Adversarial Removal of Demographic Attributes from Text Data

Yanai Elazar† and **Yoav Goldberg**†*
†Computer Science Department, Bar-Ilan University, Israel
*Allen Institute for Artificial Intelligence
{yanaiela,yoav.goldberg}@gmail.com

A Standard Fairness Definitions and Guarded Classifiers

In this work, we focus on creating a representation which is oblivious to some factor. We measure this by the term *leakage* which is defined in Section 2 and say that a classifier is *guarded* in respect to an attribute z if z is *guarded* and hasn't *leaked*. In this section, we show that this definition matches more common definitions of fairness, under our setup. Specifically, we show that under the setup we discuss, if z is guarded than the classifier satisfies *Demographic Parity*, *Equality of Odds* and *Equality of Opportunity*.[1]

For completeness, we repeat these definitions provided and redefined by Hardt et al. (2016), Beutel et al. (2017) and Zhang et al. (2018):

Demographic Parity. A predictor f satisfies *demographic parity* if f and z are independent:

$$P(f=\hat{y}|z=0) = P(f=\hat{y}|z=1)$$

Equality of Odds. A predictor f satisfies *equality of odds* if f and z are conditionally independent of the main task label Y:

$$P(f=\hat{y}|z=0, Y=y) = \\ P(f=\hat{y}|z=1, Y=y), y \in \{0,1\}$$

Equality of Opportunity. A predictor f satisfies *equality of opportunity* if f and z are conditionally independent on a particular value Y:

$$P(f=\hat{y}|z=0, Y=y) = \\ P(f=\hat{y}|z=1, Y=y), y \in \{0|1\}$$

Recall that in all our data split setups, there is an equal appearance of both y and z. We now claim the following:

[1] We note however that, as we discussed, achieving 0-leakage is far from trivial.

Lemma A.1. *If a classifier is guarded to z in our setup, it realizes demographic parity*

Proof. The classifier is oblivious to z, meaning that: $P(Z|H) = 0.5$, where H is the internal representation. As in our setup, $P(Z) = 0.5$

$$\Rightarrow P(Z) = P(Z|H)$$

therefore Z and H are independent.

From graphical models we know that Y and X (the textual input), are independent given H (symmetrically for Z and X), therefore we can overlook X when conditioning on H. Also, we can say that given H, $\hat{Y}$ and Z are independent:

$$P(\hat{Y}, Z|H) = P(\hat{Y}|H)P(Z|H)$$

using bayes rule on both sides:

$$\frac{P(H|\hat{Y}, Z)P(\hat{Y}, Z)}{P(H)} = \frac{P(H|\hat{Y})P(\hat{Y})}{P(H)}P(Z|H)$$
$$\Rightarrow P(H|\hat{Y}, Z)P(\hat{Y}, Z) = P(H|\hat{Y})P(\hat{Y})P(Z|H)$$

and from the independency of Z and H, we get that:

$$P(\hat{Y}, Z) = P(\hat{Y})P(Z|H)$$

As Z and H are independent:

$$P(\hat{Y}, Z) = P(\hat{Y})P(Z)$$

meaning that $\hat{Y}$ and Z are independent □

Lemma A.2. *If a classifier is oblivious to z in our setup, it realizes equality of odds.*

Proof. If a classifier is *oblivious* to z, this means that:

$$P(f=\hat{y}|z=0) = P(f=\hat{y}|z=1)$$

(from Lemma A.1), therefore, from the law of total probability:

$$P(f=\hat{y}|z=0,y=0)\cdot P(y=0)+$$
$$P(f=\hat{y}|z=0,y=1)\cdot P(y=1)=$$
$$P(f=\hat{y}|z=1,y=0)\cdot P(y=0)+$$
$$P(f=\hat{y}|z=1,y=1)\cdot P(y=1)$$

Since $P(y=0)=P(y=1)=0.5$ in all setups, we can get rid of $P(y)$, and we get:

$$P(f=\hat{y}|z=0,y=0)+$$
$$P(f=\hat{y}|z=0,y=1)=$$
$$P(f=\hat{y}|z=1,y=0)+$$
$$P(f=\hat{y}|z=1,y=1)$$

and due to Lemma A.1 we get

$$P(f=\hat{y}|z=0,y=1)=P(f=\hat{y}|z=1,y=1)$$

(and similarly for $y=0$). $\square$

Lemma A.3. *If a classifier is oblivious to z in our setup, it realizes equality of opportunity.*

Proof. As equality of opportunity is a relaxation of *equality of odds* and is less strict then it, from Lemma A.2, it holds automatically. $\square$

In conclusion, we showed that an *oblivious classifier* on our setup would satisfy the three fairness definitions which were introduced above.

B Implementation Details

Preprocessing We tokenize each tweet using *twokenize*, a twitter specific tokenizer (O'Connor et al., 2010; Owoputi et al., 2013), and discard duplicate tweets and tweets with less than three tokens.

Neural Network architecture and hyperparameters Unless otherwise noted, both the LSTM encoder and the MLP hidden layer have 300 hidden units with a single layer. We use randomly-initialized 300-dimensional embeddings, and train using SGD with Momentum (Qian, 1999) and a learning rate of 0.01 for 100 epochs. We use dropout (Hinton et al., 2012) of 0.2 on all hidden layers, and negative log likelihood as the loss function with 32 sized mini-batch.

C Emojis Details

Figure 1 contains the different emojis we used for defining positive and negative tweets. In addition to those emojis, we also looked for the following emoticons: :) :-) :) :D =) as positive and :(:-(: (:-(=(as negative

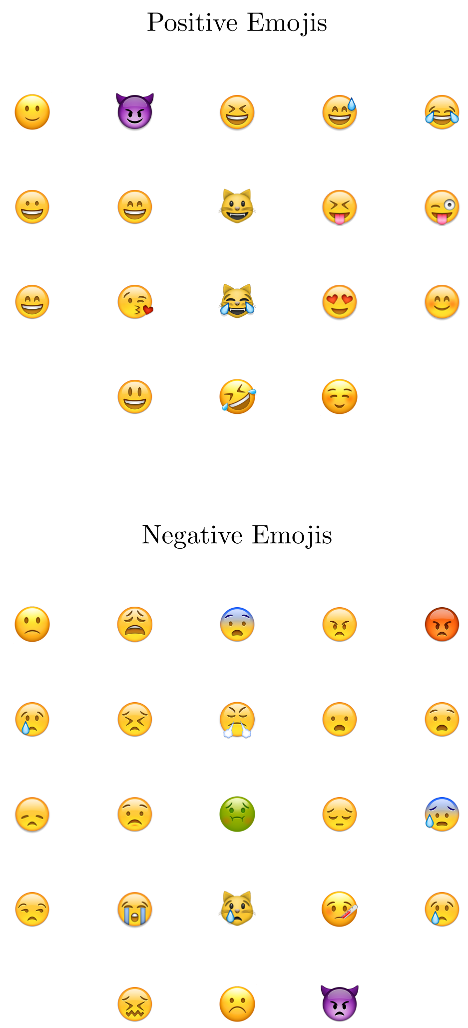

Figure 1: Emojis used as positive and negative proxies for sentiment

References

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012.

Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, pages 384–385.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Ning Qian. 1999. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. *arXiv preprint arXiv:1801.07593*.