Towards Fair and Selectively Privacy-Preserving Models Using Negative Multi-Task Learning (Student Abstract)

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

[1] Texas Tech University, 2500 Broadway, Lubbock, 79409, Texas, USA
[2] Lubbock High School, 2004 19th St, Lubbock, 79401, Texas, USA.
liygao@ttu.edu

Abstract

Deep learning models have shown great performances in natural language processing tasks. While much attention has been paid to improvements in utility, privacy leakage and social bias are two major concerns arising in trained models. In order to tackle these problems, we protect individuals' sensitive information and mitigate gender bias simultaneously. First, we propose a selective privacy-preserving method that only obscures individuals' sensitive information. Then we propose a negative multi-task learning framework to mitigate the gender bias which contains a main task and a gender prediction task. We analyze two existing word embeddings and evaluate them on sentiment analysis and a medical text classification task. Our experimental results show that our negative multi-task learning framework can mitigate the gender bias while keeping models' utility.

Introduction

Recent developments in Natural Language Processing (NLP) had 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. Models trained from the source data not only encode but even amplify the bias present in the dataset (Zhao et al. 2017). 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 (Sun et al. 2021). To address these problems, we first apply a selective privacy-preserving method on sensitive word embeddings, and then use the negative multi-task learning framework to train the model. 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 quantitatively measure the gender bias, we use a disparity score to calculate the difference of the model's accuracy between males and females. We evaluated the proposed negative multi-task learning framework on sentiment analysis and a medical text classification task.

Methodology

The sensitive individual information may include: Name, Address, Email, Phone number, Age, Gender, Marital status, Race, Nationality, Religious beliefs, etc. We first define a sensitive information detecting 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 then use the perturb function $P(E)$ to obscure the word embeddings. $P(E) = E + \mathcal{N}(\mu, \sigma, D_E)$, where E is the word embedding of X, $\mathcal{N}(\mu, \sigma, D_E)$ means normally distributed noise with mean μ and variance σ, D_E is the dimension of E.

In this paper, we only consider the "gender" bias, which is a frequently concerned factor in fairness. The process of negative multi-task learning to mitigate the gender bias is as follows. First, we obtain word embeddings from embedding generative models, and then take the text embeddings as input to 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: $Loss$ = $L_{main-task}$ - λ*$L_{gender-prediction}$. λ is the gender prediction loss constraint. We can adjust it to balance the accuracy of the main task and the gender bias. 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 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.

Experimental Setup

We use Word2vec (Mikolov et al. 2013) and GloVe (Pennington, Socher, and Manning 2014) as the basic word embeddings with 100 dimensionality. The perturb function employed on sensitive word embeddings uses (0, 1)-Gaussian noise. Based on the model performance with different λs, the default loss constraint 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

Embedding	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**

Table 1: Disparity score on sentiment analysis (λ is e^{-5} in Group-2 and Group-4).

to estimate the performance change caused by the optimisation on each set individually. In all experiments, we compare the models with the same settings. For model utility, we use F1-score to measure each sentiment class separately. We use balanced accuracy to evaluate the overall performance of the medical text classification task. For fairness evaluation, we use Disparity Score to measure the gender bias (Hardt, Price, and Srebro 2016). The average difference between males and females is described as $Disparity\ Score$:

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

where Acc is the accuracy of each model built in 5-fold cross validation. $Disparity\ Score$ is 0 means that there is no gender bias on the predictions. The closer the disparity score is to 0, the fairer the models are.

Experimental Results and Analysis

In this section, we evaluate the gender bias on sentiment analysis and a medical text classification task. We only apply selective privacy-preserving on medical dataset due to the insufficient sensitive information in sentiment dataset.

Sentiment Analysis

We test four groups to measure the gender bias difference between negative sentiment and positive sentiment. Group-1 and Group-2 have negative sentiment disparity score tested on single-task learning model and negative multi-task learning model respectively. Group-3 and Group-4 have positive sentiment disparity score tested on single-task learning model and negative multi-task learning model respectively. The models' performance on males is less than on females for both negative and positive reviews. That means that the sentiment analysis models are better at identifying sentiment from females than from males. Comparing Word2vec and GloVe, GloVe has better accuracy than Word2vec. From Table 1, we can see that positive sentiment has higher disparity score than negative sentiment on both Word2vec and GloVe, which means there exits a 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.

Medical Text Classification

Evaluated model	Average accuracy		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 2: Medical text classification disparity score using GloVe (λ is e^{-5} in Model3 and Model4).

On the medical text classification task, we investigate the impact of both the selective privacy-preserving method and the negative multi-task learning method for mitigating gender bias. 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 without privacy-preserving handling, and Model4 for the negative multi-task learning with selective privacy-preserving. Table 2 shows our experimental results using GloVe. the negative multi-task learning with the selective privacy-preserving model realizes the lowest disparity score and highest accuracy.

Conclusion

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 these two tasks. We also demonstrated that our proposed selective privacy-preserving method does protect individuals' sensitive information, and it further mitigates the gender bias with the negative multi-task learning framework. In the future, we plan to apply the negative multi-task learning framework to mitigate multiple bias simultaneously, such as age, race, etc.

References

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

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

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

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

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.