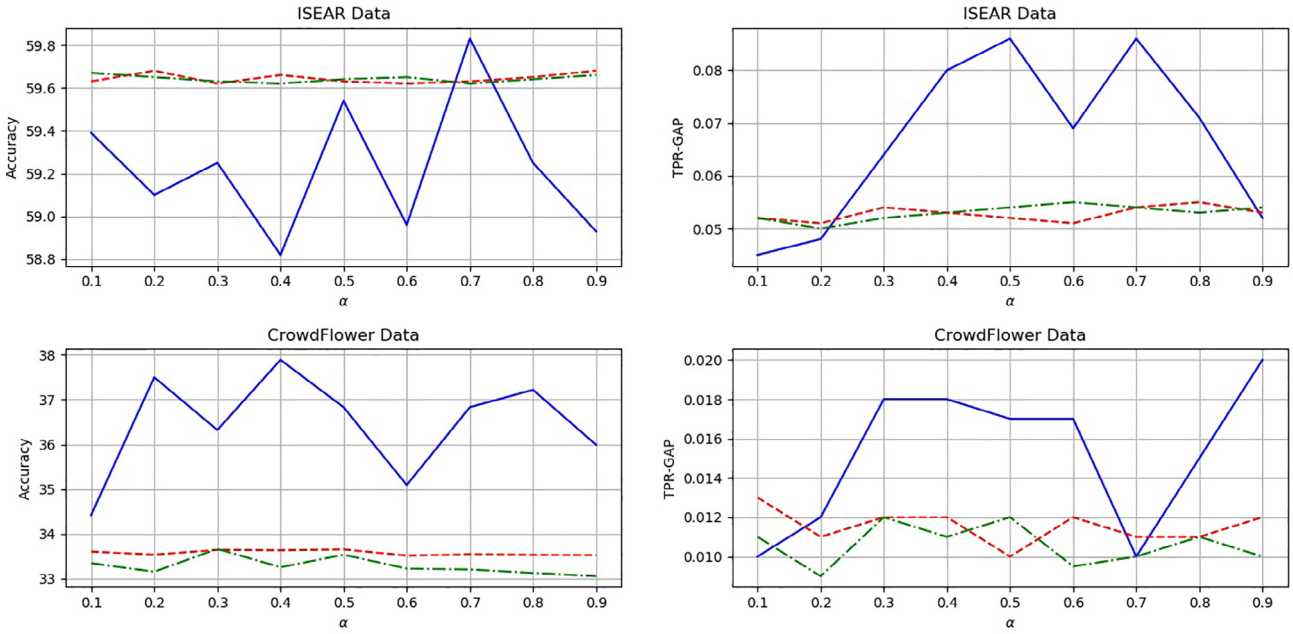

Fig. 1. Introducing Weights: Accuracy and TPR-GAP under different α.

Finally, we activate the male-model and vary α_m under $\alpha_t = \alpha_f = 0.2$. The results are shown in Figs. 1 green line. We find it has a similar trend with female-model. Thus, we choose $\alpha_t = \alpha_f = \alpha_m = 0.2$ to avoid too heavy effort on hyperparameter searching.

Variant of Focal Loss. In Figs. 2, we compare the results obtained by using different γ and α using Variant of Focal Loss. On the ISEAR data, when $\gamma = 2.0$ the improvement of accuracy increases when α grows to 0.4. However, its decrease starts from this point. There are greater fluctuations among $\gamma = 2.0$ than among $\gamma = 4.0$ and $\gamma = 6.0$ with regard to TPR-GAP. When $\gamma = 4.0$ and $\gamma = 6.0$, the accuracies of models have a similar trend between $\alpha = 0.1$ and 0.4, as well as after 0.6. TPR-GAP scores present a similar trend among all α values. Thus, when we set $\gamma = 4.0$ or $\gamma = 6.0$, a proper range of α (0.1–0.2) or (0.5–0.6) is a way to avoid gender bias with good accuracy.

On the CrowdFlower data, $\gamma = 4.0$ and $\gamma = 6.0$ have the similar ranges of accuracy, with between $\alpha = 0.1$ and 0.7. However, the results of $\gamma = 2.0$ among all α values are not fluctuated as $\gamma = 4.0$ and 6.0. In terms of TPR-GAP, these three γ lines show upward trend in the range of α: (0.1–0.3) and (0.7–0.8), $\gamma = 2.0$ and $\gamma = 4.0$ present similar trend in comparison to $\gamma = 6.0$.

In general, to reduce the effect of gender bias, we can set $\gamma = 4.0$ or $\gamma = 6.0$, and select α from 0.1 to 0.3.

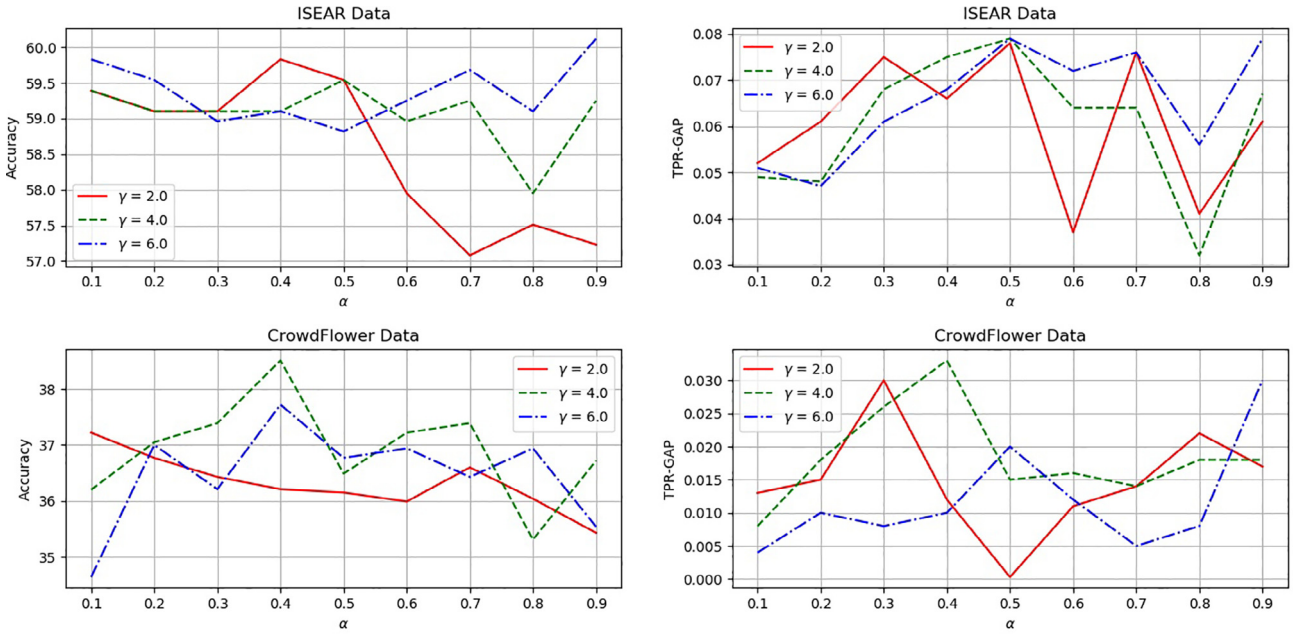

Fig. 2. Variant of Focal Loss: Accuracy and TPR-GAP under various γ and α.

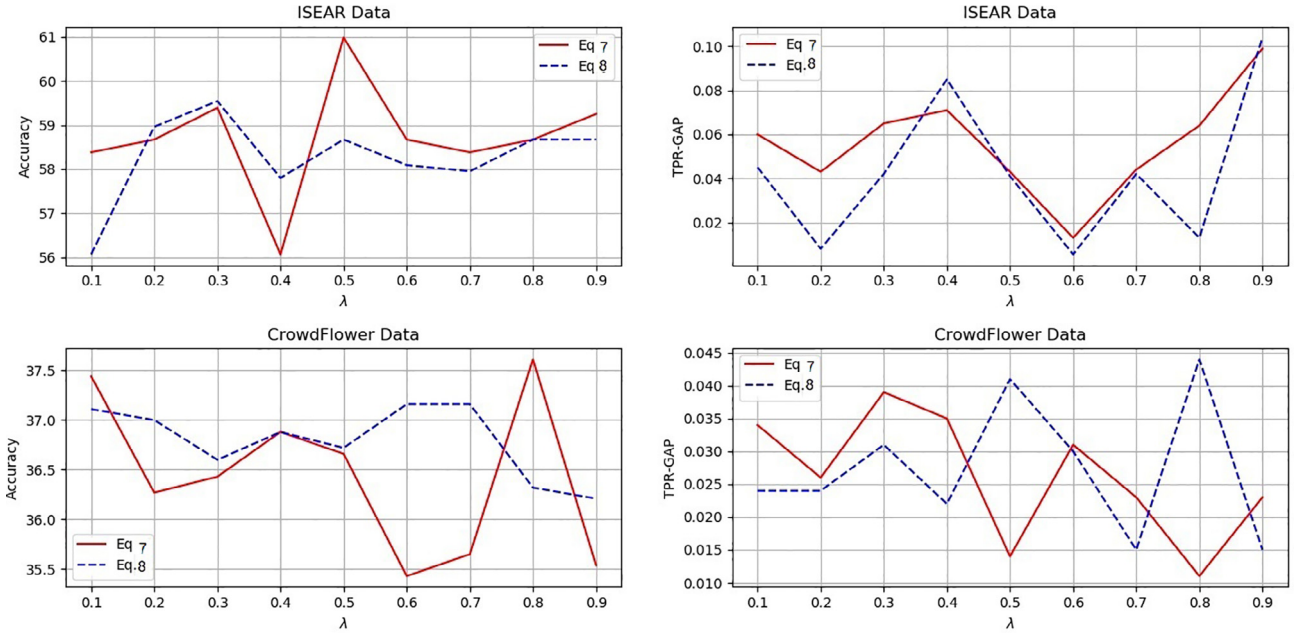

Fig. 3. adversarial training: Accuracy and TPR-GAP under different combination equations and λ.

adversarial training. We also compare our combination methods using Eq. 7 and Eq. 8, and the results of different λ. Figs. 3 shows the results obtained on different combination equations and λ. In general, both equations have shown a similar trend with regard to accuracy and tpr-gap over all λ values on the ISEAR data. Both equations achieve better accuracy and tpr-gap when $\lambda = 0.2$ and 0.6. Typically, we see that Eq. 7 achieves slightly better results than Eq. 8 when $\lambda = 0.2$, but by setting $\lambda = 0.6$, Eq. 8 gives better results than Eq. 7. However, they show more fluctuations on the CrowdFlower data than the ISEAR data. We also observe that between $\lambda = 0.2$ and 0.4, Eq. 7 and Eq. 8 show similar trend, but Eq. 8 outperforms Eq. 7 with $\lambda = 0.5$ and 0.8. Relatively good results of Eq. 7 are generated in a range of λ: (0.2–0.4) and (0.6–0.7).

It seems that the proposed architecture achieves better accuracy and tpr-gap when λ is between 0.2 and 0.4 on both datasets.

6. Conclusion

We evaluated gender bias in the emotion detection task and presented various methods to mitigate its impact. We also prepared datasets labeled by both emotion classes and gender information. To evaluate gender bias, we tested the differences in model performance when predicting on sentences from male versus female data; we further studied the role of different emotion categories in gender identification to analyze how emotion features differ between men and women. The results show that there exists gender bias in emotion detection: the models trained on female data often achieve better results than the male-models, and female-model and male-model reported the opposite trends on the recognition of some emotions.

We considered three general approaches: products of experts, introducing weights and variants of focal loss, as well as adversarial training to defend against gender bias. We reported on experiments that showed decreases of difference over baseline models in terms of TPR-GAP. These experiments demonstrate that adversarial training can be successfully exploited to overcome gender biases of the emotion detection model.*Limitations and Future work* The main shortcoming in this paper concerns the size and variety of corpus and the training procedure of adversarial training we currently apply. Due to lack of corpus labeled with both emotions and gender, we applied two datasets for our research. However, the size of the ISEAR data is limited and the quality of the CrowdFlower is not guaranteed even though we cleaned it. To keep the memory requirements manageable, we restricted the epoch and iterations of the adversarial training phase. Ideal, we would like to adversarial training involves both finding the parameters of an emotion detection model that maximize its classification accuracy and finding the parameters of an adversary that maximize its loss. We are currently investigating the applicability of adversarial techniques to reduce the gap between female and male, simultaneously there is less loss of accuracy. However, we were not able to obtain the best results after limiting the training phase.

In future work, we plan to explore more emotion datasets which contain more labels such as gender, ethics, and age etc, and further investigate whether these protected attributes affect the performance of emotion detection. Moreover, we expect to overcome the constraints of the currently implemented adversarial training by developing a training criteria which makes the optimization in terms of accuracy and true positive rate. We also plan to continue widening the scope of our study - for example, expanding our methods to include examining gender biases in scientific writings, uncertainty claims detection etc.

CRediT authorship contribution statement

Odbal: Methodology, Software, Validation, Writing - original draft. **Guanhong Zhang:** Methodology, Supervision, Funding acquisition, Funding acquisition. **Sophia Ananiadou:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the Key Research and Development Plan of Anhui Province (202104d07020006), the Natural Science Foundation of Anhui Province (2108085MF223), the Natural Science Research Project in Universities of Anhui Province (KJ2021A0991), the Key Research and Development Plan of Hefei (2021GJ030) and the China Scholarship Foundation (201804910294). Additionally, the authors would like to thank reviewers for their valuable comments and suggestions.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudık, J. Langford, H. Wallach, A reductions approach to fair classification, in: Proceedings of the 35th International Conference on Machine Learning, 2018.
[2] Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.J., Srivastava, M.B., Chang, K.W., 2018. Generating natural language adversarial examples. In EMNLP.
[3] Anthony Mulac, J. James, P.G. Bradac, Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences, Human Communication Research 27 (2001) 121–152.
[4] Bartl, M., Nissim, M., Gatt, A., 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. In Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing at COLING 2020 arxiv.org/abs/2010.14534.
[5] S.L. Blodgett, S. Barocas III, H.D. Wallach, H., Language (technology) is power: A critical survey of "bias in nlp, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5454–5476.
[6] T. Bolukbasi, K.W. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in Neural Information Processing Systems (2016) 4349–4357.
[7] L.A.M. Bostan, R. Klinger, An analysis of annotated corpora for emotion classification in text, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2104–2119.
[8] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.
[9] L. Cheng, A. Mosallanezhad, Y.N. Silva, D.L. Hall, H. Liu, Mitigating bias in session-based cyberbullying detection: A non-compromising approach, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 2158–2168.
[10] C. Clark, M. Yatskar, L. Zettlemoyer, Don't take the easy way out: Ensemble based methods for avoiding known dataset biases, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 4069–4082.
[11] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A.T. Kalai, Bias in bios: A case study of semantic representation bias in a high-stakes setting, ACM Conference on Fairness, Accountability, and Transparency (ACM FAT) (2021) 120–128.
[12] Dixon, L., Li, J., Sorensen, J., NithumThain, Vasserman, L., 2017. Measuring and mitigating unintended bias in text classification, in. In AAAI.
[13] X. Dong, Y. Zhu, Z. Fu, D. Xu, G. de Melo, Data augmentation with adversarial training for cross-lingual nli, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 5158–5167.
[14] EV, Intracultural variation of semantic and episodic emotion knowledge in estonian, Trames 10 (169–189) (2006) 2.
[15] Y. Elazar, Y. Goldberg, Adversarial removal of demographic attributes from text data, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 11–21.
[16] Gajane, P., Pechenizkiy, M., 2018. On formalizing fairness in prediction with machine learning. arXiv 1710.03184v3.
[17] Goldfarb-Tarrant, S., Marchant, R., Sanchez, R.M., Pandya, M., Lopez, A., 2021. Intrinsic bias metrics do not correlate with application bias. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 1926–1940.
[18] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, NIPS (2014) 2672–2680.
[19] Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. Machine Learning arXiv:1412.6572. https://arxiv.org/abs/1412.6572. version 3.
[20] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems (2016) 3315–3323.
[21] Hashimoto, T.B., Srivastava, M., Namkoong, H., Liang, P., 2018. Fairness without demographics in repeated loss minimization. arXiv 1806.09010v2.
[22] G.E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Computation 14 (1771–1800) (2002) 8.
[23] D. Hovy, Demographic factors improve classification performance, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, pp. 752–762.
[24] Kiritchenko, S., Mohammad, S.M., 2018. Examining gender and race bias in two hundred sentiment analysis systems. Computation and Language arXiv:1805.04508. https://arxiv.org/abs/1805.04508. version 1.
[25] Kleinberg, J., Mullainathan, S., Raghavan, M., 2016. Inherent trade-offs in the fair determination of risk scores. arXiv 1609.05807.
[26] Le, T., Wang, S., Lee, D., 2020. Generating malicious comments to attack neural fake news detection models. In IEEE ICDM.
[27] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
[28] H. Liu, W. Wang, Y. Wang, H. Liu, Z. Liu, J. Tang, Mitigating gender bias for neural dialogue generation with adversarial learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 893–903.
[29] Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A., 2019. Gender bias in neural natural language processing. Machine Learning arXiv:1807.11714. Version 2.
[30] R. Masumura, Y. Shinohara, R. Higashinaka, Y. Aono, Adversarial training for multi-task and multi-lingual joint modeling of utterance intent classification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.
[31] M.R. Mehl, J.W. Pennebaker, The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations, Journal of Personality & Social Psychology 84 (2003) 857–870.
[32] R. O'Kearney, M. Dadds, Developmental and gender differences in the language for emotions across the adolescent years, Cognition and Emotion 18 (913–938) (2004) 7.
[33] Park, J.H., Shin, J., Fung, P., 2018. Reducing gender bias in abusive language detection, in: Empirical Methods of Natural Language Processing.
[34] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[35] Rabeeh Karimi Mahabadi, J.H. Yonatan Belinkov, End-to-end bias mitigation by modelling biases in corpora, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8706–8716.
[36] M.T. Ribeiro, S. Singh, C. Guestrin, Semantically equivalent adversarial rules for debugging nlp models, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 856–865.
[37] A. Romanov, M. De-Arteaga, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. Rumshisky, A.T. Kalai, What's in a name? reducing bias in bios without access to protected attributes, Proceedings of NAACL-HLT 2019 (2019) 4187–4195.
[38] R. Rudinger, J. Naradowsky, B. Leonard, B.V. Durme, Gender bias in coreference resolution, Proceedings of NAACL-HLT 2018 (2018) 8–14.
[39] Russell, C., Kusner, M.J., Loftus, J.R., 2017. When worlds collide: Integrating different counterfactual assumptions in fairness. 31st Conference on Neural Information Processing Systems (NIPS 2017).
[40] M. Sato, J. Suzuki, Matsumoto Y. Shindo, Interpretable adversarial perturbation in input embedding space for text, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2019, pp. 5158–5167.
[41] Saunders, D., Byrne, B., 2020. Reducing gender bias in neural machine translation as a domain adaptation problem, p. 7724–7736.
[42] K.R. Scherer, H.G. Wallbott, Evidence for universality and cultural variation of differential emotion response patterning, Journal of Personality and Social Psychology 66 (1994) 310–328.
[43] G. Stanovsky, N.A. Smith, L. Zettlemoyer, Evaluating gender bias in machine translation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1679–1684.
[44] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.W., Wang, W.Y., 2019. Mitigating gender bias in natural language processing: Literature review. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1630–1640.
[45] Søgaard, A., Ebert, S., Bastings, J., Filippova, K., 2021. We need to talk about random splits. EACL 2021 arXiv:2005.00636.
[46] C. Tang, K. Ross, N. Saxena, R. Chen, Whats in a name: a study of names, gender inference, and gender behavior in facebook, in: 16th International Conference on Database Systems for Advanced Applications, 2011, pp. 344–356.
[47] Telegraph, T., 2016. Microsoft deletes "teen girl' ai after it became a hitlter-loving sex robot within 24 hours. https://goo.gl/mE8p3J.
[48] M. Thelwall, Gender bias in sentiment analysis, Online Inf. Rev 42 (1) (2018) 45–57.
[49] Utama, P.A., Moosavi, N.S., Gurevych, I., 2020. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance arXiv:2005.00315v1. https://arxiv.org/abs/2005.00315.
[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, ukasz Kaiser, Polosukhin, I., Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[51] Verma, S., Rubin, J., 2018. Fairness definitions explained. 2018 ACM/IEEE International Workshop on Software Fairness.
[52] Wallace, E., Feng, S., Kandpal, N., Gardner, M., Singh, S., 2021. Universal adversarial triggers for attacking and analyzing nlp. arXiv 1908.07125v3.
[53] Wu, Y., Bamman, D., Russell, S., 2017. Adversarial training for relation extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1778–1783.
[54] Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P., 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. arXiv 1610.08452v2.

[55] R. Zemel, Y.L. Wu, K. Swersky, T. Pitassi, Learning fair representations, in: International Conference on Machine Learning, 2013.
[56] B.H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI and Society, 2018, pp. 335–340.
[57] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.W. Chang, Men also like shopping: Reducing gender bias amplification using corpus-level constraints, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2979–2989.
[58] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W., 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods, in: NAACL.
[59] Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W., 2018b. Learning gender-neutral word embeddings, in: EMNLP.
[60] Y. Zhou, X. Zheng, C.J. Hsieh, K.W. Chang, X. Huang, Defense against synonym substitution-based adversarial attacks via dirichlet neighborhood ensemble, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 5482–5492.

Odbal received her B.A. and M.Sc. from Inner Mongolia University in 2004 and 2007, respectively. In 2017 she was awarded the Ph.D. degree from University of Science and Technology of China. She was a researcher in the Institute of Intelligent Machines, Chinese Academy of Sciences from 2007 to 2021. Between 2019 and 2021, she was a visiting postdoctoral fellow in the School of Computer Science at the University of Manchester. Currently, she is employed at Anhui Vocational and Technical College. Her research focuses on the area of emotion detection in social media, natural language processing and machine learning.

Guanhong Zhang is currently a professor at Hefei university. He received his B.A. from Anqing Normal University in 1998, and M.Sc.from Inner Mongolia University in 2003. He received his Ph.D. degree from Technological University Dublin in 2021. His research focuses on the area of machine learning, sentiment analysis, machine translation and deep learning.

Sophia Ananiadou is Professor in the School of Computer Science at the University of Manchester and is the director of The National Centre for Text Mining (NaCTeM) the only centre of its type in the world. She has led the development of the text mining tools and services currently used in NaCTeM with the aim of providing scalable text mining services: information extraction, intelligent searching, association mining, etc. She has received the IBM UIMA innovation award 3 consecutive times and is also a Daiwa award winner.