ALGORITHMIC FAIRNESS IN POST-PROCESSED TOXICITY TEXT CLASSIFICATION

Abishrant Panday
abishrantpanday@college.harvard.edu

Joyce Tian
joycetian@college.harvard.edu

May 7, 2020

1 Introduction

With increasing reliance on automated systems which make data-driven classifications, there are growing concerns in the potential that such systems learn prior unjustified prejudices and further such biases in continued classification. In this work, we will study the applicability of post-processing classification algorithms to promote fairness with respect to how the algorithm classifies certain demographics or similar individuals. A common fairness measure used is group fairness, otherwise known as statistical parity, which enforces the condition that the expected classification of members within a protected group match the overall population classification. However, in providing only a constraint over what is potentially a broad sweep of individuals, enforcing group fairness may still yield unfair outcomes; for instance, Dwork et al. (2011) outlined the concept of a self-fulfilling prophecy, wherein unqualified members within the protected group are accepted with expectations that they will fail to justify the prior bias against the group. As such, Dwork et al. (2011) proposed instead the notion of individual fairness, wherein fairness is achieved if individuals deemed similar to each other are classified similarly. We will aim to examine the relationship between both group and individual fairness and specifically attempt to enforce individual fairness, which is the stricter of the two constraints.

Kim et al. (2018) proposed the notion of post-processing a classification algorithm, which is particularly useful in our current society as many algorithms have already been implemented without prior checks for fairness, and thus would benefit from the addition of such a technique. The post-processing involves a switching subgradient descent (SSGD) which adjusts the predictive weights primarily for any fairness violations within a subset of the population and secondarily moves along the objective function to balance fairness and accuracy. Their theoretical results proved that metric multifairness, a relaxation of individual fairness which focuses on fairness within smaller collections of similar individuals, can be achieved while only sampling a fraction of the original data. However, the convergence specified by the algorithm in Kim et al. (2018) is quite large, being quadratic in the number of samples. Further, it is not immediately clear what the smaller collections of similar individuals should be to ensure more intersectional notions of fairness.

To that end, we aimed to investigate the implications of the SSGD algorithm on the Jigsaw Unintended Bias in Toxicity Classification dataset. The dataset was created following the publication of several queries of the 2017 Perspectives API, which identified "I am a man" as being 20% likely to be toxic, "I am a woman" as 41% likely to be toxic, and "I am a black man" as 85% likely to be toxic. This disparity in toxicity probabilities for very similar sentences seems to indicate a strong bias against certain identity labels; since the Perspectives API was meant to automatically hide comments deemed toxic, this represented a distinct harm to certain demographics in being unfairly censored. To combat this, the dataset published includes comments, tags for any identity labels contained within the comments, and the probability of toxicity as assigned by several impartial arbiters, with the aim of maintaining comment toxicity classification accuracy

while also allowing for tracking of the accuracy for certain demographics. Thus, it is ideal for our use-case to attempt fair classification between various identity labels and the overall population. We will specifically only examine the [].

Thus, our motivations were to first verify the validity of the similarity and individual fairness metrics we specify, and use them to examine the viability of SSGD in augmenting the fairness of text toxicity classification, and using the fairness and accuracy shifts over the course of post-processing to visualize the rate of convergence for various subset designations. Further, we would also like to examine whether the metric multi-fairness constraint in Kim et al. (2018) generalizes to increasing both individual fairness and group fairness.

2 Methods

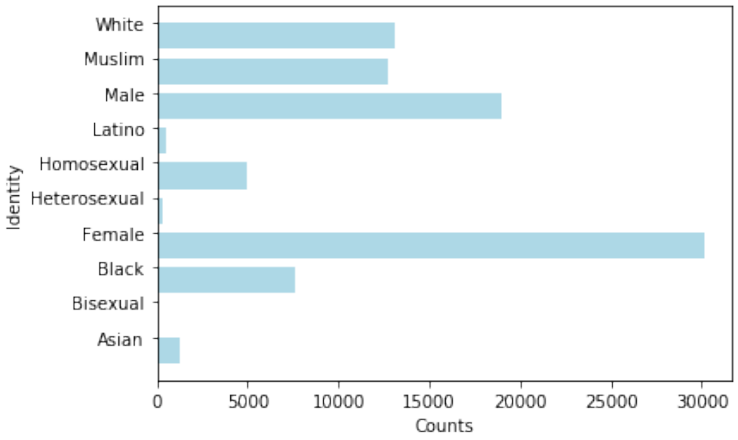


Figure 1: Frequency of Annotated Identity Labels over Entire Toxicity Dataset.

First, we trimmed our dataset to only contain identity labels for whether the comment text mentioned a definitive gender, race, or sexual orientation, leaving 10 label tags which we will treat as the protected groups within this dataset. We also culled any metadata regarding temporal or regional IDs, as they were inconsistent in placement and existence between comments. We only included annotated comments, those checked explicitly for identity labels by multiple third parties, to ensure there was a minimized chance for lack of identification of the inclusion of a protected group within the comment text. We then sampled 25% of these points to yield a dataset X with $|X| = 180487$ data points. To ensure we had enough labeled samples to be statistically relevant, we focused on the comments containing "male" and the comments containing "female" identity labels as gender was the most common identity mentioned per Figure 1.

Our similarity metric was derived from cosine similarity of the average word embedding of the comment text, which is commonly performed as a measure of sentence similarity in natural language processing. The word embeddings were taken from the GloVe-100 dataset, which yields 100-dimensional vectors for words as found within Wikipedia text. In addition to the average word embedding, we added 0-1 classifiers for whether an identity label was included or not for 110-dimensional embeddings overall. We then performed a normalization of each embedding, with the similarity metric being the Euclidean distance, since the Euclidean distance of normalized vectors gives the same relative measure as the cosine similarity distance.

For our individual fairness metric, we used the Zemel et al. (2013) consistency metric I, which defines fairness for some data point $x_n \in X$, $|X| = N$, as the discrepancy between its predicted value $\hat{y}_n$ and the predicted values for its k-nearest neighbors, $kNN(x_n)$, based on the similarity metric given:

$$I = 1 - \frac{1}{Nk} \sum_{x_n \in X} \left| \hat{y}_n - \sum_{x_j \in kNN(x_n)} \hat{y}_j \right|$$

This measure thus provides a sense of whether the comments most similar to each other are similarly classified. Here, a larger I value would indicate more similar judgements between similar individuals, and thus higher individual fairness.

We further used the Zemel et al. (2013) discrimination metric D_S as a proxy for measuring group fairness:

$$D_S = \left| \frac{\sum_{x_i \in S} \hat{y}_i}{|S|} - \frac{\sum_{x_j \in S^c} \hat{y}_j}{|S^c|} \right|$$

where S is the set of datapoints which contain the sensitive feature, thus showing the average discrepancy between points containing the feature and not. Here, a smaller D_S value would indicate less discrimination, and thus higher group fairness.

Our base model which took as input the comment text and outputted the toxicity probability. This structure emulates the basic architecture used in the traditional natural language processing classification models. We also examined an augmented model which predicts toxicity probability and 0-1 values for the 10 identity markers. This data augmentation, as described in Nguyen et al. (2011), is a common technique which has been shown to increase classification accuracy significantly. Both models consist of an embedding layer and 50 LSTM units with 0.2 dropout and sigmoid activation, used the Adam optimizer, and ran for 3 epochs with batch sizes of 128.

In postprocessing, the SSGD algorithm will shift the LSTM layer weights w_0 primarily on the gradient of any violated fairness constraints and secondarily on the usual objective function. The estimated residual query for a given subset $S \subsetneq X$, $\hat{R}_S(w)$, gives a ballpark figure for

$$E_{(x_i,x_j)\in S\times S}[\hat{y}_i - \hat{y}_j] - E_{(x_i,x_j)\in S\times S}[d(x_i, x_j)],$$

where $d(x_i, x_j)$ is the similarity distance of the comment text as defined earlier; this measure thus notes how different the prediction values are from the perceived similarity of the comments. The estimate is done by only computing the pair predictive difference and distances for the values in S which were picked from a general sampling of X. We will use a tolerance of $\tau = 0.05$ and $T = 100,000$ total steps. Every 100 steps, group fairness was calculated for the set containing all "men"-labeled comments, M, and the set containing all "women"-labeled comments, W, as well as the individual fairness metric I.

SSGD(w_0, τ, T, C){
  for($k = 0;\ k < T;\ k++$){
    if ($\exists S \in C$ such that $\hat{R}_S(w_k) > 4\tau/5$) { \\ fairness violation detected
      $S_k \leftarrow$ any $S \in C$ such that $\hat{R}_S(w_k) > 4\tau/5$; \\ pick a subset on which fairness was violated
      $w_{k+1} \leftarrow w_k - \frac{\tau}{M^2}\nabla R_{S_k}(w_k)$; \\ step per fairness gradient on the subset
    }
    else{ \\ fair weights found
      $W \leftarrow W \cup \{w_k\}$; \\ add to set of known fair weights
      $w_{k+1} \leftarrow w_k - \frac{\tau}{GM}\nabla L(w_k)$; \\ step per objective gradient
    }
  }
  return $\bar{w} = \frac{1}{|W|}\sum_{w\in W} w$ \\ output average of fair weights
}

We noted that there is significant computation involved in the number of metric samples specified by Kim et al.(2018) for convergence of SSGD, on the order of $|X|^2 \approx 10^{10}$ per step. To test whether this could be reasonably minimized, we also performed an altered switching subgradient descent, SSGD', which used an estimation technique for $\hat{R}_S(w_k)$ which, instead of considering a flat number of pairs, sampled 100 of the points within the subset and found the residual for those points. This clearly adds additional uncertainty to the computed potential unfairness, but is far less computationally expensive to compute.

We considered several sets of subsets of X, which we label C, to examine the impact of distance of points within each subset, the perceived importance of using identity labels to create the subsets, and the impact of containing overlapping sets. For our first collection C_1, we identified three basic disjoint sets: comments which included "men", comments which included "women", and comments which included neither. For the second collection, C_2, we considered a similar design, except within each segment ("men", "women", and neither-labeled), we created 3 further strict subsets using a clustering algorithm within each segment for 12 subsets total. Finally, for C_3, we created 12 clusters over the dataset such that all attempted to minimize distance such that there was overlap between clusters.

3 Results

One of the critical limitations of individual fairness is the lack of specificity within designations for a similarity metric between data points. There has not been a generalizable claim as to what metrics can be considered fair in judging the level of similarity of individuals. Thus, we preliminarily evaluated the effectiveness of our similarity metric on a series of simple sentences containing a pronoun (“I” or “You”) and identity labels.

The resulting similarity metrics rated "I am" statements much closer than "You are" statements, and vice versa, and rated inclusions of similar gender, race, or sexual orientation as making statements more similar, with penalties for opposite gender, race, or sexual orientation. This is desirable, as it shows acknowledgement of various axes of identities while also analyzing overall sentence structure similarity. Without label inclusions, the distance for the inclusion of labels was clustered by identity type (gender, race, or sexual orientation), further noting a general similarity in treatment for each label grouping.

Table 1: Base Model and Augmented Model Metrics

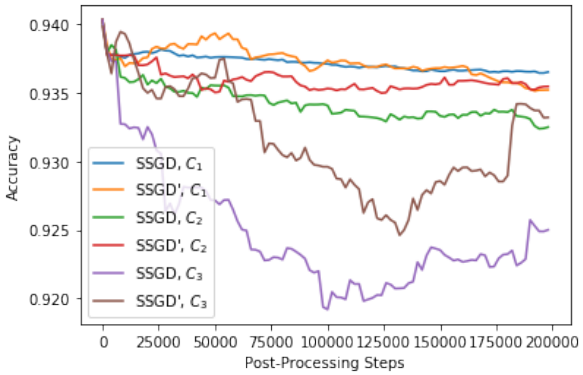
Model Type	Accuracy	D_M	D_W	I
Base	90.43%	0.0519	0.0827	0.5434
Augmented	94.03%	0.0514	0.0834	0.5585

Using this similarity metric, we found the k-nearest neighbors for each comment in X. This took a large amount of computational effort, with the algorithm running for nearly 2 days. In Table 1, we see the model accuracies, the discrimination with respect to the set of "men"-labeled comments, M, and the set of "women"-labeled comments, W, as well as the consistency as found with $k = 15$. As expected, the augmented model is more accurate than the base model Examination of this reveals that generally, the comments containing "men" has a lower discrimination value than comments containing "women", indicating that W as a set experiences more group unfairness and implying the results are generally more discrepant for "women"-labeled comments. Interestingly, we also note that the augmented model has a wider gap between D_M and D_W but a slightly higher individual fairness metric. Predicting identity labels in our augmented model necessarily emphasizes the importance of identity labels, and thus may have augmented the distinguishment between . To attempt to improve upon the best model we had, we proceeded to perform post-processing on the augmented model so as to determine the best level of individual and group fairness we could achieve using our framework.

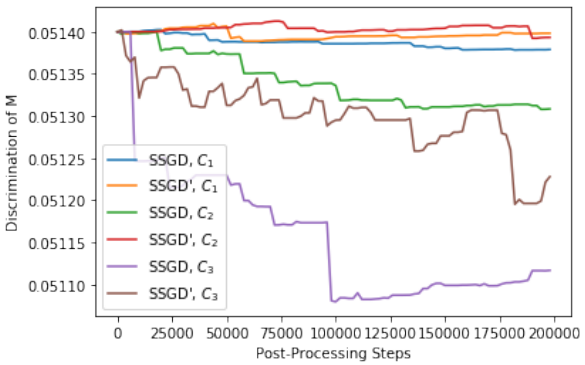
Examining the impact of post-processing generally, as shown in Figure 2, we found that the SSGD’ algorithm is much more volatile in terms of its output relative to all measures, and especially with regards to accuracy; this implies it is much more volatile about whether any fairness violation is found, and the gradient in the case where a violation is found, which makes sense given the smaller sample size. However, even the traditional SSGD algorithm performed several jumps in discrimination and consistency values, which may suggest our sampling was too small for that algorithm as well. However, we note that while accuracy decreased for all post-processed algorithms, all also had at least marginal improvements in terms of improved consistency and lessened discrimination, indicating improvements in both group fairness and individual fairness as a result of this post-processing. It also seems that generally D_M was more volatile

than D_W, indicating it was likely the sets containing more "men"-labeled comments which violated more fairness constraints; we are not entirely sure as to the reasoning, but we do note that there are significantly fewer "men"-labeled comments generally, which may mean it is easier to violate fairness constraints due to decreased sampling. We also note a generally positive trend between increasing group fairness and increasing individual fairness.

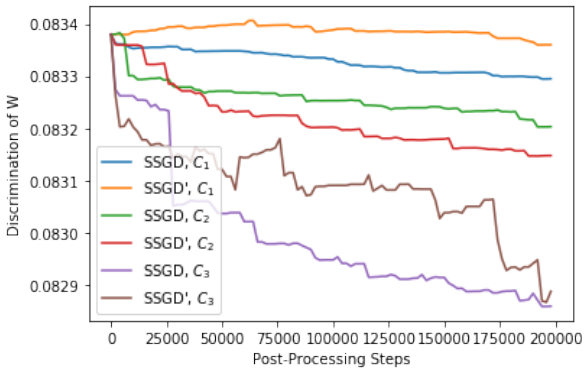
Further, the SSGD and SSGD' algorithms showed very distinct results based on which of the three collections was used. The algorithms using C_1 yielded little to no results, which makes sense as disjoint sets do not give much holistic information and further were likely too large to be useful with regards to violating fairness constraints. The algorithms using C_2 slightly decreased discrimination and increased consistency, which indicates that the notion of intersecting, or even nested, subsets are useful in improving fairness constraints. The algorithms using C_3 were by far the most visibly successful at helping the algorithm improve both individual and group fairness bounds, despite not being explicitly based on the metric of gender. This seems to indicate that the best way to create the collection of sets is to minimize distance, rather than explicitly account for disjoint sets of the features upon which we would like to improve fairness.



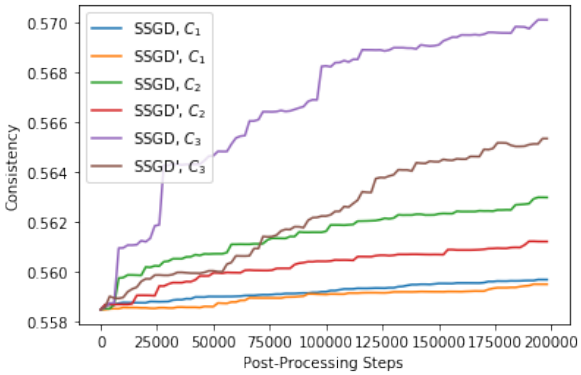
(a) Model Accuracy using SSGD or SSGD'.



(b) Model D_M using SSGD or SSGD'.



(c) Model D_W using SSGD or SSGD'.



(d) Model Consistency (I) using SSGD or SSGD'.

Figure 2: Model accuracy, discrimination, and consistency metrics based on varying post-processing techniques and sets considered.

While the SSGD algorithm was much less volatile compared to the SSGD' algorithm, it also took far longer to run; all SSGD algorithms took over a day for 200,000 steps, and the algorithm using the C_3 collection took over three, and this large computation prevented us from finding any convergence within the fairness contraints or accuracy. Meanwhile, the SSGD' algorithm took at most 8 hours for 200,000 steps regardless of the collection used, which makes it a potential to explore further with regards to whether it can converge given more steps.

Finally, we examined the consistency of the 2017 Perspectives model, our base model, and the post-processed models based on the C_3 collection based on the queries given to the 2017 Perspectives API, as listed in Table 2. We have

found a large improvement between the 2017 model and our base model, with another slight improvement between our base and augmented model. The largest increase is between the augmented model and the SSGD-postprocessed models, which performed similarly. However, we note that since these were very simple sentences where the only differences were the inclusion of varying identity labels that there is still much progress to be made, as an algorithm should distinctly be capable of obtaining near 100% consistency with respect to this simple set of queries.

Table 2: Consistency over 2017 Queries

Model Type	Consistency, k = 4
2017 Perspectives	52.44%
Base	68.35%
Augmented	70.75%
Augmented-SSGD (C_3)	75.75%
Augmented-SSGD' (C_3)	78.5%

4 Discussion

We verified the effectiveness of using our defined similarity and distance metrics for use in determining individual fairness, and showed the viability of using the post-processing SSGD algorithm proposed by Kim et al. (2018) in enforcing a proxy for individual and group fairness on the task of text classification with sensitive labels. The major obstacle we encountered was the large computational burden of calculating the k-nearest neighbors for the individual fairness consistency metric, and the large amount of pair prediction differences needed for each step of the post-processing residual estimates. Further, it is still unclear from our findings how to efficiently determine an appropriate collection of subsets for the dataset. Our C_3 collection took several hours to compute, and given our general finding that more robust collections are better for enforcing fairness constraints, creating an efficient algorithm to find such collections would be useful for furthering the applicability of this metric multi-fairness enforcing post-processing technique.

Other areas for further research include examining other metrics for text similarity; a fundamental assumption in our use-case was that the GloVe word embeddings are not overly biased and have similar semantic embedding to the word use-cases in our dataset. To make this more generalizable, it may be useful to include within the model a BART custom word embedding model, and to try to make that a fair representation using Naive Bayes as performed in Zemel et al. (2013). Orthogonally, one could also explore using counterfactual models of fairness to create a more interpretable schema for determining the current fairness of an algorithm, and perhaps use SSGD to find the impact of fairness in this framework.

References

[1] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).

[2] Kim, M., Reingold, O., & Rothblum, G. (2018). Fairness through computationally-bounded awareness. In Advances in Neural Information Processing Systems (pp. 4842-4852).

[3] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. In International Conference on Machine Learning (pp. 325-333).

[4] Nguyen, Q., Valizadegan, H., & Hauskrecht, M. (2011, December). Learning classification with auxiliary probabilistic information. In 2011 IEEE 11th International Conference on Data Mining (pp. 477-486). IEEE.

[5] Bechavod, Y., Jung, C., & Wu, Z. S. (2020). Metric-Free Individual Fairness in Online Learning. arXiv preprint arXiv:2002.05474.

[6] Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems (pp. 3992-4001).

[7] Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

[8] Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., & Varshney, K. R. (2020, February). Data Augmentation for Discrimination Prevention and Bias Disambiguation. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 358-364).

[9] Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018, July). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp.2239-2248).