We conduct the experimental evaluation on three tabular datasets and two text datasets. For tabular datasets, we trained three models for each dataset, which are MLP, SVM, and RF. For text datasets, we trained two models for each dataset, which are CNN and LR. All of these models are uploaded to the **unfair\_models folder**.

After the inputting of models and initial data into ExpGA, for a specific protected attribute, a large number of discriminatory samples will be generated. We uploaded these sample files to the **experiment\_data folder**.

Part of discriminatory samples is used to retrain models, aim at decreasing the bias on the original models without changing the accuracy of classifying. The retrained models are uploaded to the **retrain\_model folder**. The efficiency and effectiveness of ExpGA on retrained models are much lower than that on the original models.

Here we present a generated discriminatory sample on the IMDB dataset. Given the gender as the protected attribute, a pair of sentences on IMDB datasets in the following is discriminatory samples on the original CNN model, we add “male” and “female” before “master” in two sentences respectively, the predicted labels are different, the sentence with “male master” is positive, while sentence with “female master” is negative.

… Robertson is a male master of pace, camera, angles and montage…

… Robertson is a female master of pace, camera, angles and montage…

After the model retraining, both sentences are predicted as positive on the new CNN model.