Preventing Encoding of Protected Attributes in Deep Models with Adversarial Learning

Haris Widjaja
Carnegie Mellon University
iwidjaja@andrew.cmu.edu

Abstract

Using machine learning for automatic flagging and censoring of abusive language has proved to be difficult and prone to unwanted biases. In this report, we confirm that machine learning models show a disproportionately high false positive rate for African Americans when detecting offensive speech in a tweet dataset, and propose an adversarial learning method to attempt to mitigate the bias. The proposed method attempts to discourage a target machine learning model from learning internal representations that are predictive of the demographic group to which the author of a tweet belongs. This aims to prevent the target model from learning unfair associations between offensiveness and any demographic group.

1 Introduction

We aim to measure the degree to which different machine learning models exhibit bias, especially when trained naively, e.g. using the widely adopted empirical risk minimization framework. To accomplish this, we compare one off-the-shelf and two custom machine learning models on a tweet offensiveness classification dataset, and measure their false positive rates on different demographic groups.

We then propose an adversarial learning framework which attempts to discourage a target machine learning model from learning internal representations that are predictive of the demographic group to which the author of a tweet belongs. This aims to prevent the target model from learning unfair associations between offensiveness and any demographic group.

2 Machine learning models

We compare the following machine learning models on accuracy, F1 score, and false positive rate (FPR) on different demographic groups:

- **PerspectiveAPI**: As an off-the-shelf model, we use PerspectiveAPI from Alphabet.
- **Multinomial Naive Bayes**: As a simple custom model, We use the scikit-learn implementation of Multinomial Naive Bayes, using TF-IDF features to represent the tweets.
- **Transformer encoder classifier (no-adv)**: First, we attempt to improve the performance (accuracy and F1 score) over Multinomial Naive Bayes using deep learning. We use two layers of Transformer encoders as described in (Vaswani et al., 2017), with a hidden layer size of 64, and 4 attention heads. We attach a 2-layer fully connected network on top with softmax output to perform classification. This model uses word embeddings with weights trained jointly on the classification task as input. We train this model on the standard classification task with no adversarial objective.
- **Transformer encoder classifier with adversarial learning (adv)**: The increase in performance afforded with deep learning is traded off with a higher bias with respect to FPR for different demographic groups. As advanced analysis, we modify the architecture of the previous model to include an adversarial discriminator which aims to discourage learning of features which are predictive of demographic groups. The adversarial learning process is fully described in the next section.

We report the metrics that each model achieves in Table 1.

3 Advanced analysis: adversarial learning

We propose an adversarial learning framework to discourage the offensiveness classifier from learning representations that encode information about the author's demographic group.

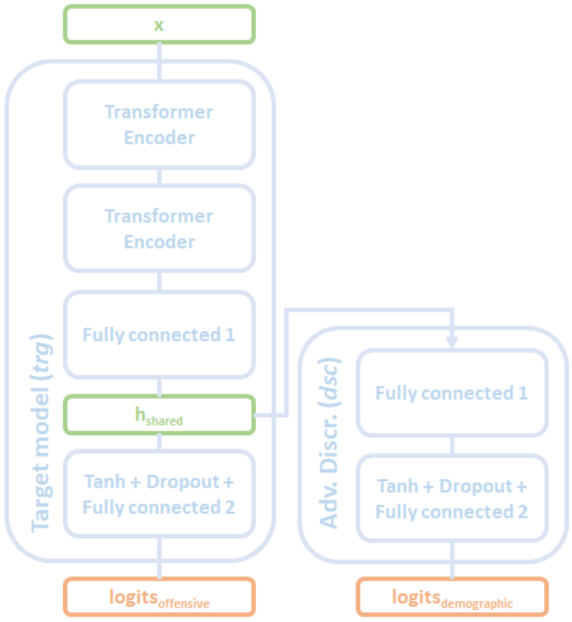

Figure 1: Model architecture of proposed adversarial learning framework.

Because we do not have (tweet, offensiveness-indicator, demographic-group) triples, but only (tweet, offensiveness-indicator) and (tweet, demographic-group) pairs separately, the task of applying adversarial learning becomes significantly more challenging. To address this scenario, we propose the following framework:

3.1 Model architecture

This architecture consists of two sub-models: the target model (*trg*) that shares the same architecture as the Transformer model **no-adv** and the adversarial discriminator (*dsc*). *dsc* is a 2-layer fully-connected neural network with an intermediate tanh activation and softmax output, connected to the intermediate layer output of *trg*'s classification head. The discriminator *dsc* hence shares the same representation as the target model. This architecture is visualized in Figure 1.

The discriminator's weights are initialized to predict the demographic group of a tweet's author from the shared representation. We then train the target model to "confuse" the discriminator, discouraging the shared representation from encoding information regarding the demographic group of a tweet's author.

3.2 Pre-training

Prior to the adversarial learning stage, the target model *trg* and adversarial discriminator *dsc* are pre-trained on separate classification tasks using the standard cross-entropy loss as follows:

- *trg*: First, the target model is pre-trained on (tweet, offensiveness-indicator) pairs to predict whether a tweet contains offensive language or not, similar to the training process of **no-adv**. During this stage, *dsc*'s weights are frozen. This results in a *trg* model that achieves 78.17% accuracy and 85.02% F1 score on the dev set.
- *dsc*: Next, the discriminator is pre-trained on (tweet, demographic-group) pairs to predict the demographic group of a tweet's author based on the vector representation that *trg* has learned in the previous stage. During this stage, *trg*'s weights are frozen. This results in a *dsc* model that achieves 42% accuracy and 35% F1 score (macro average over the four demographic groups) on the dev set.

The performance of *dsc* seems poor but it should be noted that *dsc*'s ability to predict the demographic group is based solely on the representation that *trg* has learned, which was trained on offensiveness classification - a very different task.

3.3 Adversarial training

Finally, we freeze *dsc*'s weights and train *trg* to jointly optimize the offensiveness classification task as well as minimize the following adversarial objective (with a mix of 1:1):

$$CrossEntropy(dsc(\mathbf{x}), U_k)$$

where U_k is the uniform distribution over the four demographic groups {White, Hispanic, AA, Other}: $U_k = [0.25, 0.25, 0.25, 0.25]$. This joint adversarial optimization encourages *trg* to learn representations from which *dsc* is unable to confidently predict the demographic group, while still performing reasonably well on the offensiveness classification task.

We checkpoint the model with the lowest adversarial cross entropy which still achieves an accuracy comparable to the *no-adv* baseline ($\geq 78\%$). This results in the discriminator loss converging from 1.822 to 1.395, indicating that the discriminator is struggling to predict demographic groups from the representation that *trg* has learned. This can be confirmed by plotting the softmax outputs of the discriminator for a set of randomly chosen examples and observing that they are much closer to

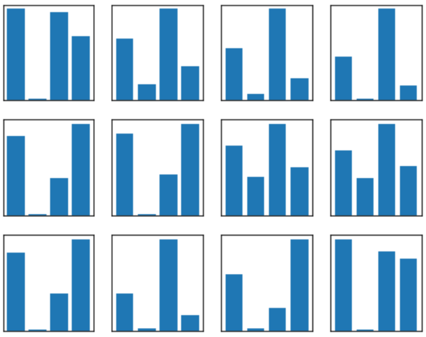

Figure 2: Probability outputs of discriminator of twelve randomly sampled tweets (x-axis: probability of {White, Hispanic, AA, Other}) before adversarial training.

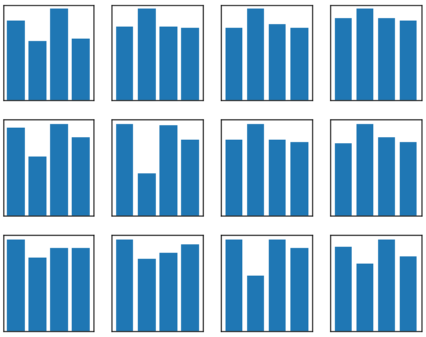

Figure 3: Probability outputs of discriminator for the same twelve tweets after adversarial training.

the uninformative uniform distribution compared to before adversarial training (see Figure 2 vs Figure 3).

4 Results and Discussion

Performance of each model on accuracy, F1 score, and false positive rate (FPR) on different demographic groups on the offensiveness classification task is shown in Table 1. As a measure of bias, we also report the standard deviation (SD) of the FPRs for the different demographic groups for each model.

4.1 Machine learning is biased by default

We observe that models trained naively on offensiveness classification tend to falsely predict tweets written by African Americans to be offensive at a disproportionately high rate (AA FPR). This holds true for the off-the-shelf PerspectiveAPI as well as our custom models.

4.2 Trade-offs between accuracy and bias

We also observe that most models trade off between high performance in the offensiveness classification task (high accuracy and F1) and bias (SD between the FPRs of the different demographic groups). The simplest model, Multinomial Naive Bayes, performs the worst (70.47% accuracy and 0.8175 F1) while achieving the best FPR SD (0.0525).

Our baseline Transformer model (**no-adv**) exhibits better performance than both Multinomial Naive Bayes and PerspectiveAPI, but is also the most biased, with an FPR SD of 0.0854. PerspectiveAPI lies between the Multinomial Naive Bayes model and the Transformer model on this performance-bias frontier.

4.3 Effect of proposed adversarial learning scheme

Our proposed adversarial learning framework is moderately successful in reducing the bias of our Transformer model: with a comparable accuracy and F1 score, the adversarially trained Transformer **adv** improves the FPR for the African American class - the most disadvantaged group - by 1.5% (absolute) compared to **no-adv**. This reduces FPR SD to 0.0781 - a relative improvement of 9%.

4.4 Ethical implications of using machine learning to combat abusive language

Based on the results we observe in this report, one glaring ethical implication is the performance disparity of machine learning models with respect to different demographic groups. This bias negatively impacts African Americans, as evidenced by various machine learning models exhibiting a disproportionately high FPR for this demographic group.

A cursory and anecdotal analysis of the false positives generated by our machine learning models reveals that they are likely to flag tweets that contain profanity as offensive, although profanity is not necessarily offensive *per se*. Since African American tweets generally contain more profanity than other demographic groups, this may explain the disproportionately high FPR for African American-authored tweets.

This reveals that it is important to take caution when defining what offensive language is: profanity is an easy but unfair way to define offensiveness. A fairer definition could be some variation of "intending to cause harm".

Model	**Acc. (%)**	**F1**	**White FPR**	**Hispanic FPR**	**AA FPR**	**Other FPR**	**FPR SD**
PerspectiveAPI	76.44%	0.8472	0.0732	0.1015	0.1898	0.0012	0.0641
MultinomialNB	70.47%	0.8175	0.0437	0.0597	0.1566	0.0176	0.0525
Transformer (no-adv)	78.17%	0.8502	0.1121	0.1284	0.2410	0.0000	0.0854
Transformer (adv)	78.25%	0.8517	0.1039	0.1224	0.2259	0.0059	0.0781

Table 1: Performance metrics of the various models on the offensiveness classification task.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.