Mitigate Gender Bias using Negative Multi-Task Learning

Liyuan Gao (liygao@ttu.edu)
Texas Tech University
Huixin Zhan
Texas Tech University
Austin Chen
Lubbock High School
Victor Sheng
Texas Tech University

Research Article

Keywords: Gender bias, Selective privacy-preserving, Negative multi-task learning, Classification

Posted Date: September 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2024101/v1

Mitigate Gender Bias using Negative Multi-Task Learning

Liyuan Gao[1*], Huixin Zhan[1], Austin Chen[2] and Victor Sheng[1]

[1*]Computer Science, Texas Tech University, 2500 Broadway, Lubbock, 79409, Texas, USA.
[2] Lubbock High School, 2004 19th St, Lubbock, 79401, Texas, USA.

*Corresponding author(s). E-mail(s): liygao@ttu.edu;
Contributing authors: huixin.zhan@ttu.edu;
austinchen2005@gmail.com; victor.sheng@ttu.edu;

Abstract

Deep learning models have shown their great performances in natural language processing tasks. While much attention has been paid towards improvements in utility, privacy leakage and social bias are two major concerns arising in trained models. In this paper, we protect individual privacy and mitigate gender bias on classification models simultaneously. First, we propose a selective privacy-preserving method that only obscure individuals' sensitive information by adding noise on word embeddings. Then we propose a negative multi-task learning framework to mitigate the gender bias which contains a main task and a gender prediction task. The main task uses a positive loss constraint to ensure utility while the gender prediction task applies a negative loss constraint to remove gender-specific features. We analyze two existing word embeddings and evaluate them on sentiment analysis and medical text classification tasks. Our experimental results show that our negative multi-task learning framework can mitigate the gender bias while keeping models' utility on both sentiment analysis and medical text classification.

Keywords: Gender bias, Selective privacy-preserving, Negative multi-task learning, Classification

1 Introduction

Recent developments in Natural Language Processing (NLP) have made significant success on enormous text data. While social biases like racism and sexism may exist in the text data, classifiers which are trained and evaluated on these data will cause unfairness. Sentiment analysis is to find opinions, identify the sentiments through customer opinions towards products and services expressed in social media or review sites [1]. It is widely used within marketing and customer relations management. However, sentiment analysis algorithms may perform differently for males and females.This will mislead decision makings. [2] and [3] have demonstrated that word embeddings which are trained on human-generated text data encode human biases in vector spaces. For example, the word "programmer" is neutral to gender by its definition, but models usually associate the word "programmer" closer with "male" than "female"[4]. Such biases will affect downstream applications. Models trained from the source data not only encode but even amplify the bias present in dataset [5].

Another major concern is how sensitive information should be used during the training and testing phases in a model. Without privacy-preserving methods, some individuals' information might be leaked from a model learned on training data, such as gender, race and age. For example, models trained on medical data may contain information about patients' disease status or other sensitive information [6]. In clinic diagnosis, it is more dangerous that models are biased in the sense that they are much more effective for texts from certain groups of users [7].

To address these problems, in this paper we propose a negative multi-task learning framework to mitigate gender bias while keeping model utility. Traditional multi-task learning frameworks jointly train several related tasks to improve their generalization performance by leveraging common knowledge among them. We use positive loss weights to ensure utility for the main classification task while applying negative loss weights to remove gender-specific features for the gender prediction task. In order to evaluate gender bias of classifiers, we use disparity score to measure the difference of accuracy between males and females. We also propose a selective privacy-preserving method to protect individuals' sensitive information. In Section 4, we conduct experiments on sentiment analysis and medical text classification. We quantify, analyse and mitigate the gender bias on the two tasks. For sentiment analysis, we only apply the negative multi-task learning method, the disparity score between males and females drops a lot compared with baseline models. For medical text classification, we first apply selective privacy-preserving method on sensitive word embeddings, and then use the negative multi-task learning framework to train the model. Our experimental results show that both privacy-preserving and negative multi-task learning methods can reduce the disparity score.

The contributions of this paper are as follows:

- We propose a negative multi-task learning framework to mitigate gender bias for text classification models.
- We selectively explore sensitive information in the text embeddings, and then perturb the information of each individual in the data.
- In order to quantitatively measure the gender bias, we propose disparity score to calculate the difference of model's accuracy between males and females.
- We evaluated the proposed negative multi-task learning framework on sentiment analysis and medical text classification. We also protect the sensitive information on medical data and thoroughly analyse our experimental results.

2 Related Works

2.1 Gender Bias Mitigation Methods

Text corpora used to train NLP models may contain gender, racial and religious biases. Word embeddings trained on these data will keep the bias. Gender bias is the most common bias which exits in many NLP applications, several works have revealed gender bias in various NLP tasks [8–10].

[4] proposes a novel training procedure for learning gender-neutral word embeddings. They generate a Gender-Neutral variant of GloVe (GN-GloVe), which tries to remove socially-biased information in certain dimensions while ensuring that other dimensions are free from this gender effect. Biases are not only contained in text data and embeddings, they may also exist in learned models even if the data itself is not biased. [11] have investigated gender biases in machine translation, it is true that social gender assignment influences translation choices.

[12] measures how removing a small part of the training corpus would affect the resulting bias. They perturb the training corpus to see what affects resulting embedding bias most, and then remove them in the training corpus. In this paper, we investigate whether a multi-task learning framework can be used to mitigate gender bias for text classification tasks. The multi-task learning frameworks proposed by previous works are used to improve the performance for several tasks at the same time. However, we use the gender prediction as a auxiliary task and apply negative loss weights to reduce the gender effect for the main task. The details are described in Section 3.2.

2.2 Privacy-Preserving Text Embeddings

Text embeddings are distributed representations of text in an n-dimensional space. Word2vec [13] and GloVe [14] are text embedding generative models which can learn word embeddings efficiently from a large text corpus containing wealth semantic relatedness between words. These word embeddings are used for solving most NLP problems. Recent research has shown that it is possible for attackers to infer information about their training data through learned

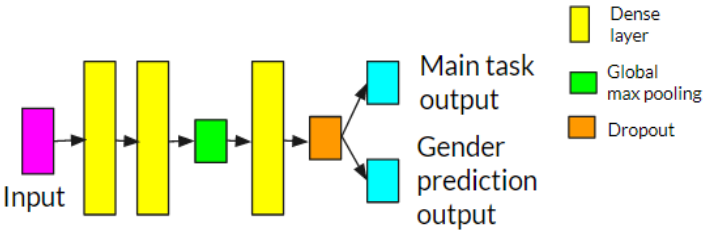

Fig. 1 Multi-task learning structure

models [15]. [16] proves that private information can be recovered only using text embeddings.

In order to protect individuals' sensitive information, a lot of methods have been proposed. Differential privacy (DP) is a mathematical definition of privacy which provides a guarantee between privacy and utility. DP usually injects noise into the data to anonymize data. It also has other forms of corruption such that using a Gaussian sanitizer to sanitize gradients and a amortized moments accountant to keep track of used privacy [17]. However, we usually need to sacrifice some utility accuracy to ensure privacy using DP.

Another previous work is using an adversarial training objective to minimise the risk of adversarial attacks in sensitive information. It can explicitly obscure users' private information [18]. [19] proposes a selective differential privacy method to provide privacy guarantees on the sensitive portion of the data for language model utility.

There is not a common metric on how privacy should be protected. In most situations, the common information doesn't need to be protected. Thus, in this paper we only focus on the sensitive information of each individual. We first detect the sensitive words, and then add noise on corresponding word embeddings to obscure them, so that the original sensitive words will be protected. The details are described in Section 3.3.

3 Methodology

3.1 Multi-task Learning Framework

Supposing there are T tasks, multi-task learning frameworks aim to solve these tasks simultaneously. It usually contains two parts of parameters: shared parameters θ and task-specific parameters $\{\psi_t\}_{t=1}^{T}$. In this paper, the shared layers are 3 dense layers with 128 units. There are a Global max pooling operation behind the second layer and a Dropout (0.5) layer behind the third layer. The multi-task learning framework has two tasks: a main task and a gender prediction task. As shown in Fig.1.

3.1.1 Shared Feature Extraction.

The basic multi-task architectures are to extract some common features in shared lower layers. The multi-task learning will generalize the model better on the tasks by sharing information between related tasks [20].

3.1.2 Task-Specific Layer.

After the shared layers, the remaining layers are split into the multiple specific tasks. The optimization objective of multi-task learning is as follows:

$$\text{Loss} = \sum_{t=1}^{T} \lambda_t(\theta, \{\psi_t\}_{t=1}^{T}) \tag{1}$$

where $\{\psi_t\}_{t=1}^{T}$ are task-specific loss weights, and the constraints $\lambda_t \geqslant 0$ in most previous works. However, in order to remove gender features for the main task, we apply a negative loss constraint for the gender prediction task.

3.2 Mitigate Gender Bias using Negative Multi-task Learning

In this paper, we only consider the "gender" bias, which is a frequently concerned factor in fairness. Algorithm 1 demonstrates the process of negative multi-task learning to mitigate the gender bias. First, we obtain word embeddings from embedding generative models (Word2vec and GolVe in our experiments), and then take the text embeddings as input of the negative multi-task learning model. After having extracted common features through shared layers, there are two outputs: one is the main task classification, and another is the gender prediction. The final loss is: $Loss = L_{main-task}$ - $\lambda^* L_{gender-prediction}$. λ is the gender prediction loss constraint. We can adjust it to balance the accuracy of the main task and the gender bias. We will discuss the impact of the value of λ in Section 4.

The objective of the negative multi-task learning framework is to improve the main task classification accuracy while reducing the gender prediction accuracy. In this way, the text classification model can remove gender-specific features and be distributed without exposing the gender information. This allows the model to avoid learning biases from training data while still being adequately trained to perform the main task.

3.3 Selective Privacy-Preserving Text Embeddings

Machine learning algorithms are used to make decisions in various applications. These algorithms rely on large amounts of sensitive individual information to work properly. The sensitive individual information may include: Name, Address, Email, Phone number, Age, Sex, Marital status, Race Nationality, Religious beliefs, and so on. We first define a sensitive information detecting

Algorithm 1 Mitigate Gender bias using Negative Multi-Task Learning

Input: Trained text embeddings
Initialize: Gender prediction loss weight λ
create batch inputs B
for epoch in 1,2, ..max **do**
  for b_i in B **do**
    Compute main task loss: $L_{main}(\theta)$;
    Compute gender prediction loss: $L_{gender}(\theta)$;
    Final Loss: L(θ) = L_{main} - λ*$L_{gender}(\theta)$
    Compute gradient: $\nabla(\theta)$
    Update model: $\theta = \theta$ - $\eta * \nabla(\theta)$
  end for
end for

function $S(X)$, where X is a word in a dataset. For example, $S(X)$ will search keywords for four kinds of privacy attributes: Gender, Age, Race, and Weight. Each attribute has a list of sensitive words. If a word is in the list, $S(X)$ will return 1, otherwise return 0. If $S(X)$ returns 1, we will use perturb function P to obscure the word embeddings. $P(E) = E + \mathcal{N}(\mu, \sigma, D_E)$, where E is the word embedding of X, and D_E is the dimension of E.

As shown in Algorithm 2, given a dataset D, each sample is a text sequence X_i. We perturb the sensitive word embeddings by adding Gaussian noise. After perturbing text embeddings, the sensitive information will be changed to other non-sensitive word embeddings. For example, for a sentence 'She is 84 years old, 148 pounds history of hypertension and diabetes', after having perturbed its text embeddings, its recovered text is 'the is load years old, diagnosed pounds history of hypertension and diabetes'. The sensitive information have been protected.

Algorithm 2 Selective Privacy-Preserving Text Embeddings

Require: Input texts $\{X_1^t, ..., X_N^t\}$, a max sequence length L, a word embeddings Model, sensitive word detecting function S, word embedding perturb function P.
for each text sequence X_i **do**
  for each word X_i^t **do**
    if S(X_i^t) is true **then** ▷ Detect sensitive word.
      E_i^t=Embedding Model(X_i^t) ▷ Sensitive word embedding.
      E_i^t=P(E_i^t) ▷ Apply perturbation.
    else
      E_i^t=Model(X_i^t) ▷ Non-sensitive word embedding.
    end if
  end for
end for
Output: Selective Privacy-Preserving Text Embeddings

Besides, we compare two word embeddings Word2vec and GloVe as the input of our negative multi-task learning models. Word2vec is one of the most popular techniques to learn word embeddings. It includes a two-layer neural networks which is trained to reconstruct linguistic contexts of words with each unique word in the corpus being assigned a corresponding vector in the space. GloVe is an unsupervised learning algorithm for obtaining vector representations for words and have been shown to perform well across a variety of NLP tasks. It is based on ratios of probabilities from the word-word co-occurrence matrix. It combines the intuitions of count-based models while also capturing the linear structures used by methods like Word2vec. Compared with Word2vec, GloVe does not rely just on local statistics (local context information of words), but incorporates global statistics (word co-occurrence) to obtain word embeddings.

4 Evaluation

In this section, we evaluate the gender bias protection using both the sentiment analysis task and the medical text classification task and analyze the effect of the negative multi-task learning framework on mitigating gender bias. We also compare two word embeddings and apply the selective privacy-preserving method on the medical text classification task.

4.1 Dataset and Settings

4.1.1 Sentiment Dataset.

The sentiment dataset is extracted from TripAdvisor reviews of restaurants in UK. The reviews were authored by males and females. The sensitive information of authors is not displayed with their reviews, so we don't apply our selective-privacy preserving method on this dataset. The review ratings are on a five-point decile scale (10, 20, 30, 40, 50). According to the ratings, we separate the data into positive reviews (review ratings>30) and negative reviews (review ratings<=30), as shown in Table 1.

Table 1 The Characteristics of the TripleAdvisor Dataset

Sentiment type	Male-authored reviews	Female-authored reviews	Total
Positive	800	800	1600
Negative	1200	1200	2400
Total	2000	2000	4000

4.1.2 Medical Dataset.

Medical Transcriptions contain sample medical transcriptions for various medical specialties which were scraped from 'mtsamples.com'. The dataset is highly imbalanced as shown in Fig. 2. In order to mitigate the effect of class imbalance

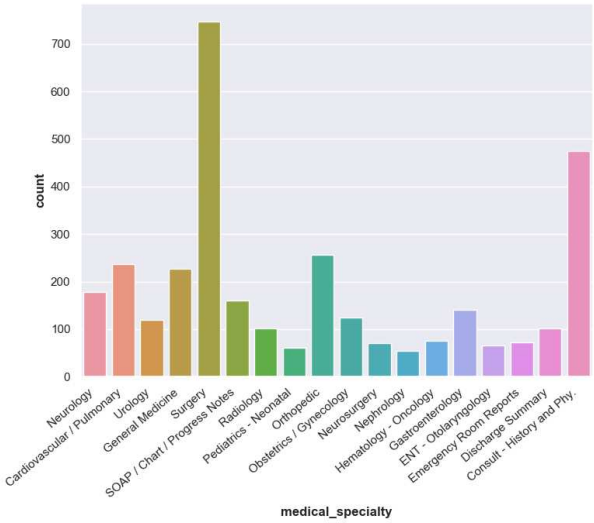

Fig. 2 MTsample dataset statistics (only showing the specialties with more than 50 instances)

problems on experiment results, we picked the most two specialties for binary classification to simplify the task. We first removed any invalid sample (either transcription or label is empty), then transformed all the texts to lower case, deleted punctuations and removed stopwords. The processed dataset statistic is shown in Table 2.

Table 2 The Characteristics of the Medical Dataset

Medical specialty	Male	Female	Total
Surgery	348	399	747
Consult-History and Phy.	185	290	475
Total	533	689	1222

5 Experimental Settings

We use Word2vec and GloVe as the basic word embeddings with 100 dimensionality. The max length of the input sequences to train word embeddings is set to 250. The perturb function employed on sensitive word embeddings uses (0, 1)-Gaussian noise. The default loss constrain λ for gender predication in negative multi-task learning frameworks is set to e^{-5}. All of the models are trained and tested using 5-fold cross-validation to estimate the performance change caused by the optimisation on each set individually. All of the models are trained for 100 epochs with batch size 32. We train single task learning models as the baseline models. We analyse the impact of both the selective privacy-preserving method and the negative multi-task learning framework on

mitigating gender bias. In all the experiments, we compare the models with the same settings.

5.1 Evaluation Metrics

5.1.1 Accuracy Evaluation.

For model utility, we use different metrics to evaluate the sentiment analysis task and the medical text classification task. In order to evaluate each sentiment class separately, we use F1-score to measure sentiment analysis models' accuracy. We calculate the F1-score for both negative and positive sentiments. We use balanced accuracy to evaluate the overall performance of the medical text classification task, defined as follows:

$$\text{Balanced Accuracy} = \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{2}$$

5.1.2 Gender Bias Evaluation.

We measure the gender bias using Disparity Score. Equality of Opportunity Evaluation was proposed by [21]. A predictor $\hat{Y}$ satisfies equality of opportunity with respect to a class y if $\hat{Y}$ and Z are independent conditioned on Y=y. This says that the true positive rates should be the same for males and females. To measure the gender bias similarly with Equality of Opportunity, we use the pairwise difference in accuracy for predictions. The average difference between males and females is described as *Disparity Score*:

$$\textit{Disparity Score} = \frac{1}{k} \sum_{n=1}^{k} (Acc_{female,k} - Acc_{male,k}) \tag{3}$$

where *Acc* is the accuracy of each model built in k-fold cross validation. In our experiment, k=5. *Disparity Score* is 0 means there is no gender bias on the predictions. The closer the disparity score is to 0, the fairer the models are.

5.2 Sentiment Analysis

We test four groups to measure the difference of gender bias between negative sentiment and positive sentiment. In negative multi-task learning framework, the gender prediction loss constrain λ is e^{-5}. The gender differences from sentiment analysis results are shown in Fig. 3, which shows that sentiment analysis models perform differently from males and females. The models' performance on males is less than on females for both negative reviews and positive reviews. That means the sentiment analysis models are better at identifying sentiment from females than from males, so it is more difficult to detect male sentiment. Comparing Word2vec and GloVe embeddings, GloVe has better accuracy than Word2vec in terms of sentiment analysis.

Table 3 reports the disparity score on four groups. Group-1 is negative sentiment disparity score tested on the single-task learning model, Group-2 is

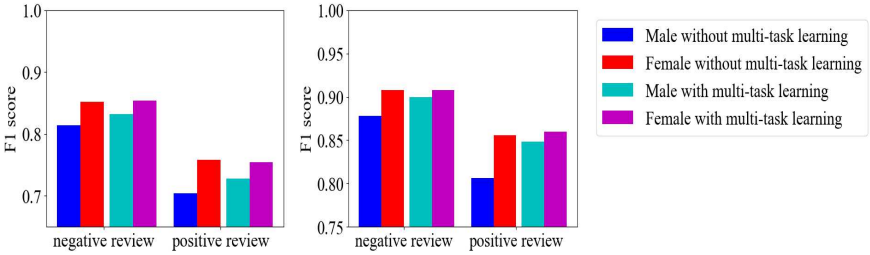

Fig. 3 F1 score of sentiment analysis models for Word2vec (left) and GloVe (right). λ is e^{-5} in the negative multi-task learning framework.

negative sentiment disparity score tested on the negative multi-task learning model, Group-3 is positive sentiment disparity score tested on the single-task learning model, Group-4 is positive sentiment disparity score tested on the negative multi-task learning model. From Table 3, we can see that positive sentiment has higher disparity score than negative sentiment on both Word2vec and GloVe, which means there exits higher gender bias in positive sentiment than that in negative sentiment. This might be that females use more positive words than males which cause easier to detect females' positive sentiment. While in negative sentiment, females and males use closer negative words which has less bias to detect negative sentiment. For Word2vec, negative multi-task learning model's disparity score drops 1.6 (Group-2) on negative sentiment and drops 2.8 (Group-4) on positive sentiment. For Glove, negative multi-task learning model's disparity score drops 2.2 (Group-2) on negative sentiment and drops 3.8 (Group-4) on positive sentiment. GloVe performs better on mitigating gender bias.

Table 3 Disparity score on sentiment analysis (λ is e^{-5} in Group-2 and Group-4)

Embedding type	Negative sentiment		Positive sentiment	
	Group-1	Group-2	Group-3	Group-4
Word2vec	3.8	**2.2**	5.4	**2.6**
GloVe	3.0	**0.8**	5.0	**1.2**

We also test the impact of loss constrains for gender prediction on the negative multi-task learning framework. As shown in Table 4 and Table 5, λ with e^{-6} has the highest disparity score. As the value of λ increases, the disparity score will drop, and gets the lowest disparity score at e^{-5}. An ideal model maximizes the classification performance (measured in terms of F1 score) and minimizes the disparity score (gender gap), so we choose e^{-5} as the default value of λ.

5.3 Medical Text Classification

On the medical text classification task, we investigate the impact of both the selective privacy-preserving method and the negative multi-task learning

Table 4 Negative multi-task learning models' performance with different λs for negative sentiment (left) and positive (right) using Word2vec

λ value	F1 score		Disparity score (%)	λ value	F1 score		Disparity score (%)
e^{-6}	male	0.820	3.4	e^{-6}	male	0.718	4.60
	female	0.854			female	0.764	
e^{-5}	male	**0.832**	**2.2**	e^{-5}	male	**0.728**	**2.60**
	female	**0.854**			female	**0.754**	
e^{-4}	male	0.812	3.2	e^{-4}	male	0.702	4.60
	female	0.844			female	0.748	
e^{-3}	male	0.818	2.6	e^{-3}	male	0.708	3.20
	female	0.844			female	0.740	

Table 5 Negative multi-task learning models' performance with different λs for negative sentiment (left) and positive (right) using GloVe

λ value	F1 score		Disparity score (%)	λ value	F1 score		Disparity score (%)
e^{-6}	male	0.884	1.6	e^{-6}	male	0.822	2.8
	female	0.90			female	0.850	
e^{-5}	male	**0.90**	**0.8**	e^{-5}	male	**0.848**	**1.2**
	female	**0.908**			female	**0.860**	
e^{-4}	male	0.892	1.2	e^{-4}	male	0.832	2.8
	female	0.904			female	0.860	
e^{-3}	male	0.888	1.4	e^{-3}	male	0.828	2.0
	female	0.902			female	0.848	

method for mitigating gender bias. The comparison results of different models are presented in Table 6 and Table 7 respectively. We train Model1 for the single task learning without privacy-preserving handling as the baseline model, Model2 for a single task learning with selective privacy-preserving, Model3 for the negative multi-task learning framework without privacy-preserving handling, and Model4 for the negative multi-task learning framework with selective privacy-preserving.

Our experimental results using Word2vec are presented in Table 6. Comparing Model1 with Model2, we can see that there is a little performance drop because of the noise addition. Comparing Model1 with Model2 and Model3 with Model4 respectively, we can conclude that the selective privacy-preserving method can decrease the disparity score, mitigating gender biases. Furthermore, Model4 has the lowest disparity score while keeping the accuracy. Comparing Model1 with Model3 and Model2 with Model4 respectively, we can see that negative multi-task learning not only mitigates the disparity score but also improves the accuracy. Table 7 shows our experimental results using GloVe. Similar with Word2vec models, the negative multi-task learning

Table 6 Medical text classification disparity score using Word2vec (λ is e^{-5} in Model3 and Model4)

Evaluated type	5-folds average accuracy		Average disparity score (%)
Model1	male	0.9378	2.38
	female	**0.9616**	
Model2	male	0.9224	1.08
	female	0.9332	
Model3	male	0.9428	0.32
	female	0.9460	
Model4	male	**0.9490**	**-0.28**
	female	**0.9462**	

Table 7 Medical text classification disparity score using GloVe (λ is e^{-5} in Model3 and Model4)

Evaluated type	5-folds average accuracy		Average disparity score (%)
Model1	male	0.9466	2.14
	female	**0.9680**	
Model2	male	0.9498	0.96
	female	0.9594	
Model3	male	0.9492	0.32
	female	0.9524	
Model4	male	**0.9656**	**-0.28**
	female	**0.9628**	

Table 8 Medical text classification model's performance with different λs with Word2vec (left) and GloVe (right) using privacy-preserving and negative multi-task learning

λ value	Accuracy		Disparity score (%)	λ value	Accuracy		Disparity score (%)
e^{-6}	male	0.9446	0.71	e^{-6}	male	0.9484	1.45
	female	0.9517			female	0.9629	
e^{-5}	male	**0.9490**	**-0.28**	e^{-5}	male	**0.9656**	**-0.28**
	female	**0.9462**			female	**0.9628**	
e^{-4}	male	0.9342	-1.68	e^{-4}	male	0.9487	1.46
	female	0.9174			female	0.9633	
e^{-3}	male	0.9322	0.5	e^{-3}	male	0.9431	2.38
	female	0.9372			female	0.9669	

with the selective privacy-preserving model realizes the lowest disparity score and highest accuracy. Comparing with Word2vec, GloVe embeddings have a better performance on gender bias mitigation and utility.

Table 8 shows the impact of loss constrains for medical text classification models. Similar with sentiment analysis, λ with e^{-5} achieves the lowest disparity score and good accuracy.

6 Conclusions

In this paper, we presented a negative multi-task learning framework to mitigate gender bias in sentiment analysis and medical text classification. We have demonstrated the effectiveness of our approach by applying our model to the sentiment analysis task and the medical text classification task. We compared Word2vec and Glove word embeddings, and our experimental results showed that Glove performs better on both the accuracy and gender bias mitigation. Our experimental results also showed that our negative multi-task learning framework mitigates the gender bias. It significantly reduced the disparity score on both negative and positive sentiment (2.2 and 3.8 respectively) while achieving the highest F1 score. For medical text classification, our selective privacy-preserving method does protect individuals' sensitive information, and our experimental results showed that integrating with our negative multi-task learning framework, it further mitigates the gender bias. The disparity score reduced 1.86 while achieving good accuracy.

References

[1] Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal **5**(4), 1093–1113 (2014)

[2] Nissim, M., van Noord, R., van der Goot, R.: Fair is better than sensational: Man is to doctor as woman is to doctor. Computational Linguistics **46**(2), 487–497 (2020)

[3] Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems **29** (2016)

[4] Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.-W.: Learning gender-neutral word embeddings. arXiv preprint arXiv:1809.01496 (2018)

[5] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.-W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457 (2017)

[6] Sun, Y., Liu, J., Yu, K., Alazab, M., Lin, K.: Pmrss: privacy-preserving medical record searching scheme for intelligent diagnosis in iot healthcare. IEEE Transactions on Industrial Informatics **18**(3), 1981–1990 (2021)

[7] Hovy, D.: 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 (volume 1: Long Papers), pp. 752–762 (2015)

[8] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.-W.: Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876 (2018)

[9] Sheng, E., Chang, K.-W., Natarajan, P., Peng, N.: The woman worked as a babysitter: On biases in language generation. arXiv preprint arXiv:1909.01326 (2019)

[10] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., Wang, W.Y.: Mitigating gender bias in natural language processing: Literature review. arXiv preprint arXiv:1906.08976 (2019)

[11] Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., Turchi, M.: Gender bias in machine translation. Transactions of the Association for Computational Linguistics **9**, 845–874 (2021)

[12] Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., Zemel, R.: Understanding the origins of bias in word embeddings. In: International Conference on Machine Learning, pp. 803–811 (2019). PMLR

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)

[14] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[15] Leino, K., Fredrikson, M.: Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference. In: 29th USENIX Security Symposium (USENIX Security 20), pp. 1605–1622 (2020)

[16] Song, C., Shmatikov, V.: Overlearning reveals sensitive attributes. arXiv preprint arXiv:1905.11742 (2019)

[17] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)

[18] Coavoux, M., Narayan, S., Cohen, S.B.: Privacy-preserving neural representations of text. arXiv preprint arXiv:1808.09408 (2018)

[19] Shi, W., Cui, A., Li, E., Jia, R., Yu, Z.: Selective differential privacy for language modeling. arXiv preprint arXiv:2108.12944 (2021)

[20] Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)

[21] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)